Math Refresher A

Basic Mathematical Tools

This Math Refresher covers some basic mathematics that are used in econometric analysis. We summarize various properties of the summation operator, study properties of linear and certain nonlinear equations, and review proportions and percentages. We also present some special functions that often arise in applied econometrics, including quadratic functions and the natural logarithm. The first four sections require only basic algebra skills. Section A-5 contains a brief review of differential calculus; although a knowledge of calculus is not necessary to understand most of the text, it is used in some end-of-chapter appendices and in several of the more advanced chapters in Part 3.

A-1 The Summation Operator and Descriptive Statistics

The **summation operator** is a useful shorthand for manipulating expressions involving the sums of many numbers, and it plays a key role in statistics and econometric analysis. If $\{x_i: i = 1, \ldots, n\}$ denotes a sequence of n numbers, then we write the sum of these numbers as

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n. \qquad [A.1]$$

With this definition, the summation operator is easily shown to have the following properties:

Property Sum.1: For any constant c,

$$\sum_{i=1}^{n} c = nc. \qquad [A.2]$$

Property Sum.2: For any constant c,

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i. \qquad [A.3]$$

Property Sum.3: If $\{(x_i, y_i): i = 1, 2, \ldots, n\}$ is a set of n pairs of numbers, and a and b are constants, then

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}y_i. \qquad \text{[A.4]}$$

It is also important to be aware of some things that *cannot* be done with the summation operator. Let $\{(x_i, y_i): i = 1, 2, \ldots, n\}$ again be a set of n pairs of numbers with $y_i \neq 0$ for each i. Then,

$$\sum_{i=1}^{n}(x_i/y_i) \neq \left(\sum_{i=1}^{n}x_i\right)\Big/\left(\sum_{i=1}^{n}y_i\right).$$

In other words, the sum of the ratios is not the ratio of the sums. In the $n = 2$ case, the application of familiar elementary algebra also reveals this lack of equality: $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. Similarly, the sum of the squares is not the square of the sum: $\sum_{i=1}^{n}x_i^2 \neq (\sum_{i=1}^{n}x_i)^2$, except in special cases. That these two quantities are not generally equal is easiest to see when $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Given n numbers $\{x_i: i = 1, \ldots, n\}$, we compute their **average** or *mean* by adding them up and dividing by n:

$$\bar{x} = (1/n)\sum_{i=1}^{n}x_i. \qquad \text{[A.5]}$$

When the x_i are a sample of data on a particular variable (such as years of education), we often call this the *sample average* (or *sample mean*) to emphasize that it is computed from a particular set of data. The sample average is an example of a **descriptive statistic**; in this case, the statistic describes the central tendency of the set of points x_i.

There are some basic properties about averages that are important to understand. First, suppose we take each observation on x and subtract off the average: $d_i \equiv x_i - \bar{x}$ (the "d" here stands for *deviation* from the average). Then, the sum of these deviations is always zero:

$$\sum_{i=1}^{n}d_i = \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x} = \sum_{i=1}^{n}x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

We summarize this as

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0. \qquad \text{[A.6]}$$

A simple numerical example shows how this works. Suppose $n = 5$ and $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$, and $x_5 = 5$. Then, $\bar{x} = 2$, and the demeaned sample is $\{4, -1, -4, -2, 3\}$. Adding these gives zero, which is just what equation (A.6) says.

In our treatment of regression analysis in Chapter 2, we need to know some additional algebraic facts involving deviations from sample averages. An important one is that the sum of squared deviations is the sum of the squared x_i minus n times the square of $\bar{x}$:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2. \qquad \text{[A.7]}$$

This can be shown using basic properties of the summation operator:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$
$$= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n(\bar{x})^2$$
$$= \sum_{i=1}^{n}x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2.$$

Given a data set on two variables, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, it can also be shown that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}x_i(y_i - \bar{y}) \qquad \text{[A.8]}$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}x_iy_i - n(\bar{x}\cdot\bar{y});$$

this is a generalization of equation (A.7). (There, $y_i = x_i$ for all i.)

The average is the measure of central tendency that we will focus on in most of this text. However, it is sometimes informative to use the **median** (or *sample median*) to describe the central value. To obtain the median of the n numbers $\{x_1, \ldots, x_n\}$, we first order the values of the x_i from smallest to largest. Then, if n is odd, the sample median is the middle number of the ordered observations. For example, given the numbers $\{-4, 8, 2, 0, 21, -10, 18\}$, the median value is 2 (because the ordered sequence is $\{-10, -4, 0, 2, 8, 18, 21\}$). If we change the largest number in this list, 21, to twice its value, 42, the median is still 2. By contrast, the sample average would increase from 5 to 8, a sizable change. Generally, the median is less sensitive than the average to changes in the extreme values (large or small) in a list of numbers. This is why "median incomes" or "median housing values" are often reported, rather than averages, when summarizing income or housing values in a city or county.

If n is even, there is no unique way to define the median because there are two numbers at the center. Usually, the median is defined to be the average of the two middle values (again, after ordering the numbers from smallest to largest). Using this rule, the median for the set of numbers $\{4, 12, 2, 6\}$ would be $(4 + 6)/2 = 5$.

A-2 Properties of Linear Functions

Linear functions play an important role in econometrics because they are simple to interpret and manipulate. If x and y are two variables related by

$$y = \beta_0 + \beta_1 x, \qquad \text{[A.9]}$$

then we say that y is a **linear function** of x, and β_0 and β_1 are two parameters (numbers) describing this relationship. The **intercept** is β_0, and the **slope** is β_1.

The defining feature of a linear function is that the change in y is always β_1 times the change in x:

$$\Delta y = \beta_1 \Delta x, \qquad \text{[A.10]}$$

where Δ denotes "change." In other words, the **marginal effect** of x on y is constant and equal to β_1.

EXAMPLE A.1 **Linear Housing Expenditure Function**

Suppose that the relationship between monthly housing expenditure and monthly income is

$$housing = 164 + .27\, income. \qquad [A.11]$$

Then, for each additional dollar of income, 27 cents is spent on housing. If family income increases by \$200, then housing expenditure increases by $(.27)200 = \$54$. This function is graphed in Figure A.1.

According to equation (A.11), a family with no income spends \$164 on housing, which of course cannot be literally true. For low levels of income, this linear function would not describe the relationship between *housing* and *income* very well, which is why we will eventually have to use other types of functions to describe such relationships.

In (A.11), the *marginal propensity to consume* (MPC) housing out of income is .27. This is different from the *average propensity to consume* (APC), which is

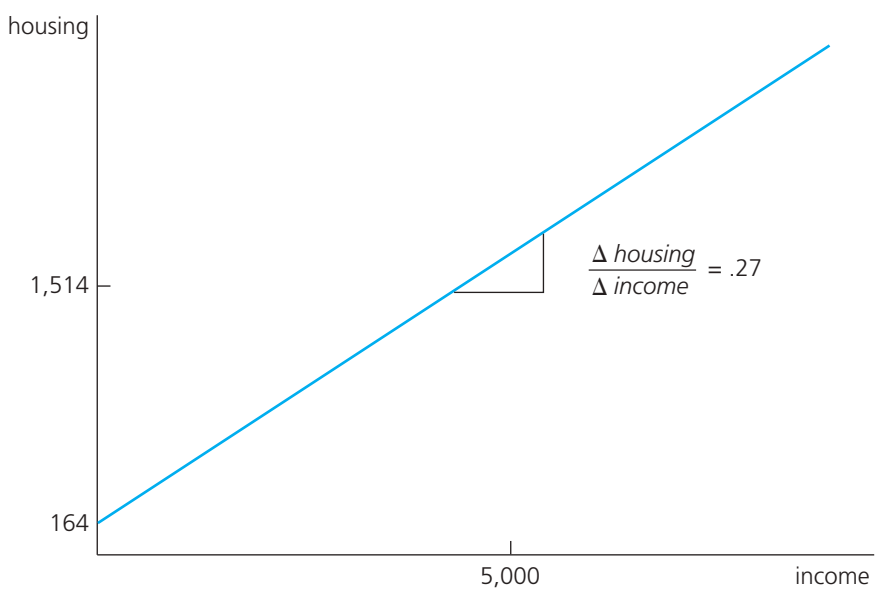
$$\frac{housing}{income} = 164/income + .27.$$

The APC is not constant; it is always larger than the MPC, and it gets closer to the MPC as income increases.

Linear functions are easily defined for more than two variables. Suppose that y is related to two variables, x_1 and x_2, in the general form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \qquad [A.12]$$

FIGURE A.1 Graph of *housing* = 164 + .27 *income*.



It is rather difficult to envision this function because its graph is three-dimensional. Nevertheless, β_0 is still the intercept (the value of y when $x_1 = 0$ and $x_2 = 0$), and β_1 and β_2 measure particular slopes. From (A.12), the change in y, for given changes in x_1 and x_2, is

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \qquad [A.13]$$

If x_2 does not change, that is, $\Delta x_2 = 0$, then we have

$$\Delta y = \beta_1 \Delta x_1 \text{ if } \Delta x_2 = 0,$$

so that β_1 is the slope of the relationship in the direction of x_1:

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ if } \Delta x_2 = 0.$$

Because it measures how y changes with x_1, holding x_2 fixed, β_1 is often called the **partial effect** of x_1 on y. Because the partial effect involves holding other factors fixed, it is closely linked to the notion of **ceteris paribus**. The parameter β_2 has a similar interpretation: $\beta_2 = \Delta y/\Delta x_2$ if $\Delta x_1 = 0$, so that β_2 is the partial effect of x_2 on y.

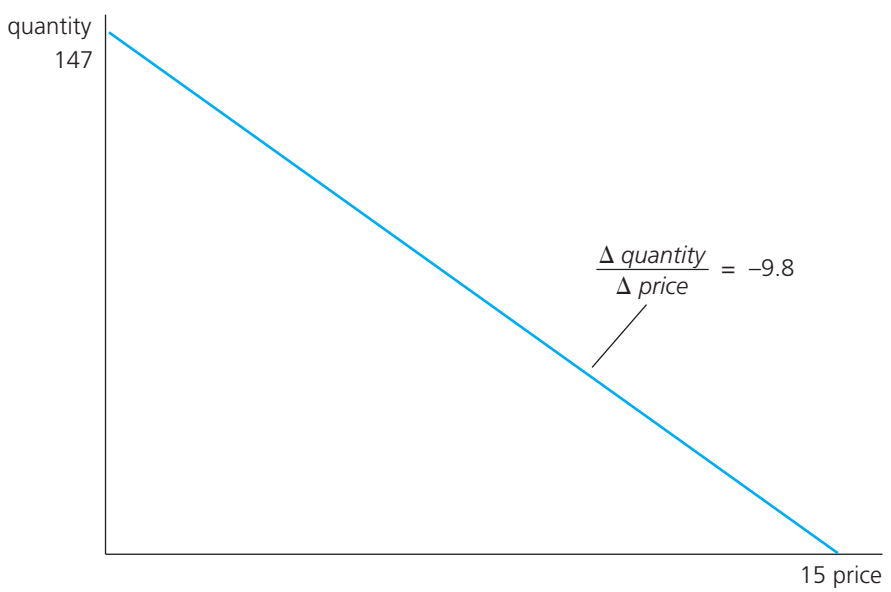
EXAMPLE A.2 Demand for Compact Discs

For college students, suppose that the monthly quantity demanded of compact discs is related to the price of compact discs and monthly discretionary income by

$$quantity = 120 - 9.8\, price + .03\, income,$$

where *price* is dollars per disc and *income* is measured in dollars. The *demand curve* is the relationship between *quantity* and *price*, holding *income* (and other factors) fixed. This is graphed in two dimensions in Figure A.2 at an income level of $900. The slope of the demand curve, -9.8, is the *partial effect* of price on quantity: holding income fixed, if the price of compact discs increases by one dollar, then the quantity demanded falls by 9.8. (We abstract from the fact that CDs can only be purchased in discrete units.) An increase in income simply shifts the demand curve up (changes the intercept), but the slope remains the same.

FIGURE A.2 Graph of *quantity* = 120 − 9.8 *price* + .03 *income*, with income fixed at $900.



A-3 Proportions and Percentages

Proportions and percentages play such an important role in applied economics that it is necessary to become very comfortable in working with them. Many quantities reported in the popular press are in the form of percentages; a few examples are interest rates, unemployment rates, and high school graduation rates.

An important skill is being able to convert proportions to percentages and vice versa. A percentage is easily obtained by multiplying a proportion by 100. For example, if the proportion of adults in a county with a high school degree is .82, then we say that 82% (82 percent) of adults have a high school degree. Another way to think of percentages and proportions is that a proportion is the decimal form of a percentage. For example, if the marginal tax rate for a family earning $30,000 per year is reported as 28%, then the proportion of the next dollar of income that is paid in income taxes is .28 (or 28¢).

When using percentages, we often need to convert them to decimal form. For example, if a state sales tax is 6% and $200 is spent on a taxable item, then the sales tax paid is 200(.06) = $12. If the annual return on a certificate of deposit (CD) is 7.6% and we invest $3,000 in such a CD at the beginning of the year, then our interest income is 3,000(.076) = $228. As much as we would like it, the interest income is not obtained by multiplying 3,000 by 7.6.

We must be wary of proportions that are sometimes incorrectly reported as percentages in the popular media. If we read, "The percentage of high school students who drink alcohol is .57," we know that this really means 57% (not just over one-half of a percent, as the statement literally implies). College volleyball fans are probably familiar with press clips containing statements such as "Her hitting percentage was .372." This really means that her hitting percentage was 37.2%.

In econometrics, we are often interested in measuring the *changes* in various quantities. Let x denote some variable, such as an individual's income, the number of crimes committed in a community, or the profits of a firm. Let x_0 and x_1 denote two values for x: x_0 is the initial value, and x_1 is the subsequent value. For example, x_0 could be the annual income of an individual in 1994 and x_1 the income of the same individual in 1995. The **proportionate change** in x in moving from x_0 to x_1, sometimes called the **relative change,** is simply

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \qquad \text{[A.14]}$$

assuming, of course, that $x_0 \neq 0$. In other words, to get the proportionate change, we simply divide the change in x by its initial value. This is a way of standardizing the change so that it is free of units. For example, if an individual's income goes from $30,000 per year to $36,000 per year, then the proportionate change is 6,000/30,000 = .20.

It is more common to state changes in terms of percentages. The **percentage change** in x in going from x_0 to x_1 is simply 100 times the proportionate change:

$$\%\Delta x = 100(\Delta x/x_0); \qquad \text{[A.15]}$$

the notation "$\%\Delta x$" is read as "the percentage change in x." For example, when income goes from $30,000 to $33,750, income has increased by 12.5%; to get this, we simply multiply the proportionate change, .125, by 100.

Again, we must be on guard for proportionate changes that are reported as percentage changes. In the previous example, for instance, reporting the percentage change in income as .125 is incorrect and could lead to confusion.

When we look at changes in things like dollar amounts or population, there is no ambiguity about what is meant by a percentage change. By contrast, interpreting percentage change calculations can be tricky when the variable of interest is itself a percentage, something that happens often in economics and other social sciences. To illustrate, let x denote the percentage of adults in a particular city having a college education. Suppose the initial value is $x_0 = 24$ (24% have a college education), and the new

value is $x_1 = 30$. We can compute two quantities to describe how the percentage of college-educated people has changed. The first is the change in x, Δx. In this case, $\Delta x = x_1 - x_0 = 6$: the percentage of people with a college education has increased by six *percentage points*. On the other hand, we can compute the percentage change in x using equation (A.15): $\%\Delta x = 100[(30 - 24)/24] = 25$.

In this example, the percentage point change and the percentage change are very different. The **percentage point change** is just the change in the percentages. The percentage change is the change relative to the initial value. Generally, we must pay close attention to which number is being computed. The careful researcher makes this distinction perfectly clear; unfortunately, in the popular press as well as in academic research, the type of reported change is often unclear.

EXAMPLE A.3 **Michigan Sales Tax Increase**

In March 1994, Michigan voters approved a sales tax increase from 4% to 6%. In political advertisements, supporters of the measure referred to this as a two percentage point increase, or an increase of two cents on the dollar. Opponents of the tax increase called it a 50% increase in the sales tax rate. Both claims are correct; they are simply different ways of measuring the increase in the sales tax. Naturally, each group reported the measure that made its position most favorable.

For a variable such as salary, it makes no sense to talk of a "percentage point change in salary" because salary is not measured as a percentage. We can describe a change in salary either in dollar or percentage terms.

A-4 Some Special Functions and Their Properties

In Section A-2, we reviewed the basic properties of linear functions. We already indicated one important feature of functions like $y = \beta_0 + \beta_1 x$: a one-unit change in x results in the *same* change in y, regardless of the initial value of x. As we noted earlier, this is the same as saying the marginal effect of x on y is constant, something that is not realistic for many economic relationships. For example, the important economic notion of *diminishing marginal returns* is not consistent with a linear relationship.

In order to model a variety of economic phenomena, we need to study several nonlinear functions. A **nonlinear function** is characterized by the fact that the change in y for a given change in x depends on the starting value of x. Certain nonlinear functions appear frequently in empirical economics, so it is important to know how to interpret them. A complete understanding of nonlinear functions takes us into the realm of calculus. Here, we simply summarize the most significant aspects of the functions, leaving the details of some derivations for Section A-5.

A-4a Quadratic Functions

One simple way to capture diminishing returns is to add a quadratic term to a linear relationship. Consider the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \qquad [A.16]$$

where β_0, β_1, and β_2 are parameters. When $\beta_1 > 0$ and $\beta_2 < 0$, the relationship between y and x has the parabolic shape given in Figure A.3, where $\beta_0 = 6$, $\beta_1 = 8$, and $\beta_2 = -2$.

When $\beta_1 > 0$ and $\beta_2 < 0$, it can be shown (using calculus in the next section) that the *maximum* of the function occurs at the point

$$x^* = \beta_1/(-2\beta_2). \qquad \text{[A.17]}$$

For example, if $y = 6 + 8x - 2x^2$ (so $\beta_1 = 8$ and $\beta_2 = -2$), then the largest value of y occurs at $x^* = 8/4 = 2$, and this value is $6 + 8(2) - 2(2)^2 = 14$ (see Figure A.3).

The fact that equation (A.16) implies a **diminishing marginal effect** of x on y is easily seen from its graph. Suppose we start at a low value of x and then increase x by some amount, say, c. This has a larger effect on y than if we start at a higher value of x and increase x by the same amount c. In fact, once $x > x^*$, an increase in x actually decreases y.

The statement that x has a diminishing marginal effect on y is the same as saying that the slope of the function in Figure A.3 decreases as x increases. Although this is clear from looking at the graph, we usually want to quantify how quickly the slope is changing. An application of calculus gives the approximate slope of the quadratic function as

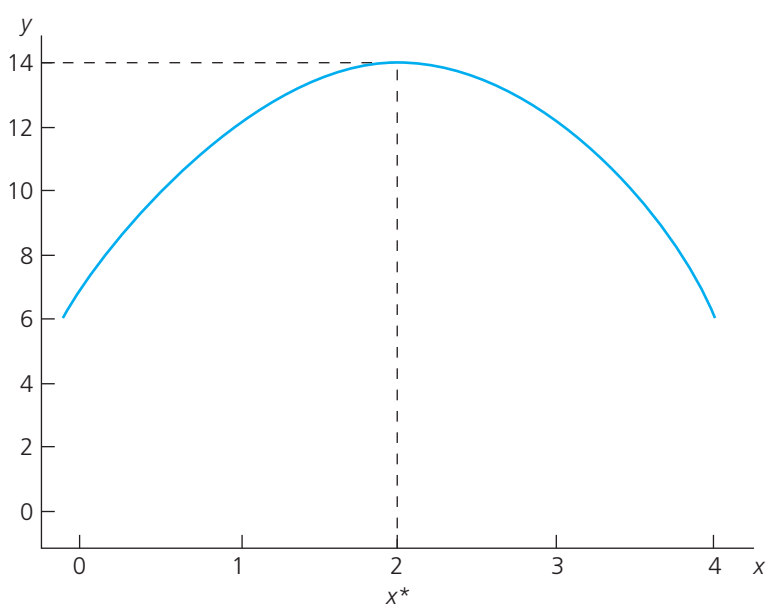
$$slope = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x, \qquad \text{[A.18]}$$

for "small" changes in x. [The right-hand side of equation (A.18) is the **derivative** of the function in equation (A.16) with respect to x.] Another way to write this is

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \text{ for "small" } \Delta x. \qquad \text{[A.19]}$$

To see how well this approximation works, consider again the function $y = 6 + 8x - 2x^2$. Then, according to equation (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Now, suppose we start at $x = 1$ and change x by $\Delta x = .1$. Using (A.19), $\Delta y \approx (8 - 4)(.1) = .4$. Of course, we can compute the change exactly by finding the values of y when $x = 1$ and $x = 1.1$: $y_0 = 6 + 8(1) - 2(1)^2 = 12$ and $y_1 = 6 + 8(1.1) - 2(1.1)^2 = 12.38$, so the exact change in y is .38. The approximation is pretty close in this case.

FIGURE A.3 Graph of $y = 6 + 8x - 2x^2$.



Now, suppose we start at $x = 1$ but change x by a larger amount: $\Delta x = .5$. Then, the approximation gives $\Delta y \approx 4(.5) = 2$. The exact change is determined by finding the difference in y when $x = 1$ and $x = 1.5$. The former value of y was 12, and the latter value is $6 + 8(1.5) - 2(1.5)^2 = 13.5$, so the actual change is 1.5 (not 2). The approximation is worse in this case because the change in x is larger.

For many applications, equation (A.19) can be used to compute the approximate marginal effect of x on y for any initial value of x and small changes. And, we can always compute the exact change if necessary.

EXAMPLE A.4 A Quadratic Wage Function

Suppose the relationship between hourly wages and years in the workforce (*exper*) is given by

$$wage = 5.25 + .48\, exper - .008\, exper^2. \qquad [A.20]$$

This function has the same general shape as the one in Figure A.3. Using equation (A.17), *exper* has a positive effect on wage up to the turning point, $exper^* = .48/[2(.008)] = 30$. The first year of experience is worth approximately .48, or 48 cents [see (A.19) with $x = 0$, $\Delta x = 1$]. Each additional year of experience increases wage by less than the previous year—reflecting a diminishing marginal return to experience. At 30 years, an additional year of experience would actually lower the wage. This is not very realistic, but it is one of the consequences of using a quadratic function to capture a diminishing marginal effect: at some point, the function must reach a maximum and curve downward. For practical purposes, the point at which this happens is often large enough to be inconsequential, but not always.

The graph of the quadratic function in (A.16) has a U-shape if $\beta_1 < 0$ and $\beta_2 > 0$, in which case there is an increasing marginal return. The minimum of the function is at the point $-\beta_1/(2\beta_2)$.

A-4b The Natural Logarithm

The nonlinear function that plays the most important role in econometric analysis is the **natural logarithm**. In this text, we denote the natural logarithm, which we often refer to simply as the **log function**, as

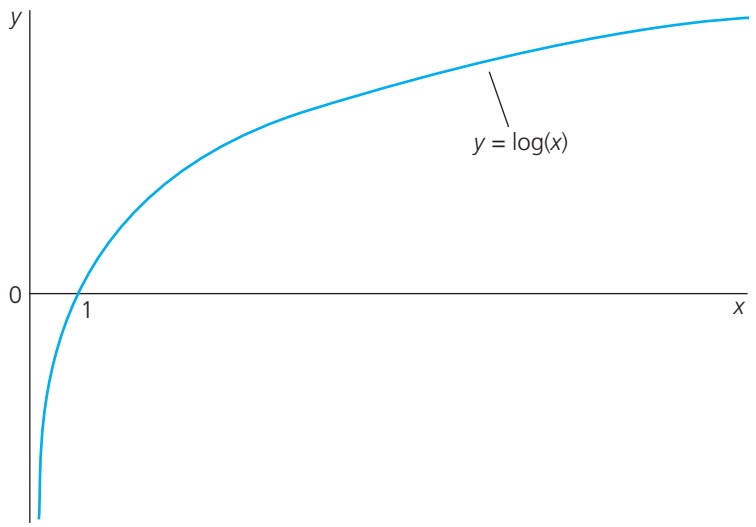
$$y = \log(x). \qquad [A.21]$$

You might remember learning different symbols for the natural log; $\ln(x)$ or $\log_e(x)$ are the most common. These different notations are useful when logarithms with several different bases are being used. For our purposes, only the natural logarithm is important, and so $\log(x)$ denotes the natural logarithm throughout this text. This corresponds to the notational usage in many statistical packages, although some use $\ln(x)$ [and most calculators use $\ln(x)$]. Economists use both $\log(x)$ and $\ln(x)$, which is useful to know when you are reading papers in applied economics.

The function $y = \log(x)$ is defined only for $x > 0$, and it is plotted in Figure A.4. It is not very important to know how the values of $\log(x)$ are obtained. For our purposes, the function can be thought of as a black box: we can plug in any $x > 0$ and obtain $\log(x)$ from a calculator or a computer.

Several things are apparent from Figure A.4. First, when $y = \log(x)$, the relationship between y and x displays diminishing marginal returns. One important difference between the log and the quadratic function in Figure A.3 is that when $y = \log(x)$, the effect of x on y never becomes negative: the

FIGURE A.4 Graph of $y = \log(x)$.



slope of the function gets closer and closer to zero as x gets large, but the slope never quite reaches zero and certainly never becomes negative.

The following are also apparent from Figure A.4:

$$\log(x) < 0 \text{ for } 0 < x < 1$$
$$\log(1) = 0$$
$$\log(x) > 0 \text{ for } x > 1.$$

In particular, $\log(x)$ can be positive or negative. Some useful algebraic facts about the log function are

$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2), x_1, x_2 > 0$$
$$\log(x_1/x_2) = \log(x_1) - \log(x_2), x_1, x_2 > 0$$
$$\log(x^c) = c \log(x), x > 0, c \text{ any number.}$$

Occasionally, we will need to rely on these properties.

The logarithm can be used for various approximations that arise in econometric applications. First, $\log(1 + x) \approx x$ for $x \approx 0$. You can try this with $x = .02$, .1, and .5 to see how the quality of the approximation deteriorates as x gets larger. Even more useful is the fact that the difference in logs can be used to approximate proportionate changes. Let x_0 and x_1 be positive values. Then, it can be shown (using calculus) that

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \qquad \text{[A.22]}$$

for small changes in x. If we multiply equation (A.22) by 100 and write $\Delta\log(x) = \log(x_1) - \log(x_0)$, then

$$100 \cdot \Delta\log(x) \approx \%\Delta x \qquad \text{[A.23]}$$

for small changes in x. The meaning of "small" depends on the context, and we will encounter several examples throughout this text.

Why should we approximate the percentage change using (A.23) when the exact percentage change is so easy to compute? Momentarily, we will see why the approximation in (A.23) is useful in econometrics. First, let us see how good the approximation is in two examples.

First, suppose $x_0 = 40$ and $x_1 = 41$. Then, the percentage change in x in moving from x_0 to x_1 is 2.5%, using $100(x_1 - x_0)/x_0$. Now, $\log(41) - \log(40) = .0247$ (to four decimal places), which when multiplied by 100 is very close to 2.5. The approximation works pretty well. Now, consider a much bigger change: $x_0 = 40$ and $x_1 = 60$. The exact percentage change is 50%. However, $\log(60) - \log(40) \approx .4055$, so the approximation gives 40.55%, which is much farther off.

Why is the approximation in (A.23) useful if it is only satisfactory for small changes? To build up to the answer, we first define the **elasticity** of y with respect to x as

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}. \qquad \text{[A.24]}$$

In other words, the elasticity of y with respect to x is the percentage change in y when x increases by 1%. This notion should be familiar from introductory economics.

If y is a linear function of x, $y = \beta_0 + \beta_1 x$, then the elasticity is

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \qquad \text{[A.25]}$$

which clearly depends on the value of x. (This is a generalization of the well-known result from basic demand theory: the elasticity is not constant along a straight-line demand curve.)

Elasticities are of critical importance in many areas of applied economics, not just in demand theory. It is convenient in many situations to have *constant* elasticity models, and the log function allows us to specify such models. If we use the approximation in (A.23) for both x and y, then the elasticity is approximately equal to $\Delta\log(y)/\Delta\log(x)$. Thus, a constant elasticity model is approximated by the equation

$$\log(y) = \beta_0 + \beta_1 \log(x), \qquad \text{[A.26]}$$

and β_1 is the elasticity of y with respect to x (assuming that $x, y > 0$).

EXAMPLE A.5 **Constant Elasticity Demand Function**

If q is quantity demanded and p is price and these variables are related by

$$\log(q) = 4.7 - 1.25 \log(p),$$

then the price elasticity of demand is -1.25. Roughly, a 1% increase in price leads to a 1.25% fall in the quantity demanded.

For our purposes, the fact that β_1 in (A.26) is only close to the elasticity is not important. In fact, when the elasticity is defined using calculus—as in Section A-5—the definition is exact. For the purposes of econometric analysis, (A.26) defines a **constant elasticity model**. Such models play a large role in empirical economics.

Other possibilities for using the log function often arise in empirical work. Suppose that $y > 0$ and

$$\log(y) = \beta_0 + \beta_1 x. \qquad \text{[A.27]}$$

Then, $\Delta\log(y) = \beta_1\Delta x$, so $100\cdot\Delta\log(y) = (100\cdot\beta_1)\Delta x$. It follows that, when y and x are related by equation (A.27),

$$\%\Delta y \approx (100 \cdot \beta_1)\Delta x. \qquad [A.28]$$

EXAMPLE A.6 **Logarithmic Wage Equation**

Suppose that hourly wage and years of education are related by

$$\log(wage) = 2.78 + .094\, educ.$$

Then, using equation (A.28),

$$\%\Delta wage \approx 100(.094)\Delta educ = 9.4\ \Delta educ.$$

It follows that one more year of education increases hourly wage by about 9.4%.

Generally, the quantity $\%\Delta y/\Delta x$ is called the **semi-elasticity** of y with respect to x. The semi-elasticity is the percentage change in y when x increases by one *unit*. What we have just shown is that, in model (A.27), the semi-elasticity is constant and equal to $100\cdot\beta_1$. In Example A.6, we can conveniently summarize the relationship between wages and education by saying that one more year of education—starting from any amount of education—increases the wage by about 9.4%. This is why such models play an important role in economics.

Another relationship of some interest in applied economics is

$$y = \beta_0 + \beta_1\log(x), \qquad [A.29]$$

where $x > 0$. How can we interpret this equation? If we take the change in y, we get $\Delta y = \beta_1\Delta\log(x)$, which can be rewritten as $\Delta y = (\beta_1/100)[100\cdot\Delta\log(x)]$. Thus, using the approximation in (A.23), we have

$$\Delta y \approx (\beta_1/100)(\%\Delta x). \qquad [A.30]$$

In other words, $\beta_1/100$ is the unit change in y when x increases by 1%.

EXAMPLE A.7 **Labor Supply Function**

Assume that the labor supply of a worker can be described by

$$hours = 33 + 45.1\log(wage),$$

where *wage* is hourly wage and *hours* is hours worked per week. Then, from (A.30),

$$\Delta hours \approx (45.1/100)(\%\Delta wage) = .451\ \%\Delta wage.$$

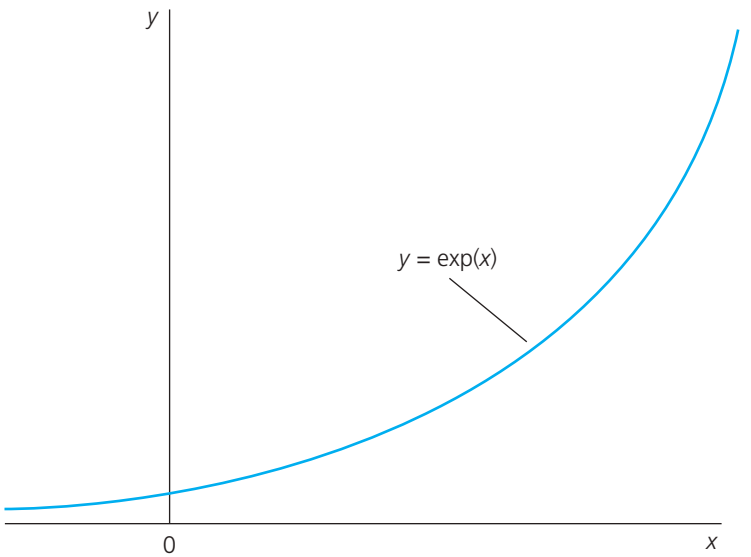
In other words, a 1% increase in *wage* increases the weekly hours worked by about .45, or slightly less than one-half hour. If the wage increases by 10%, then $\Delta hours = .451(10) = 4.51$, or about four and one-half hours. We would not want to use this approximation for much larger percentage changes in wages.

A-4c The Exponential Function

Before leaving this section, we need to discuss a special function that is related to the log. As motivation, consider equation (A.27). There, $\log(y)$ is a linear function of x. But how do we find y itself as a function of x? The answer is given by the **exponential function**.

We will write the exponential function as $y = \exp(x)$, which is graphed in Figure A.5. From Figure A.5, we see that $\exp(x)$ is defined for any value of x and is always greater than zero. Sometimes,

FIGURE A.5 Graph of $y = \exp(x)$.



the exponential function is written as $y = e^x$, but we will not use this notation. Two important values of the exponential function are $\exp(0) = 1$ and $\exp(1) = 2.7183$ (to four decimal places).

The exponential function is the inverse of the log function in the following sense: $\log[\exp(x)] = x$ for all x, and $\exp[\log(x)] = x$ for $x > 0$. In other words, the log "undoes" the exponential, and vice versa. (This is why the exponential function is sometimes called the *anti-log* function.) In particular, note that $\log(y) = \beta_0 + \beta_1 x$ is equivalent to

$$y = \exp(\beta_0 + \beta_1 x).$$

If $\beta_1 > 0$, the relationship between x and y has the same shape as in Figure A.5. Thus, if $\log(y) = \beta_0 + \beta_1 x$ with $\beta_1 > 0$, then x has an *increasing* marginal effect on y. In Example A.6, this means that another year of education leads to a larger change in wage than the previous year of education.

Two useful facts about the exponential function are $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ and $\exp[c \cdot \log(x)] = x^c$.

A-5 Differential Calculus

In the previous section, we asserted several approximations that have foundations in calculus. Let $y = f(x)$ for some function f. Then, for small changes in x,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \qquad \text{[A.31]}$$

where df/dx is the derivative of the function f, evaluated at the initial point x_0. We also write the derivative as dy/dx.

For example, if $y = \log(x)$, then $dy/dx = 1/x$. Using (A.31), with dy/dx evaluated at x_0, we have $\Delta y \approx (1/x_0)\Delta x$, or $\Delta\log(x) \approx \Delta x/x_0$, which is the approximation given in (A.22).

In applying econometrics, it helps to recall the derivatives of a handful of functions because we use the derivative to define the slope of a function at a given point. We can then use (A.31) to find the approximate change in y for small changes in x. In the linear case, the derivative is simply the slope of the line, as we would hope: if $y = \beta_0 + \beta_1 x$, then $dy/dx = \beta_1$.

If $y = x^c$, then $dy/dx = cx^{c-1}$. The derivative of a sum of two functions is the sum of the derivatives: $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. The derivative of a constant times any function is that same constant times the derivative of the function: $d[cf(x)]/dx = c[df(x)/dx]$. These simple rules allow us to find derivatives of more complicated functions. Other rules, such as the product, quotient, and chain rules, will be familiar to those who have taken calculus, but we will not review those here.

Some functions that are often used in economics, along with their derivatives, are

$$y = \beta_0 + \beta_1 x + \beta_2 x^2;\ dy/dx = \beta_1 + 2\beta_2 x$$
$$y = \beta_0 + \beta_1/x;\ dy/dx = -\beta_1/(x^2)$$
$$y = \beta_0 + \beta_1\sqrt{x};\ dy/dx = (\beta_1/2)x^{-1/2}$$
$$y = \beta_0 + \beta_1\log(x);\ dy/dx = \beta_1/x$$
$$y = \exp(\beta_0 + \beta_1 x);\ dy/dx = \beta_1\exp(\beta_0 + \beta_1 x).$$

If $\beta_0 = 0$ and $\beta_1 = 1$ in this last expression, we get $dy/dx = \exp(x)$, when $y = \exp(x)$.

In Section A-4, we noted that equation (A.26) defines a constant elasticity model when calculus is used. The calculus definition of elasticity is $(dy/dx)\cdot(x/y)$. It can be shown using properties of logs and exponentials that, when (A.26) holds, $(dy/dx)\cdot(x/y) = \beta_1$.

When y is a function of multiple variables, the notion of a **partial derivative** becomes important. Suppose that

$$y = f(x_1, x_2). \qquad \text{[A.32]}$$

Then, there are two partial derivatives, one with respect to x_1 and one with respect to x_2. The partial derivative of y with respect to x_1, denoted here by $\partial y/\partial x_1$, is just the usual derivative of (A.32) with respect to x_1, where x_2 is treated as a *constant*. Similarly, $\partial y/\partial x_2$ is just the derivative of (A.32) with respect to x_2, holding x_1 fixed.

Partial derivatives are useful for much the same reason as ordinary derivatives. We can approximate the change in y as

$$\Delta y \approx \frac{\partial y}{\partial x_1}\cdot\Delta x_1, \text{ holding } x_2 \text{ fixed.} \qquad \text{[A.33]}$$

Thus, calculus allows us to define partial effects in nonlinear models just as we could in linear models. In fact, if

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

then

$$\frac{\partial y}{\partial x_1} = \beta_1, \frac{\partial y}{\partial x_2} = \beta_2.$$

These can be recognized as the partial effects defined in Section A-2.

A more complicated example is

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1\cdot x_2. \qquad \text{[A.34]}$$

Now, the derivative of (A.34), with respect to x_1 (treating x_2 as a constant), is simply

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2;$$

note how this depends on x_1 and x_2. The derivative of (A.34), with respect to x_2, is $\partial y/\partial x_2 = -3 + 7x_1$, so this depends only on x_1.

EXAMPLE A.8 Wage Function with Interaction

A function relating wages to years of education and experience is

$$wage = 3.10 + .41\ educ + .19\ exper - .004\ exper^2 + .007\ educ \cdot exper. \qquad [A.35]$$

The partial effect of *exper* on *wage* is the partial derivative of (A.35):

$$\frac{\partial wage}{\partial exper} = .19 - .008\ exper + .007\ educ.$$

This is the approximate change in wage due to increasing experience by one year. Notice that this partial effect depends on the initial level of *exper* and *educ*. For example, for a worker who is starting with $educ = 12$ and $exper = 5$, the next year of experience increases wage by about $.19 - .008(5) + .007(12) = .234$, or 23.4 cents per hour. The exact change can be calculated by computing (A.35) at $exper = 5$, $educ = 12$ and at $exper = 6$, $educ = 12$, and then taking the difference. This turns out to be .23, which is very close to the approximation.

Differential calculus plays an important role in minimizing and maximizing functions of one or more variables. If $f(x_1, x_2, \ldots, x_k)$ is a differentiable function of k variables, then a necessary condition for $x_1^*, x_2^*, \ldots, x_k^*$ to either minimize or maximize f over all possible values of x_j is

$$\frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \ldots, x_k^*) = 0, j = 1, 2, \ldots, k. \qquad [A.36]$$

In other words, all of the partial derivatives of f must be zero when they are evaluated at the x_h*. These are called the *first order conditions* for minimizing or maximizing a function. Practically, we hope to solve equation (A.36) for the x_h*. Then, we can use other criteria to determine whether we have minimized or maximized the function. We will not need those here. [See Sydsaeter and Hammond (1995) for a discussion of multivariable calculus and its use in optimizing functions.]

Summary

The math tools reviewed here are crucial for understanding regression analysis and the probability and statistics that are covered in Appendices B and C. The material on nonlinear functions—especially quadratic, logarithmic, and exponential functions—is critical for understanding modern applied economic research. The level of comprehension required of these functions does not include a deep knowledge of calculus, although calculus is needed for certain derivations.

Key Terms

Average
Ceteris Paribus
Constant Elasticity Model
Derivative
Descriptive Statistic
Diminishing Marginal Effect
Elasticity
Exponential Function
Intercept
Linear Function
Log Function
Marginal Effect
Median
Natural Logarithm
Nonlinear Function
Partial Derivative
Partial Effect
Percentage Change
Percentage Point Change
Proportionate Change
Relative Change
Semi-Elasticity
Slope
Summation Operator

Problems

1 The following table contains monthly housing expenditures for 10 families.

Family	Monthly Housing Expenditures (Dollars)
1	300
2	440
3	350
4	1,100
5	640
6	480
7	450
8	700
9	670
10	530

(i) Find the average monthly housing expenditure.
(ii) Find the median monthly housing expenditure.
(iii) If monthly housing expenditures were measured in hundreds of dollars, rather than in dollars, what would be the average and median expenditures?
(iv) Suppose that family number 8 increases its monthly housing expenditure to $900, but the expenditures of all other families remain the same. Compute the average and median housing expenditures.

2 Suppose the following equation describes the relationship between the average number of classes missed during a semester (*missed*) and the distance from school (*distance*, measured in miles):

$$missed = 3 + 0.2\, distance.$$

(i) Sketch this line, being sure to label the axes. How do you interpret the intercept in this equation?
(ii) What is the average number of classes missed for someone who lives five miles away?
(iii) What is the difference in the average number of classes missed for someone who lives 10 miles away and someone who lives 20 miles away?

3 In Example A.2, quantity of compact discs was related to price and income by $quantity = 120 - 9.8\,price + .03\,income$. What is the demand for CDs if $price = 15$ and $income = 200$? What does this suggest about using linear functions to describe demand curves?

4 Suppose the unemployment rate in the United States goes from 6.4% in one year to 5.6% in the next.
(i) What is the percentage point decrease in the unemployment rate?
(ii) By what percentage has the unemployment rate fallen?

5 Suppose that the return from holding a particular firm's stock goes from 15% in one year to 18% in the following year. The majority shareholder claims that "the stock return only increased by 3%," while the chief executive officer claims that "the return on the firm's stock increased by 20%." Reconcile their disagreement.

6 Suppose that Person A earns $35,000 per year and Person B earns $42,000.
(i) Find the exact percentage by which Person B's salary exceeds Person A's.
(ii) Now, use the difference in natural logs to find the approximate percentage difference.

7 Suppose the following model describes the relationship between annual salary (*salary*) and the number of previous years of labor market experience (*exper*):

$$\log(salary) = 10.6 + .027\,exper.$$

(i) What is *salary* when $exper = 0$? When $exper = 5$? (*Hint*: You will need to exponentiate.)
(ii) Use equation (A.28) to approximate the percentage increase in *salary* when *exper* increases by five years.
(iii) Use the results of part (i) to compute the exact percentage difference in salary when $exper = 5$ and $exper = 0$. Comment on how this compares with the approximation in part (ii).

8 Let *grthemp* denote the proportionate growth in employment, at the county level, from 1990 to 1995, and let *salestax* denote the county sales tax rate, stated as a proportion. Interpret the intercept and slope in the equation

$$grthemp = .043 - .78\,salestax.$$

9 Suppose the yield of a certain crop (in bushels per acre) is related to fertilizer amount (in pounds per acre) as

$$yield = 120 + .19\sqrt{fertilizer}.$$

(i) Graph this relationship by plugging in several values for *fertilizer*.
(ii) Describe how the shape of this relationship compares with a linear relationship between *yield* and *fertilizer*.

10 Suppose that in a particular state a standardized test is given to all graduating seniors. Let *score* denote a student's score on the test. Someone discovers that performance on the test is related to the size of the student's graduating high school class. The relationship is quadratic:

$$score = 45.6 + .082\,class - .000147\,class^2,$$

where *class* is the number of students in the graduating class.
(i) How do you literally interpret the value 45.6 in the equation? By itself, is it of much interest? Explain.
(ii) From the equation, what is the optimal size of the graduating class (the size that maximizes the test score)? (Round your answer to the nearest integer.) What is the highest achievable test score?

(iii) Sketch a graph that illustrates your solution in part (ii).
(iv) Does it seem likely that *score* and *class* would have a deterministic relationship? That is, is it realistic to think that once you know the size of a student's graduating class you know, with certainty, his or her test score? Explain.

11 Consider the line

$$y = \beta_0 + \beta_1 x.$$

(i) Let (x_1, y_1) and (x_2, y_2) be two points on the line. Show that $(\bar{x}, \bar{y})$ is also on the line, where $\bar{x} = (x_1 + x_2)/2$ is the average of the two values and $\bar{y} = (y_1 + y_2)/2$.
(ii) Extend the result of part (i) to n points on the line, $\{(x_i, y_i): i = 1, \ldots, n\}$.

12 (i) Let $\{x_i: i = 1, 2, \ldots, n\}$ be a set of n data points, and let $\bar{x}$ be the average. Suppose that the units i are divided into two groups of sizes n_1 and n_2, with $n_1 + n_2 = n$. Without loss of generality, order the observations as

$$\{x_1, x_2, \ldots, x_{n1}, x_{n1+1}, x_{n1+2}, \ldots, x_n\},$$

so that the data points for the first group appear first. Let

$$\bar{x}_1 = n_1^{-1} \sum_{i=1}^{n_1} x_i, \quad \bar{x}_2 = n_2^{-1} \sum_{i=n_1+1}^{n} x_i$$

be the averages for the two groups. Show that

$$\bar{x} = \left(\frac{n_1}{n}\right)\bar{x}_1 + \left(\frac{n_2}{n}\right)\bar{x}_2 = w_1\bar{x}_1 + w_2\bar{x}_2,$$

so that $\bar{x}$ can be expressed as a weighted average of the averages from the two subgroups.
(ii) Do the weights w_1 and w_2 in part (i) make intuitive sense? Explain.
(iii) How does the finding in part (i) extend the case of g groups, where the group sizes are $n_1, n_2, \ldots, n_g$?

13 (i) Let $\{x_i: i = 1, 2, \ldots, n\}$ be a set of n data points with $x_i > 0$ for all i. Is it always true that

$$\sum_{i=1}^{n} \frac{1}{x_i} = \frac{1}{\sum_{i=1}^{n} x_i}?$$

(ii) Is the equality in part (i) always true if $x_i = c$ for all i,where $c > 0$?

Math Refresher B

Fundamentals of Probability

This Math Refresher covers key concepts from basic probability. Appendices B and C are primarily for review; they are not intended to replace a course in probability and statistics. However, all of the probability and statistics concepts that we use in the text are covered in these appendices.

Probability is of interest in its own right for students in business, economics, and other social sciences. For example, consider the problem of an airline trying to decide how many reservations to accept for a flight that has 100 available seats. If fewer than 100 people want reservations, then these should all be accepted. But what if more than 100 people request reservations? A safe solution is to accept at most 100 reservations. However, because some people book reservations and then do not show up for the flight, there is some chance that the plane will not be full even if 100 reservations are booked. This results in lost revenue to the airline. A different strategy is to book more than 100 reservations and to hope that some people do not show up, so the final number of passengers is as close to 100 as possible. This policy runs the risk of the airline having to compensate people who are necessarily bumped from an overbooked flight.

A natural question in this context is: Can we decide on the optimal (or best) number of reservations the airline should make? This is a nontrivial problem. Nevertheless, given certain information (on airline costs and how frequently people show up for reservations), we can use basic probability to arrive at a solution.

B-1 Random Variables and Their Probability Distributions

Suppose that we flip a coin 10 times and count the number of times the coin turns up heads. This is an example of an **experiment**. Generally, an experiment is any procedure that can, at least in theory, be infinitely repeated and has a well-defined set of outcomes. We could, in principle, carry out the coin-flipping procedure again and again. Before we flip the coin, we know that the number of heads appearing is an integer from 0 to 10, so the outcomes of the experiment are well defined.

A **random variable** is one that takes on numerical values and has an outcome that is determined by an experiment. In the coin-flipping example, the number of heads appearing in 10 flips of a coin is an example of a random variable. Before we flip the coin 10 times, we do not know how many

times the coin will come up heads. Once we flip the coin 10 times and count the number of heads, we obtain the outcome of the random variable for this particular trial of the experiment. Another trial can produce a different outcome.

In the airline reservation example mentioned earlier, the number of people showing up for their flight is a random variable: before any particular flight, we do not know how many people will show up.

To analyze data collected in business and the social sciences, it is important to have a basic understanding of random variables and their properties. Following the usual conventions in probability and statistics throughout Appendices B and C, we denote random variables by uppercase letters, usually W, X, Y, and Z; particular outcomes of random variables are denoted by the corresponding lowercase letters, w, x, y, and z. For example, in the coin-flipping experiment, let X denote the number of heads appearing in 10 flips of a coin. Then, X is not associated with any particular value, but we know X will take on a value in the set $\{0, 1, 2, \ldots, 10\}$. A particular outcome is, say, $x = 6$.

We indicate large collections of random variables by using subscripts. For example, if we record last year's income of 20 randomly chosen households in the United States, we might denote these random variables by $X_1, X_2, \ldots, X_{20}$; the particular outcomes would be denoted $x_1, x_2, \ldots, x_{20}$.

As stated in the definition, random variables are always defined to take on numerical values, even when they describe qualitative events. For example, consider tossing a single coin, where the two outcomes are heads and tails. We can define a random variable as follows: $X = 1$ if the coin turns up heads, and $X = 0$ if the coin turns up tails.

A random variable that can only take on the values zero and one is called a **Bernoulli** (or **binary**) **random variable**. In basic probability, it is traditional to call the event $X = 1$ a "success" and the event $X = 0$ a "failure." For a particular application, the success-failure nomenclature might not correspond to our notion of a success or failure, but it is a useful terminology that we will adopt.

B-1a Discrete Random Variables

A **discrete random variable** is one that takes on only a finite or countably infinite number of values. The notion of "countably infinite" means that even though an infinite number of values can be taken on by a random variable, those values can be put in a one-to-one correspondence with the positive integers. Because the distinction between "countably infinite" and "uncountably infinite" is somewhat subtle, we will concentrate on discrete random variables that take on only a finite number of values. Larsen and Marx (1986, Chapter 3) provide a detailed treatment.

A Bernoulli random variable is the simplest example of a discrete random variable. The only thing we need to completely describe the behavior of a Bernoulli random variable is the probability that it takes on the value one. In the coin-flipping example, if the coin is "fair," then $P(X = 1) = 1/2$ (read as "the probability that X equals one is one-half"). Because probabilities must sum to one, $P(X = 0) = 1/2$, also.

Social scientists are interested in more than flipping coins, so we must allow for more general situations. Again, consider the example where the airline must decide how many people to book for a flight with 100 available seats. This problem can be analyzed in the context of several Bernoulli random variables as follows: for a randomly selected customer, define a Bernoulli random variable as $X = 1$ if the person shows up for the reservation, and $X = 0$ if not.

There is no reason to think that the probability of any particular customer showing up is 1/2; in principle, the probability can be any number between 0 and 1. Call this number θ, so that

$$P(X = 1) = \theta \qquad \text{[B.1]}$$

$$P(X = 0) = 1 - \theta. \qquad \text{[B.2]}$$

For example, if $\theta = .75$, then there is a 75% chance that a customer shows up after making a reservation and a 25% chance that the customer does not show up. Intuitively, the value of θ is crucial in determining the airline's strategy for booking reservations. Methods for *estimating* θ, given historical data on airline reservations, are a subject of mathematical statistics, something we turn to in Math Refresher C.

More generally, any discrete random variable is completely described by listing its possible values and the associated probability that it takes on each value. If X takes on the k possible values $\{x_1, \ldots, x_k\}$, then the probabilities $p_1, p_2, \ldots, p_k$ are defined by

$$p_j = \mathrm{P}(X = x_j), j = 1, 2, \ldots, k, \qquad \text{[B.3]}$$

where each p_j is between 0 and 1 and

$$p_1 + p_2 + \cdots + p_k = 1. \qquad \text{[B.4]}$$

Equation (B.3) is read as: "The probability that X takes on the value x_j is equal to p_j."

Equations (B.1) and (B.2) show that the probabilities of success and failure for a Bernoulli random variable are determined entirely by the value of θ. Because Bernoulli random variables are so prevalent, we have a special notation for them: $X \sim \text{Bernoulli}(\theta)$ is read as "X has a Bernoulli distribution with probability of success equal to θ."

The **probability density function (pdf)** of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, j = 1, 2, \ldots, k, \qquad \text{[B.5]}$$

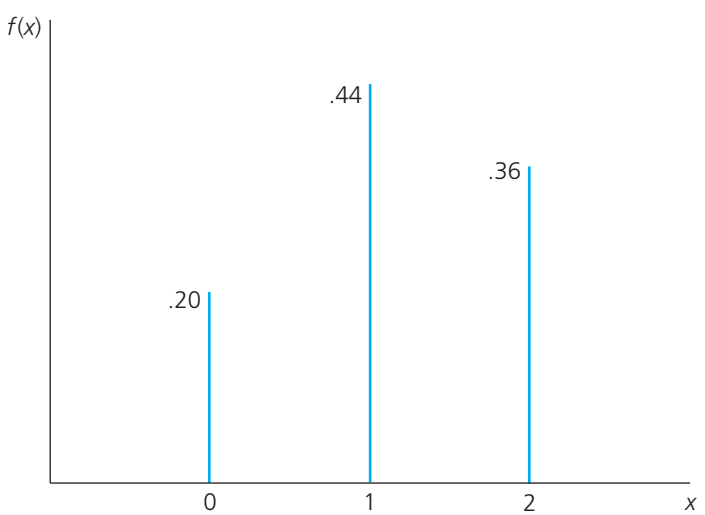
with $f(x) = 0$ for any x not equal to x_j for some j. In other words, for any real number x, $f(x)$ is the probability that the random variable X takes on the particular value x. When dealing with more than one random variable, it is sometimes useful to subscript the pdf in question: f_X is the pdf of X, f_Y is the pdf of Y, and so on.

Given the pdf of any discrete random variable, it is simple to compute the probability of any event involving that random variable. For example, suppose that X is the number of free throws made by a basketball player out of two attempts, so that X can take on the three values $\{0, 1, 2\}$. Assume that the pdf of X is given by

$$f(0) = .20, f(1) = .44, \text{ and } f(2) = .36.$$

The three probabilities sum to one, as they must. Using this pdf, we can calculate the probability that the player makes *at least* one free throw: $\mathrm{P}(X \geq 1) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = .44 + .36 = .80$. The pdf of X is shown in Figure B.1.

FIGURE B.1 The pdf of the number of free throws made out of two attempts.



B-1b Continuous Random Variables

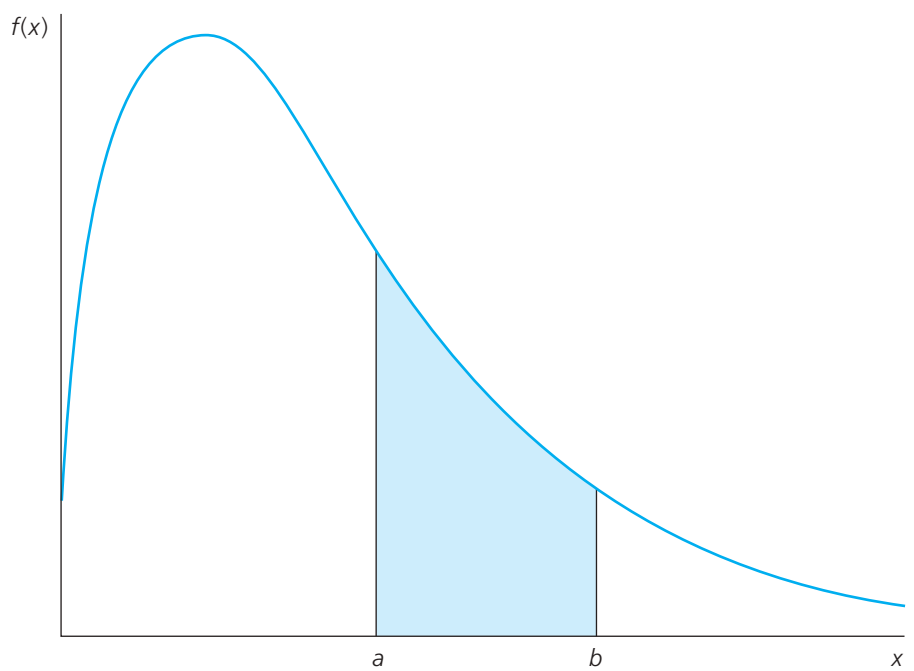
A variable X is a **continuous random variable** if it takes on any real value with *zero* probability. This definition is somewhat counterintuitive because in any application we eventually observe some outcome for a random variable. The idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero. While measurements are always discrete in practice, random variables that take on numerous values are best treated as continuous. For example, the most refined measure of the price of a good is in terms of cents. We can imagine listing all possible values of price in order (even though the list may continue indefinitely), which technically makes price a discrete random variable. However, there are so many possible values of price that using the mechanics of discrete random variables is not feasible.

We can define a probability density function for continuous random variables, and, as with discrete random variables, the pdf provides information on the likely outcomes of the random variable. However, because it makes no sense to discuss the probability that a continuous random variable takes on a particular value, we use the pdf of a continuous random variable only to compute events involving a range of values. For example, if a and b are constants where $a < b$, the probability that X lies between the numbers a and b, $\mathrm{P}(a \leq X \leq b)$, is the *area* under the pdf between points a and b, as shown in Figure B.2. If you are familiar with calculus, you recognize this as the *integral* of the function f between the points a and b. The entire area under the pdf must always equal one.

When computing probabilities for continuous random variables, it is easiest to work with the **cumulative distribution function (cdf)**. If X is any random variable, then its cdf is defined for any real number x by

$$F(x) \equiv \mathrm{P}(X \leq x). \qquad \text{[B.6]}$$

FIGURE B.2 The probability that *X* lies between the points *a* and *b*.



For discrete random variables, (B.6) is obtained by summing the pdf over all values x_j such that $x_j \leq x$. For a continuous random variable, $F(x)$ is the area under the pdf, f, to the left of the point x. Because $F(x)$ is simply a probability, it is always between 0 and 1. Further, if $x_1 < x_2$, then $P(X \leq x_1) \leq P(X \leq x_2)$, that is, $F(x_1) \leq F(x_2)$. This means that a cdf is an increasing (or at least a nondecreasing) function of x.

Two important properties of cdfs that are useful for computing probabilities are the following:

$$\text{For any number } c,\ P(X > c) = 1 - F(c). \qquad \text{[B.7]}$$

$$\text{For any numbers } a < b,\ P(a < X \leq b) = F(b) - F(a). \qquad \text{[B.8]}$$

In our study of econometrics, we will use cdfs to compute probabilities only for continuous random variables, in which case it does not matter whether inequalities in probability statements are strict or not. That is, for a continuous random variable X,

$$P(X \geq c) = P(X > c), \qquad \text{[B.9]}$$

and

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \qquad \text{[B.10]}$$

Combined with (B.7) and (B.8), equations (B.9) and (B.10) greatly expand the probability calculations that can be done using continuous cdfs.

Cumulative distribution functions have been tabulated for all of the important continuous distributions in probability and statistics. The most well known of these is the normal distribution, which we cover along with some related distributions in Section B-5.

B-2 Joint Distributions, Conditional Distributions, and Independence

In economics, we are usually interested in the occurrence of events involving more than one random variable. For example, in the airline reservation example referred to earlier, the airline might be interested in the probability that a person who makes a reservation shows up *and* is a business traveler; this is an example of a *joint probability*. Or, the airline might be interested in the following *conditional probability*: conditional on the person being a business traveler, what is the probability of his or her showing up? In the next two subsections, we formalize the notions of joint and conditional distributions and the important notion of *independence* of random variables.

B-2a Joint Distributions and Independence

Let X and Y be discrete random variables. Then, (X, Y) have a **joint distribution**, which is fully described by the *joint probability density function* of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y), \qquad \text{[B.11]}$$

where the right-hand side is the probability that $X = x$ and $Y = y$. When X and Y are continuous, a joint pdf can also be defined, but we will not cover such details because joint pdfs for continuous random variables are not used explicitly in this text.

In one case, it is easy to obtain the joint pdf if we are given the pdfs of X and Y. In particular, random variables X and Y are said to be independent if, and only if,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{[B.12]}$$

for all x and y, where f_X is the pdf of X and f_Y is the pdf of Y. In the context of more than one random variable, the pdfs f_X and f_Y are often called *marginal probability density functions* to distinguish them from the joint pdf $f_{X,Y}$. This definition of independence is valid for discrete and continuous random variables.

To understand the meaning of (B.12), it is easiest to deal with the discrete case. If X and Y are discrete, then (B.12) is the same as

$$P(X = x, Y = y) = P(X = x)P(Y = y); \qquad [B.13]$$

in other words, the probability that $X = x$ and $Y = y$ is the product of the two probabilities $P(X = x)$ and $P(Y = y)$. One implication of (B.13) is that joint probabilities are fairly easy to compute, because they only require knowledge of $P(X = x)$ and $P(Y = y)$.

If random variables are not independent, then they are said to be *dependent*.

EXAMPLE B.1 **Free Throw Shooting**

Consider a basketball player shooting two free throws. Let X be the Bernoulli random variable equal to one if she or he makes the first free throw, and zero otherwise. Let Y be a Bernoulli random variable equal to one if he or she makes the second free throw. Suppose that she or he is an 80% free throw shooter, so that $P(X = 1) = P(Y = 1) = .8$. What is the probability of the player making both free throws?

If X and Y are independent, we can easily answer this question: $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = (.8)(.8) = .64$. Thus, there is a 64% chance of making both free throws. If the chance of making the second free throw depends on whether the first was made—that is, X and Y are not independent—then this simple calculation is not valid.

Independence of random variables is a very important concept. In the next subsection, we will show that if X and Y are independent, then knowing the outcome of X does not change the probabilities of the possible outcomes of Y, and vice versa. One useful fact about independence is that if X and Y are independent and we define new random variables $g(X)$ and $h(Y)$ for any functions g and h, then these new random variables are also independent.

There is no need to stop at two random variables. If $X_1, X_2, \ldots, X_n$ are discrete random variables, then their joint pdf is $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are **independent random variables** if, and only if, their joint pdf is the product of the individual pdfs for any $(x_1, x_2, \ldots, x_n)$. This definition of independence also holds for continuous random variables.

The notion of independence plays an important role in obtaining some of the classic distributions in probability and statistics. Earlier, we defined a Bernoulli random variable as a zero-one random variable indicating whether or not some event occurs. Often, we are interested in the number of successes in a sequence of *independent* Bernoulli trials. A standard example of independent Bernoulli trials is flipping a coin again and again. Because the outcome on any particular flip has nothing to do with the outcomes on other flips, independence is an appropriate assumption.

Independence is often a reasonable approximation in more complicated situations. In the airline reservation example, suppose that the airline accepts n reservations for a particular flight. For each $i = 1, 2, \ldots, n$, let Y_i denote the Bernoulli random variable indicating whether customer i shows up: $Y_i = 1$ if customer i appears, and $Y_i = 0$ otherwise. Letting θ again denote the probability of success (using reservation), each Y_i has a Bernoulli(θ) distribution. As an approximation, we might assume that the Y_i are independent of one another, although this is not exactly true in reality: some people travel in groups, which means that whether or not a person shows up is not truly independent of whether all others show up. Modeling this kind of dependence is complex, however, so we might be willing to use independence as an approximation.

The variable of primary interest is the total number of customers showing up out of the n reservations; call this variable X. Because each Y_i is unity when a person shows up, we can write

$X = Y_1 + Y_2 + \cdots + Y_n$. Now, assuming that each Y_i has probability of success θ and that the Y_i are independent, X can be shown to have a **binomial distribution**. That is, the probability density function of X is

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, 2, \ldots, n, \qquad \text{[B.14]}$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$, and for any integer n, $n!$ (read "n factorial") is defined as $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$. By convention, $0! = 1$. When a random variable X has the pdf given in (B.14), we write $X \sim \text{Binomial}(n, \theta)$. Equation (B.14) can be used to compute $P(X = x)$ for any value of x from 0 to n.

If the flight has 100 available seats, the airline is interested in $P(X > 100)$. Suppose, initially, that $n = 120$, so that the airline accepts 120 reservations, and the probability that each person shows up is $\theta = .85$. Then, $P(X > 100) = P(X = 101) + P(X = 102) + \cdots + P(X = 120)$, and each of the probabilities in the sum can be found from equation (B.14) with $n = 120$, $\theta = .85$, and the appropriate value of x (101 to 120). This is a difficult hand calculation, but many statistical packages have commands for computing this kind of probability. In this case, the probability that more than 100 people will show up is about .659, which is probably more risk of overbooking than the airline wants to tolerate. If, instead, the number of reservations is 110, the probability of more than 100 passengers showing up is only about .024.

B-2b Conditional Distributions

In econometrics, we are usually interested in how one random variable, call it Y, is related to one or more other variables. For now, suppose that there is only one variable whose effects we are interested in, call it X. The most we can know about how X affects Y is contained in the **conditional distribution** of Y given X. This information is summarized by the *conditional probability density function*, defined by

$$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x) \qquad \text{[B.15]}$$

for all values of x such that $f_X(x) > 0$. The interpretation of (B.15) is most easily seen when X and Y are discrete. Then,

$$f_{Y|X}(y|x) = P(Y = y|X = x), \qquad \text{[B.16]}$$

where the right-hand side is read as "the probability that $Y = y$ given that $X = x$." When Y is continuous, $f_{Y|X}(y|x)$ is not interpretable directly as a probability, for the reasons discussed earlier, but conditional probabilities are found by computing areas under the conditional pdf.

An important feature of conditional distributions is that, if X and Y are independent random variables, knowledge of the value taken on by X tells us nothing about the probability that Y takes on various values (and vice versa). That is, $f_{Y|X}(y|x) = f_Y(y)$, and $f_{X|Y}(x|y) = f_X(x)$.

EXAMPLE B.2 **Free Throw Shooting**

Consider again the basketball-shooting example, where two free throws are to be attempted. Assume that the conditional density is

$$f_{Y|X}(1|1) = .85, f_{Y|X}(0|1) = .15$$
$$f_{Y|X}(1|0) = .70, f_{Y|X}(0|0) = .30.$$

This means that the probability of the player making the second free throw depends on whether the first free throw was made: if the first free throw is made, the chance of making the second is .85; if the

first free throw is missed, the chance of making the second is .70. This implies that X and Y are *not* independent; they are dependent.

We can still compute $P(X = 1, Y = 1)$ provided we know $P(X = 1)$. Assume that the probability of making the first free throw is .8, that is, $P(X = 1) = .8$. Then, from (B.15), we have

$$P(X = 1, Y = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = (.85)(.8) = .68.$$

B-3 Features of Probability Distributions

For many purposes, we will be interested in only a few aspects of the distributions of random variables. The features of interest can be put into three categories: measures of central tendency, measures of variability or spread, and measures of association between two random variables. We cover the last of these in Section B-4.

B-3a A Measure of Central Tendency: The Expected Value

The expected value is one of the most important probabilistic concepts that we will encounter in our study of econometrics. If X is a random variable, the **expected value** (or expectation) of X, denoted $E(X)$ and sometimes μX or simply μ, is a weighted average of all possible values of X. The weights are determined by the probability density function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that X represents some variable in a population.

The precise definition of expected value is simplest in the case that X is a discrete random variable taking on a finite number of values, say, $\{x_1, \ldots, x_k\}$. Let $f(x)$ denote the probability density function of X. The expected value of X is the weighted average

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \qquad [B.17]$$

This is easily computed given the values of the pdf at each possible outcome of X.

EXAMPLE B.3 Computing an Expected Value

Suppose that X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8, respectively. Then,

$$E(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

This example illustrates something curious about expected values: the expected value of X can be a number that is not even a possible outcome of X. We know that X takes on the values -1, 0, or 2, yet its expected value is 5/8. This makes the expected value deficient for summarizing the central tendency of certain discrete random variables, but calculations such as those just mentioned can be useful, as we will see later.

If X is a continuous random variable, then $E(X)$ is defined as an integral:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \qquad [B.18]$$

which we assume is well defined. This can still be interpreted as a weighted average. For the most common continuous distributions, $E(X)$ is a number that is a possible outcome of X. In this text, we will not need to compute expected values using integration, although we will draw on some well-known results from probability for expected values of special random variables.

Given a random variable X and a function $g(\cdot)$, we can create a new random variable $g(X)$. For example, if X is a random variable, then so is X^2 and $\log(X)$(if $X > 0$). The expected value of $g(X)$ is, again, simply a weighted average:

$$\mathrm{E}[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j) \qquad [B.19]$$

or, for a continuous random variable,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \qquad [B.20]$$

EXAMPLE B.4 Expected Value of X^2

For the random variable in Example B.3, let $g(X) = X^2$. Then,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

In Example B.3, we computed $\mathrm{E}(X) = 5/8$, so that $[\mathrm{E}(X)]^2 = 25/64$. This shows that $\mathrm{E}(X^2)$ is *not* the same as $[\mathrm{E}(X)]^2$. In fact, for a nonlinear function $g(X)$, $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (except in very special cases).

If X and Y are random variables, then $g(X, Y)$ is a random variable for any function g, and so we can define its expectation. When X and Y are both discrete, taking on values $\{x_1, x_2, \ldots, x_k\}$ and $\{y_1, y_2, \ldots, y_m\}$, respectively, the expected value is

$$\mathrm{E}[g(X, Y)] = \sum_{h=1}^{k} \sum_{j=1}^{m} g(x_h, y_j) f_{X,Y}(x_h, y_j),$$

where $f_{X,Y}$ is the joint pdf of (X, Y). The definition is more complicated for continuous random variables because it involves integration; we do not need it here. The extension to more than two random variables is straightforward.

B-3b Properties of Expected Values

In econometrics, we are not so concerned with computing expected values from various distributions; the major calculations have been done many times, and we will largely take these on faith. We will need to manipulate some expected values using a few simple rules. These are so important that we give them labels:

Property E.1: For any constant c, $\mathrm{E}(c) = c$.

Property E.2: For any constants a and b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

One useful implication of E.2 is that, if $\mu = \mathrm{E}(X)$, and we define a new random variable as $Y = X - \mu$, then $\mathrm{E}(Y) = 0$; in E.2, take $a = 1$ and $b = -\mu$.

As an example of Property E.2, let X be the temperature measured in Celsius at noon on a particular day at a given location; suppose the expected temperature is $\mathrm{E}(X) = 25$. If Y is the temperature measured in Fahrenheit, then $Y = 32 + (9/5)X$. From Property E.2, the expected temperature in Fahrenheit is $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

Generally, it is easy to compute the expected value of a linear function of many random variables.

Property E.3: If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\mathrm{E}(X_1) + a_2\mathrm{E}(X_2) + \cdots + a_n\mathrm{E}(X_n).$$

Or, using summation notation,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i\mathrm{E}(X_i). \qquad \text{[B.21]}$$

As a special case of this, we have (with each $a_i = 1$)

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i), \qquad \text{[B.22]}$$

so that the expected value of the sum is the sum of expected values. This property is used often for derivations in mathematical statistics.

EXAMPLE B.5 **Finding Expected Revenue**

Let X_1, X_2, and X_3 be the numbers of small, medium, and large pizzas, respectively, sold during the day at a pizza parlor. These are random variables with expected values $\mathrm{E}(X_1) = 25$, $\mathrm{E}(X_2) = 57$, and $\mathrm{E}(X_3) = 40$. The prices of small, medium, and large pizzas are \$5.50, \$7.60, and \$9.15. Therefore, the expected revenue from pizza sales on a given day is

$$\begin{aligned}\mathrm{E}(5.50\,X_1 + 7.60\,X_2 + 9.15\,X_3) &= 5.50\,\mathrm{E}(X_1) + 7.60\,\mathrm{E}(X_2) + 9.15\,\mathrm{E}(X_3)\\ &= 5.50(25) + 7.60(57) + 9.15(40) = 936.70,\end{aligned}$$

that is, \$936.70. The actual revenue on any particular day will generally differ from this value, but this is the *expected* revenue.

We can also use Property E.3 to show that if $X \sim \text{Binomial}(n, \theta)$, then $\mathrm{E}(X) = n\theta$. That is, the expected number of successes in n Bernoulli trials is simply the number of trials times the probability of success on any particular trial. This is easily seen by writing X as $X = Y_1 + Y_2 + \cdots + Y_n$, where each $Y_i \sim \text{Bernoulli}(\theta)$. Then,

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

We can apply this to the airline reservation example, where the airline makes $n = 120$ reservations, and the probability of showing up is $\theta = .85$. The *expected* number of people showing up is $120(.85) = 102$. Therefore, if there are 100 seats available, the expected number of people showing up is too large; this has some bearing on whether it is a good idea for the airline to make 120 reservations.

Actually, what the airline should do is define a profit function that accounts for the net revenue earned per seat sold and the cost per passenger bumped from the flight. This profit function is random because the actual number of people showing up is random. Let r be the net revenue from each passenger. (You can think of this as the price of the ticket for simplicity.) Let c be the compensation owed to any passenger bumped from the flight. Neither r nor c is random; these are assumed to be known to the airline. Let Y denote profits for the flight. Then, with 100 seats available,

$$\begin{aligned}Y &= rX \text{ if } X \leq 100\\ &= 100r - c(X - 100) \text{ if } X > 100.\end{aligned}$$

The first equation gives profit if no more than 100 people show up for the flight; the second equation is profit if more than 100 people show up. (In the latter case, the net revenue from ticket sales is $100r$, because all 100 seats are sold, and then $c(X - 100)$ is the cost of making more than 100 reservations.) Using the fact that X has a Binomial(n,.85) distribution, where n is the number of reservations made, expected profits, E(Y), can be found as a function of n (and r and c). Computing E(Y) directly would be quite difficult, but it can be found quickly using a computer. Once values for r and c are given, the value of n that maximizes expected profits can be found by searching over different values of n.

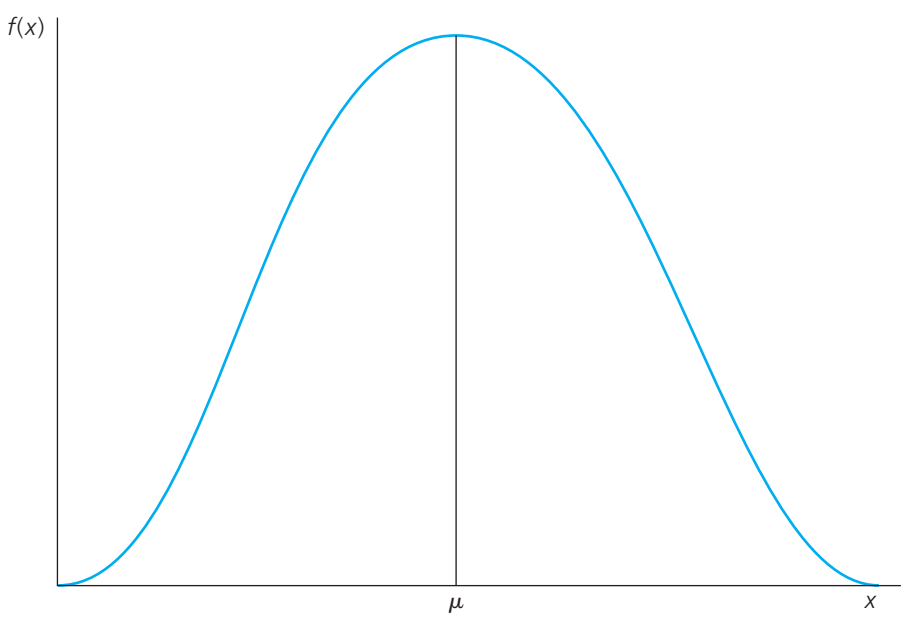
B-3c Another Measure of Central Tendency: The Median

The expected value is only one possibility for defining the central tendency of a random variable. Another measure of central tendency is the **median**. A general definition of *median* is too complicated for our purposes. If X is continuous, then the median of X, say, m, is the value such that one-half of the area under the pdf is to the left of m, and one-half of the area is to the right of m.

When X is discrete and takes on a finite number of odd values, the median is obtained by ordering the possible values of X and then selecting the value in the middle. For example, if X can take on the values $\{-4, 0, 2, 8, 10, 13, 17\}$, then the median value of X is 8. If X takes on an even number of values, there are really two median values; sometimes, these are averaged to get a unique median value. Thus, if X takes on the values $\{-5, 3, 9, 17\}$, then the median values are 3 and 9; if we average these, we get a median equal to 6.

In general, the median, sometimes denoted Med(X), and the expected value, E(X), are different. Neither is "better" than the other as a measure of central tendency; they are both valid ways to measure the center of the distribution of X. In one special case, the median and expected value (or mean) are the same. If X has a **symmetric distribution** about the value μ, then μ is both the expected value and the median. Mathematically, the condition is $f(\mu + x) = f(\mu - x)$ for all x. This case is illustrated in Figure B.3.

FIGURE B.3 A symmetric probability distribution.



B-3d Measures of Variability: Variance and Standard Deviation

Although the central tendency of a random variable is valuable, it does not tell us everything we want to know about the distribution of a random variable. Figure B.4 shows the pdfs of two random variables with the same mean. Clearly, the distribution of X is more tightly centered about its mean than is the distribution of Y. We would like to have a simple way of summarizing differences in the spreads of distributions.

B-3e Variance

For a random variable X, let $\mu = \mathrm{E}(X)$. There are various ways to measure how far X is from its expected value, but the simplest one to work with algebraically is the squared difference, $(X - \mu)^2$. (The squaring eliminates the sign from the distance measure; the resulting positive value corresponds to our intuitive notion of distance and treats values above and below μ symmetrically.) This distance is itself a random variable because it can change with every outcome of X. Just as we needed a number to summarize the central tendency of X, we need a number that tells us how far X is from μ, *on average*. One such number is the **variance**, which tells us the expected distance from X to its mean:

$$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mu)^2]. \qquad \text{[B.23]}$$

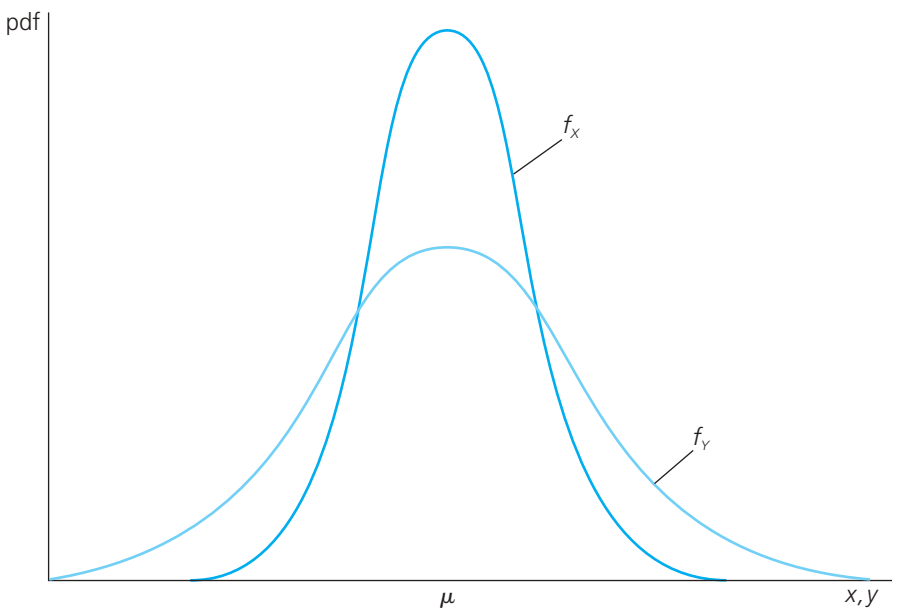
Variance is sometimes denoted σ_X^2, or simply σ^2, when the context is clear. From (B.23), it follows that the variance is always nonnegative.

As a computational device, it is useful to observe that

$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \qquad \text{[B.24]}$$

In using either (B.23) or (B.24), we need not distinguish between discrete and continuous random variables: the definition of variance is the same in either case. Most often, we first compute $\mathrm{E}(X)$, then $\mathrm{E}(X^2)$, and then we use the formula in (B.24). For example, if $X \sim \text{Bernoulli}(\theta)$, then $\mathrm{E}(X) = \theta$, and, because $X^2 = X$, $\mathrm{E}(X^2) = \theta$. It follows from equation (B.24) that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

FIGURE B.4 Random variables with the same mean but different distributions.



Two important properties of the variance follow.

Property VAR.1: $\text{Var}(X) = 0$ if, and only if, there is a constant c such that $P(X = c) = 1$, in which case $\text{E}(X) = c$.

This first property says that the variance of any constant is zero and if a random variable has zero variance, then it is essentially constant.

Property VAR.2: For any constants a and b, $\text{Var}(aX + b) = a^2\text{Var}(X)$.

This means that adding a constant to a random variable does not change the variance, but multiplying a random variable by a constant increases the variance by a factor equal to the *square* of that constant. For example, if X denotes temperature in Celsius and $Y = 32 + (9/5)X$ is temperature in Fahrenheit, then $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

B-3f Standard Deviation

The **standard deviation** of a random variable, denoted $\text{sd}(X)$, is simply the positive square root of the variance: $\text{sd}(X) \equiv +\sqrt{\text{Var}(X)}$. The standard deviation is sometimes denoted σ_X, or simply σ, when the random variable is understood. Two standard deviation properties immediately follow from Properties VAR.1 and VAR.2.

Property SD.1: For any constant c, $\text{sd}(c) = 0$.

Property SD.2: For any constants a and b,

$$\text{sd}(aX + b) = |a|\text{sd}(X).$$

In particular, if $a > 0$, then $\text{sd}(aX) = a \cdot \text{sd}(X)$.

This last property makes the standard deviation more natural to work with than the variance. For example, suppose that X is a random variable measured in thousands of dollars, say, income. If we define $Y = 1{,}000X$, then Y is income measured in dollars. Suppose that $\text{E}(X) = 20$, and $\text{sd}(X) = 6$. Then, $\text{E}(Y) = 1{,}000\text{E}(X) = 20{,}000$, and $\text{sd}(Y) = 1{,}000 \cdot \text{sd}(X) = 6{,}000$, so that the expected value and standard deviation both increase by the same factor, 1,000. If we worked with variance, we would have $\text{Var}(Y) = (1{,}000)^2\text{Var}(X)$, so that the variance of Y is one million times larger than the variance of X.

B-3g Standardizing a Random Variable

As an application of the properties of variance and standard deviation—and a topic of practical interest in its own right—suppose that given a random variable X, we define a new random variable by subtracting off its mean m and dividing by its standard deviation σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \text{[B.25]}$$

which we can write as $Z = aX + b$, where $a \equiv (1/\sigma)$ and $b \equiv -(\mu/\sigma)$. Then, from Property E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

From Property VAR.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Thus, the random variable Z has a mean of zero and a variance (and therefore a standard deviation) equal to one. This procedure is sometimes known as *standardizing* the random variable X, and Z is called a **standardized random variable**. (In introductory statistics courses, it is sometimes called the *z-transform* of X.) It is important to remember that the standard deviation, not the variance, appears in the denominator of (B.25). As we will see, this transformation is frequently used in statistical inference.

As a specific example, suppose that $E(X) = 2$, and $Var(X) = 9$. Then, $Z = (X - 2)/3$ has expected value zero and variance one.

B-3h Skewness and Kurtosis

We can use the standardized version of a random variable to define other features of the distribution of a random variable. These features are described by using what are called *higher order moments*. For example, the third moment of the random variable Z in (B.25) is used to determine whether a distribution is symmetric about its mean. We can write

$$E(Z^3) = E[(X - \mu)^3]/\sigma^3.$$

If X has a symmetric distribution about μ, then Z has a symmetric distribution about zero. (The division by σ^3 does not change whether the distribution is symmetric.) That means the density of Z at any two points z and $-z$ is the same, which means that, in computing $E(Z^3)$, positive values z^3 when $z > 0$ are exactly offset with the negative value $(-z)^3 = -z^3$. It follows that, if X is symmetric about zero, then $E(Z) = 0$. Generally, $E[(X - \mu)^3]/\sigma^3$ is viewed as a measure of **skewness** in the distribution of X. In a statistical setting, we might use data to estimate $E(Z^3)$ to determine whether an underlying population distribution appears to be symmetric. (Computer Exercise C4 in Chapter 5 provides an illustration.)

It also can be informative to compute the fourth moment of Z,

$$E(Z^4) = E[(X - \mu)^4]/\sigma^4.$$

Because $Z^4 \geq 0$, $E(Z^4) \geq 0$ (and, in any interesting case, strictly greater than zero). Without having a reference value, it is difficult to interpret values of $E(Z^4)$, but larger values mean that the tails in the distribution of X are thicker. The fourth moment $E(Z^4)$ is called a measure of **kurtosis** in the distribution of X. In Section B-5, we will obtain $E(Z^4)$ for the normal distribution.

B-4 Features of Joint and Conditional Distributions

B-4a Measures of Association: Covariance and Correlation

While the joint pdf of two random variables completely describes the relationship between them, it is useful to have summary measures of how, on average, two random variables vary with one another. As with the expected value and variance, this is similar to using a single number to summarize something about an entire distribution, which in this case is a joint distribution of two random variables.

B-4b Covariance

Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$ and consider the random variable $(X - \mu_X)(Y - \mu_Y)$. Now, if X is above its mean and Y is above its mean, then $(X - \mu_X)(Y - \mu_Y) > 0$. This is also true if $X < \mu_X$ and $Y < \mu_Y$. On the other hand, if $X > \mu_X$ and $Y < \mu_Y$, or vice versa, then $(X - \mu_X)(Y - \mu_Y) < 0$. How, then, can this product tell us anything about the relationship between X and Y?

The **covariance** between two random variables X and Y, sometimes called the *population covariance* to emphasize that it concerns the relationship between two variables describing a population, is defined as the expected value of the product $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X, Y) \equiv \text{E}[(X - \mu_X)(Y - \mu_Y)], \qquad \text{[B.26]}$$

which is sometimes denoted σ_{XY}. If $\sigma_{XY} > 0$, then, on average, when X is above its mean, Y is also above its mean. If $\sigma_{XY} < 0$, then, on average, when X is above its mean, Y is below its mean.

Several expressions useful for computing $\text{Cov}(X, Y)$ are as follows:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}[(X - \mu_X)Y] \\ &= \text{E}[X(Y - \mu_Y)] = \text{E}(XY) - \mu_X\mu_Y. \qquad \text{[B.27]}\end{aligned}$$

It follows from (B.27), that if $\text{E}(X) = 0$ or $\text{E}(Y) = 0$, then $\text{Cov}(X, Y) = \text{E}(XY)$.

Covariance measures the amount of *linear* dependence between two random variables. A positive covariance indicates that two random variables move in the same direction, while a negative covariance indicates they move in opposite directions. Interpreting the *magnitude* of a covariance can be a little tricky, as we will see shortly.

Because covariance is a measure of how two random variables are related, it is natural to ask how covariance is related to the notion of independence. This is given by the following property.

Property COV.1: If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

This property follows from equation (B.27) and the fact that $\text{E}(XY) = \text{E}(X)\text{E}(Y)$ when X and Y are independent. It is important to remember that the converse of COV.1 is *not* true: zero covariance between X and Y does not imply that X and Y are independent. In fact, there are random variables X such that, if $Y = X^2$, $\text{Cov}(X, Y) = 0$. [Any random variable with $\text{E}(X) = 0$ and $\text{E}(X^3) = 0$ has this property.] If $Y = X^2$, then X and Y are clearly not independent: once we know X, we know Y. It seems rather strange that X and X^2 could have zero covariance, and this reveals a weakness of covariance as a general measure of association between random variables. The covariance is useful in contexts when relationships are at least approximately linear.

The second major property of covariance involves covariances between linear functions.

Property COV.2: For any constants a_1, b_1, a_2, and b_2,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y). \qquad \text{[B.28]}$$

An important implication of COV.2 is that the covariance between two random variables can be altered simply by multiplying one or both of the random variables by a constant. This is important in economics because monetary variables, inflation rates, and so on can be defined with different units of measurement without changing their meaning.

Finally, it is useful to know that the absolute value of the covariance between any two random variables is bounded by the product of their standard deviations; this is known as the *Cauchy-Schwartz inequality*.

Property COV.3: $|\text{Cov}(X, Y)| \leq \text{sd}(X)\text{sd}(Y)$.

B-4c Correlation Coefficient

Suppose we want to know the relationship between amount of education and annual earnings in the working population. We could let X denote education and Y denote earnings and then compute their covariance. But the answer we get will depend on how we choose to measure education and

earnings. Property COV.2 implies that the covariance between education and earnings depends on whether earnings are measured in dollars or thousands of dollars, or whether education is measured in months or years. It is pretty clear that how we measure these variables has no bearing on how strongly they are related. But the covariance between them does depend on the units of measurement.

The fact that the covariance depends on units of measurement is a deficiency that is overcome by the **correlation coefficient** between X and Y:

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{sd}(X)\cdot\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \qquad \text{[B.29]}$$

the correlation coefficient between X and Y is sometimes denoted ρ_{XY} (and is sometimes called the *population correlation*).

Because σ_X and σ_Y are positive, $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ always have the same sign, and $\text{Corr}(X, Y) = 0$ if, and only if, $\text{Cov}(X, Y) = 0$. Some of the properties of covariance carry over to correlation. If X and Y are independent, then $\text{Corr}(X, Y) = 0$, but zero correlation does not imply independence. (Like the covariance, the correlation coefficient is also a measure of linear dependence.) However, the magnitude of the correlation coefficient is easier to interpret than the size of the covariance due to the following property.

Property CORR.1: $-1 \leq \text{Corr}(X, Y) \leq 1$.

If $\text{Corr}(X, Y) = 0$, or equivalently $\text{Cov}(X, Y) = 0$, then there is no linear relationship between X and Y, and X and Y are said to be **uncorrelated random variables**; otherwise, X and Y are *correlated.* $\text{Corr}(X, Y) = 1$ implies a perfect positive linear relationship, which means that we can write $Y = a + bX$ for some constant a and some constant $b > 0$. $\text{Corr}(X, Y) = -1$ implies a perfect negative linear relationship, so that $Y = a + bX$ for some $b < 0$. The extreme cases of positive or negative 1 rarely occur. Values of ρ_{XY} closer to 1 or -1 indicate stronger linear relationships.

As mentioned earlier, the correlation between X and Y is invariant to the units of measurement of either X or Y. This is stated more generally as follows.

Property CORR.2: For constants a_1, b_1, a_2, and b_2, with $a_1a_2 > 0$,

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = \text{Corr}(X, Y).$$

If $a_1a_2 < 0$, then

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = -\text{Corr}(X, Y).$$

As an example, suppose that the correlation between earnings and education in the working population is .15. This measure does not depend on whether earnings are measured in dollars, thousands of dollars, or any other unit; it also does not depend on whether education is measured in years, quarters, months, and so on.

B-4d Variance of Sums of Random Variables

Now that we have defined covariance and correlation, we can complete our list of major properties of the variance.

Property VAR.3: For constants a and b,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y).$$

It follows immediately that, if X and Y are uncorrelated—so that $\text{Cov}(X, Y) = 0$—then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \qquad \text{[B.30]}$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \qquad \text{[B.31]}$$

In the latter case, note how the variance of the difference is the *sum of the variances*, not the difference in the variances.

As an example of (B.30), let X denote profits earned by a restaurant during a Friday night and let Y be profits earned on the following Saturday night. Then, $Z = X + Y$ is profits for the two nights. Suppose X and Y each have an expected value of \$300 and a standard deviation of \$15 (so that the variance is 225). Expected profits for the two nights is $\text{E}(Z) = \text{E}(X) + \text{E}(Y) = 2 \cdot (300) = 600$ dollars. If X and Y are independent, and therefore uncorrelated, then the variance of total profits is the sum of the variances: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2 \cdot (225) = 450$. It follows that the standard deviation of total profits is $\sqrt{450}$ or about \$21.21.

Expressions (B.30) and (B.31) extend to more than two random variables. To state this extension, we need a definition. The random variables $\{X_1, \ldots, X_n\}$ are **pairwise uncorrelated random variables** if each variable in the set is uncorrelated with every other variable in the set. That is, $\text{Cov}(X_i, X_j) = 0$, for all $i \neq j$.

Property VAR.4: If $\{X_1, \ldots, X_n\}$ are pairwise uncorrelated random variables and a_i: $i = 1, \ldots, n$ are constants, then

$$\text{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n).$$

In summation notation, we can write

$$\text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(X_i). \qquad \text{[B.32]}$$

A special case of Property VAR.4 occurs when we take $a_i = 1$ for all i. Then, for pairwise uncorrelated random variables, the variance of the sum is the sum of the variances:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \qquad \text{[B.33]}$$

Because independent random variables are uncorrelated (see Property COV.1), the variance of a sum of independent random variables is the sum of the variances.

If the X_i are not pairwise uncorrelated, then the expression for $\text{Var}(\sum_{i=1}^{n} a_iX_i)$ is much more complicated; we must add to the right-hand side of (B.32) the terms $2a_ia_j\text{Cov}(x_i, x_j)$ for all $i > j$.

We can use (B.33) to derive the variance for a binomial random variable. Let $X \sim \text{Binomial}(n, \theta)$ and write $X = Y_1 + \cdots + Y_n$, where the Y_i are independent Bernoulli (θ) random variables. Then, by (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \cdots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

In the airline reservation example with $n = 120$ and $\theta = .85$, the variance of the number of passengers arriving for their reservations is $120(.85)(.15) = 15.3$, so the standard deviation is about 3.9.

B-4e Conditional Expectation

Covariance and correlation measure the linear relationship between two random variables and treat them symmetrically. More often in the social sciences, we would like to explain one variable, called Y, in terms of another variable, say, X. Further, if Y is related to X in a nonlinear fashion, we would like

to know this. Call Y the explained variable and X the explanatory variable. For example, Y might be hourly wage, and X might be years of formal education.

We have already introduced the notion of the conditional probability density function of Y given X. Thus, we might want to see how the distribution of wages changes with education level. However, we usually want to have a simple way of summarizing this distribution. A single number will no longer suffice, because the distribution of Y given $X = x$ generally depends on the value of x. Nevertheless, we can summarize the relationship between Y and X by looking at the **conditional expectation** of Y given X, sometimes called the *conditional mean.* The idea is this. Suppose we know that X has taken on a particular value, say, x. Then, we can compute the expected value of Y, given that we know this outcome of X. We denote this expected value by $\mathrm{E}(Y|X = x)$, or sometimes $\mathrm{E}(Y|x)$ for shorthand. Generally, as x changes, so does $\mathrm{E}(Y|x)$.

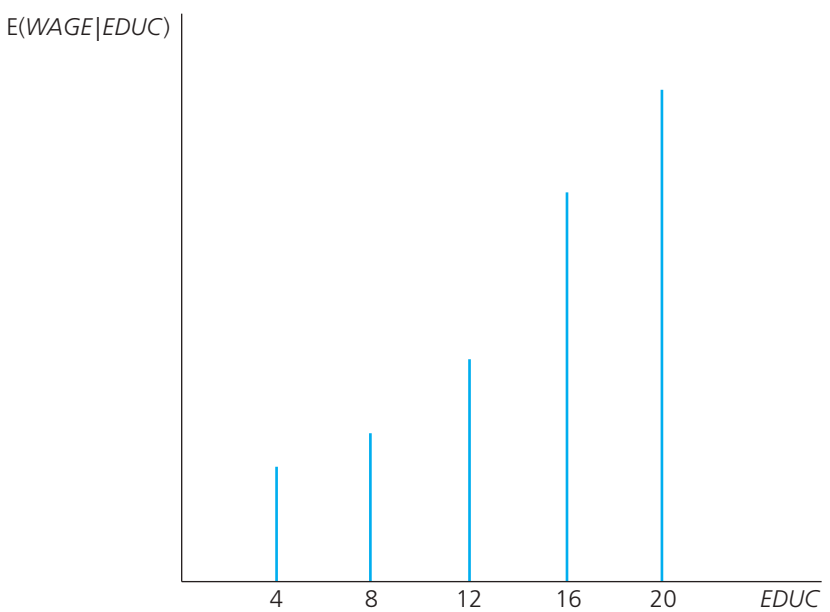
When Y is a discrete random variable taking on values $\{y_1, \ldots, y_m\}$, then

$$\mathrm{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

When Y is continuous, $\mathrm{E}(Y|x)$ is defined by integrating $yf_{Y|X}(y|x)$ over all possible values of y. As with unconditional expectations, the conditional expectation is a weighted average of possible values of Y, but now the weights reflect the fact that X has taken on a specific value. Thus, $E(Y|x)$ is just some function of x, which tells us how the expected value of Y varies with x.

As an example, let (X, Y) represent the population of all working individuals, where X is years of education and Y is hourly wage. Then, $\mathrm{E}(Y|X = 12)$ is the average hourly wage for all people in the population with 12 years of education (roughly a high school education). $\mathrm{E}(Y|X = 16)$ is the average hourly wage for all people with 16 years of education. Tracing out the expected value for various levels of education provides important information on how wages and education are related. See Figure B.5 for an illustration.

FIGURE B.5 The expected value of hourly wage given various levels of education.



In principle, the expected value of hourly wage can be found at each level of education, and these expectations can be summarized in a table. Because education can vary widely—and can even be measured in fractions of a year—this is a cumbersome way to show the relationship between average wage and amount of education. In econometrics, we typically specify simple functions that capture this relationship. As an example, suppose that the expected value of *WAGE* given *EDUC* is the linear function

$$\mathrm{E}(WAGE|EDUC) = 1.05 + .45\,EDUC.$$

If this relationship holds in the population of working people, the average wage for people with eight years of education is $1.05 + .45(8) = 4.65$, or \$4.65. The average wage for people with 16 years of education is 8.25, or \$8.25. The coefficient on *EDUC* implies that each year of education increases the expected hourly wage by .45, or 45¢.

Conditional expectations can also be nonlinear functions. For example, suppose that $\mathrm{E}(Y|x) = 10/x$, where X is a random variable that is always greater than zero. This function is graphed in Figure B.6. This could represent a demand function, where Y is quantity demanded and X is price. If Y and X are related in this way, an analysis of linear association, such as correlation analysis, would be incomplete.

B-4f Properties of Conditional Expectation

Several basic properties of conditional expectations are useful for derivations in econometric analysis.

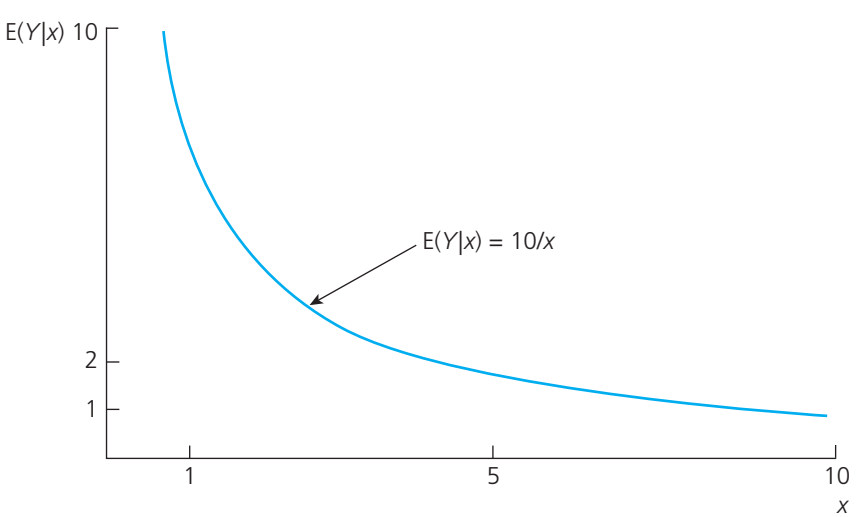
Property CE.1: $\mathrm{E}[c(X)|X] = c(X)$, for any function $c(X)$.

This first property means that functions of X behave as constants when we compute expectations conditional on X. For example, $\mathrm{E}(X^2|X) = X^2$. Intuitively, this simply means that if we know X, then we also know X^2.

Property CE.2: For functions $a(X)$ and $b(X)$,

$$\mathrm{E}[a(X)Y + b(X)|X] = a(X)\mathrm{E}(Y|X) + b(X).$$

FIGURE B.6 Graph of $\mathrm{E}(Y|x) = 10/x$.



For example, we can easily compute the conditional expectation of a function such as $XY + 2X^2$: $\mathrm{E}(XY + 2X^2|X) = X\mathrm{E}(Y|X) + 2X^2$.

The next property ties together the notions of independence and conditional expectations.

Property CE.3: If X and Y are independent, then $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

This property means that, if X and Y are independent, then the expected value of Y given X does not depend on X, in which case, $\mathrm{E}(Y|X)$ always equals the (unconditional) expected value of Y. In the wage and education example, if wages were independent of education, then the average wages of high school and college graduates would be the same. Because this is almost certainly false, we cannot assume that wage and education are independent.

A special case of Property CE.3 is the following: if U and X are independent and $\mathrm{E}(U) = 0$, then $\mathrm{E}(U|X) = 0$.

There are also properties of the conditional expectation that have to do with the fact that $\mathrm{E}(Y|X)$ is a function of X, say, $\mathrm{E}(Y|X) = \mu(X)$. Because X is a random variable, $\mu(X)$ is also a random variable. Furthermore, $\mu(X)$ has a probability distribution and therefore an expected value. Generally, the expected value of $\mu(X)$ could be very difficult to compute directly. The **law of iterated expectations** says that the expected value of $\mu(X)$ is simply equal to the expected value of Y. We write this as follows.

Property CE.4: $\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$.

This property is a little hard to grasp at first. It means that, if we first obtain $\mathrm{E}(Y|X)$ as a function of X and take the expected value of this (with respect to the distribution of X, of course), then we end up with $\mathrm{E}(Y)$. This is hardly obvious, but it can be derived using the definition of expected values.

As an example of how to use Property CE.4, let $Y = WAGE$ and $X = EDUC$, where $WAGE$ is measured in hours and $EDUC$ is measured in years. Suppose the expected value of $WAGE$ given $EDUC$ is $\mathrm{E}(WAGE|EDUC) = 4 + .60\,EDUC$. Further, $\mathrm{E}(EDUC) = 11.5$. Then, the law of iterated expectations implies that $\mathrm{E}(WAGE) = \mathrm{E}(4 + .60\,EDUC) = 4 + .60\,\mathrm{E}(EDUC) = 4 + .60(11.5) = 10.90$, or \$10.90 an hour.

The next property states a more general version of the law of iterated expectations.

Property CE.4': $\mathrm{E}(Y|X) = \mathrm{E}[\mathrm{E}(Y|X, Z)|X]$.

In other words, we can find $\mathrm{E}(Y|X)$ in two steps. First, find $\mathrm{E}(Y|X, Z)$ for any other random variable Z. Then, find the expected value of $\mathrm{E}(Y|X, Z)$, conditional on X.

Property CE.5: If $\mathrm{E}(Y|X) = \mathrm{E}(Y)$, then $\mathrm{Cov}(X, Y) = 0$ [and so $\mathrm{Corr}(X, Y) = 0$]. In fact, *every* function of X is uncorrelated with Y.

This property means that, if knowledge of X does not change the expected value of Y, then X and Y *must* be uncorrelated, which implies that if X and Y are correlated, then $\mathrm{E}(Y|X)$ must depend on X. The converse of Property CE.5 is not true: if X and Y are uncorrelated, $\mathrm{E}(Y|X)$ *could* still depend on X. For example, suppose $Y = X^2$. Then, $\mathrm{E}(Y|X) = X^2$, which is clearly a function of X. However, as we mentioned in our discussion of covariance and correlation, it is possible that X and X^2 are uncorrelated. The conditional expectation captures the nonlinear relationship between X and Y that correlation analysis would miss entirely.

Properties CE.4 and CE.5 have two important implications: if U and X are random variables such that $\mathrm{E}(U|X) = 0$, then $\mathrm{E}(U) = 0$, and U and X are uncorrelated.

Property CE.6: If $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}[g(X)^2] < \infty$ for some function g, then $\mathrm{E}\{[Y - \mu(X)]^2|X\} \leq \mathrm{E}\{[Y - g(X)]^2|X\}$ and $\mathrm{E}\{[Y - \mu(X)]^2\} \leq \mathrm{E}\{[Y - g(X)]^2\}$.

Property CE.6 is very useful in predicting or forecasting contexts. The first inequality says that, if we measure prediction inaccuracy as the *expected* squared prediction error, conditional on X, then the conditional mean is better than any other function of X for predicting Y. The conditional mean also minimizes the unconditional expected squared prediction error.

B-4g Conditional Variance

Given random variables X and Y, the variance of Y, conditional on $X = x$, is simply the variance associated with the conditional distribution of Y, given $X = x$: $\mathrm{E}\{[Y - \mathrm{E}(Y|x)]^2|x\}$. The formula

$$\mathrm{Var}(Y|X = x) = \mathrm{E}(Y^2|x) - [\mathrm{E}(Y|x)]^2$$

is often useful for calculations. Only occasionally will we have to compute a conditional variance. But we will have to make assumptions about and manipulate conditional variances for certain topics in regression analysis.

As an example, let $Y = SAVING$ and $X = INCOME$ (both of these measured annually for the population of all families). Suppose that $\mathrm{Var}(SAVING|INCOME) = 400 + .25\,INCOME$. This says that, as income increases, the variance in saving levels also increases. It is important to see that the relationship between the variance of *SAVING* and *INCOME* is totally separate from that between the *expected value* of *SAVING* and *INCOME*.

We state one useful property about the conditional variance.

Property CV.1: If X and Y are independent, then $\mathrm{Var}(Y|X) = \mathrm{Var}(Y)$.

This property is pretty clear, as the distribution of Y given X does not depend on X, and $\mathrm{Var}(Y|X)$ is just one feature of this distribution.

B-5 The Normal and Related Distributions

B-5a The Normal Distribution

The normal distribution and those derived from it are the most widely used distributions in statistics and econometrics. Assuming that random variables defined over populations are normally distributed simplifies probability calculations. In addition, we will rely heavily on the normal and related distributions to conduct inference in statistics and econometrics—even when the underlying population is not necessarily normal. We must postpone the details, but be assured that these distributions will arise many times throughout this text.

A normal random variable is a continuous random variable that can take on any value. Its probability density function has the familiar bell shape graphed in Figure B.7.

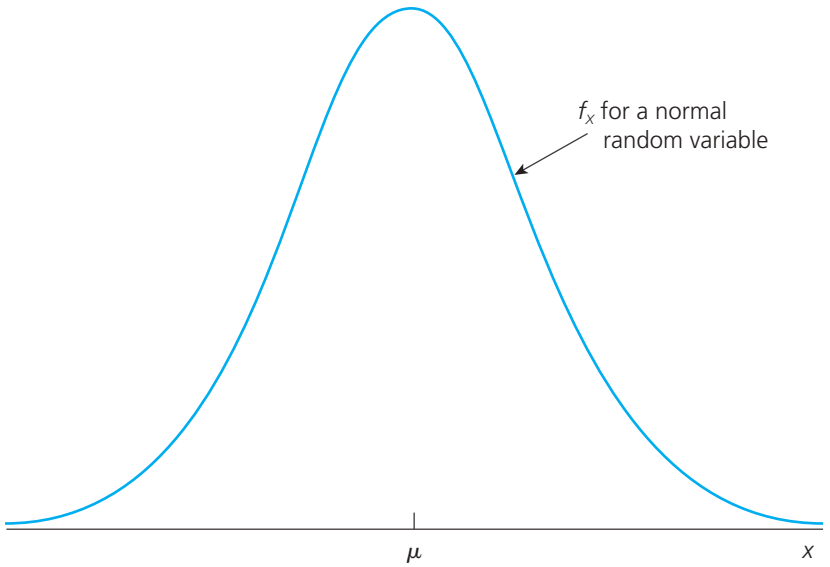
Mathematically, the pdf of X can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty, \qquad [B.34]$$

where $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. We say that X has a **normal distribution** with expected value μ and variance σ^2, written as $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Because the normal distribution is symmetric about μ, μ is also the median of X. The normal distribution is sometimes called the *Gaussian distribution* after the famous mathematician C. F. Gauss.

Certain random variables appear to roughly follow a normal distribution. Human heights and weights, test scores, and county unemployment rates have pdfs roughly the shape in Figure B.7. Other distributions, such as income distributions, do not appear to follow the normal probability function. In most countries, income is not symmetrically distributed about any value; the distribution is skewed toward the upper tail. In some cases, a variable can be transformed to achieve normality. A popular transformation is

FIGURE B.7 The general shape of the normal probability density function.



the natural log, which makes sense for positive random variables. If X is a positive random variable, such as income, and $Y = \log(X)$ has a normal distribution, then we say that X has a *lognormal distribution*. It turns out that the lognormal distribution fits income distribution pretty well in many countries. Other variables, such as prices of goods, appear to be well described as lognormally distributed.

B-5b The Standard Normal Distribution

One special case of the normal distribution occurs when the mean is zero and the variance (and, therefore, the standard deviation) is unity. If a random variable Z has a Normal(0,1) distribution, then we say it has a **standard normal distribution**. The pdf of a standard normal random variable is denoted $\phi(z)$; from (B.34), with $\mu = 0$ and $\sigma^2 = 1$, it is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2), \quad -\infty < z < \infty. \qquad \text{[B.35]}$$

The standard normal cumulative distribution function is denoted $\Phi(z)$ and is obtained as the area under ϕ, to the left of z; see Figure B.8. Recall that $\Phi(z) = P(Z \leq z)$; because Z is continuous, $\Phi(z) = P(Z < z)$ as well.

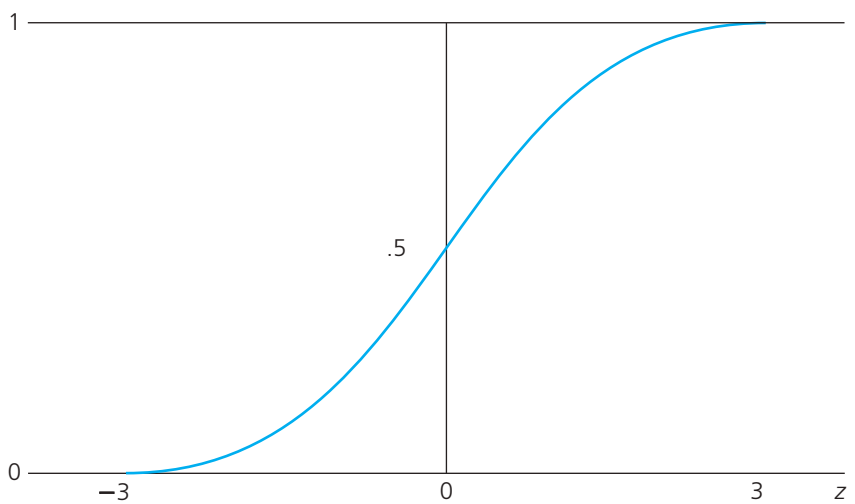
No simple formula can be used to obtain the values of $\Phi(z)$ [because $\Phi(z)$ is the integral of the function in (B.35), and this integral has no closed form]. Nevertheless, the values for $\Phi(z)$ are easily tabulated; they are given for z between −3.1 and 3.1 in Table G.1 in Statistical Tables. For $z \leq -3.1$, $\Phi(z)$ is less than .001, and for $z \geq 3.1$, $\Phi(z)$ is greater than .999. Most statistics and econometrics software packages include simple commands for computing values of the standard normal cdf, so we can often avoid printed tables entirely and obtain the probabilities for any value of z.

Using basic facts from probability—and, in particular, properties (B.7) and (B.8) concerning cdfs—we can use the standard normal cdf for computing the probability of any event involving a standard normal random variable. The most important formulas are

$$P(Z > z) = 1 - \Phi(z), \qquad \text{[B.36]}$$

$$P(Z < -z) = P(Z > z), \qquad \text{[B.37]}$$

FIGURE B.8 The standard normal cumulative distribution function.



and

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \qquad [B.38]$$

Because Z is a continuous random variable, all three formulas hold whether or not the inequalities are strict. Some examples include $P(Z > .44) = 1 - .67 = .33$, $P(Z < -.92) = P(Z > .92) = 1 - .821 = .179$, and $P(-1 < Z \leq .5) = .692 - .159 = .533$.

Another useful expression is that, for any $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \qquad [B.39] \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned}$$

Thus, the probability that the absolute value of Z is bigger than some positive constant c is simply twice the probability $P(Z > c)$; this reflects the symmetry of the standard normal distribution.

In most applications, we start with a normally distributed random variable, $X \sim \text{Normal}(\mu, \sigma^2)$, where μ is different from zero and $\sigma^2 \neq 1$. Any normal random variable can be turned into a standard normal using the following property.

Property Normal.1: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim \text{Normal}(0, 1)$.

Property Normal.1 shows how to turn any normal random variable into a standard normal. Thus, suppose $X \sim \text{Normal}(3, 4)$, and we would like to compute $P(X \leq 1)$. The steps always involve the normalization of X to a standard normal:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = .159. \end{aligned}$$

EXAMPLE B.6 **Probabilities for a Normal Random Variable**

First, let us compute $P(2 < X \leq 6)$ when $X \sim \text{Normal}(4,9)$ (whether we use $<$ or $\leq$ is irrelevant because X is a continuous random variable). Now,

$$\begin{aligned} P(2 < X \leq 6) &= P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3) \\ &= \Phi(.67) - \Phi(-.67) = .749 - .251 = .498. \end{aligned}$$

Now, let us compute $P(|X| > 2)$:

$$\begin{aligned} P(|X| > 2) &= P(X > 2) + P(X < -2) \\ &= P[(X-4)/3 > (2-4)/3] + P[(X-4)/3 < (-2-4)/3] \\ &= 1 - \Phi(-2/3) + \Phi(-2) \\ &= 1 - .251 + .023 = .772. \end{aligned}$$

B-5c Additional Properties of the Normal Distribution

We end this subsection by collecting several other facts about normal distributions that we will later use.

Property Normal.2: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Thus, if $X \sim \text{Normal}(1,9)$, then $Y = 2X + 3$ is distributed as normal with mean $2E(X) + 3 = 5$ and variance $2^2 \cdot 9 = 36$; $\text{sd}(Y) = 2\text{sd}(X) = 2 \cdot 3 = 6$.

Earlier, we discussed how, in general, zero correlation and independence are not the same. In the case of normally distributed random variables, it turns out that zero correlation suffices for independence.

Property Normal.3: If X and Y are jointly normally distributed, then they are independent if, and only if, $\text{Cov}(X, Y) = 0$.

Property Normal.4: Any linear combination of independent, identically distributed normal random variables has a normal distribution.

For example, let X_i, for $i = 1, 2$, and 3, be independent random variables distributed as $\text{Normal}(\mu, \sigma^2)$. Define $W = X_1 + 2X_2 - 3X_3$. Then, W is normally distributed; we must simply find its mean and variance. Now,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

Also,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

Property Normal.4 also implies that the average of independent, normally distributed random variables has a normal distribution. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as $\text{Normal}(\mu, \sigma^2)$, then

$$\overline{Y} \sim \text{Normal}(\mu, \sigma^2/n). \qquad \text{[B.40]}$$

This result is critical for statistical inference about the mean in a normal population.

Other features of the normal distribution are worth knowing, although they do not play a central role in the text. Because a normal random variable is symmetric about its mean, it has zero skewness, that is, $E[(X - \mu)^3] = 0$. Further, it can be shown that

$$E[(X - \mu)^4]/\sigma^4 = 3,$$

or $E(Z^4) = 3$, where Z has a standard normal distribution. Because the normal distribution is so prevalent in probability and statistics, the measure of kurtosis for any given random variable X (whose fourth moment exists) is often defined to be $E[(X - \mu)^4]/\sigma^4 - 3$, that is, relative to the value for the standard normal distribution. If $E[(X - \mu)^4]/\sigma^4 > 3$, then the distribution of X has fatter tails than the normal distribution (a somewhat common occurrence, such as with the t distribution to be introduced shortly); if $E[(X - \mu)^4]/\sigma^4 < 3$, then the distribution has thinner tails than the normal (a rarer situation).

B-5d The Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \qquad [B.41]$$

Then, X has what is known as a **chi-square distribution** with n **degrees of freedom** (or *df* for short). We write this as $X \sim \chi_n^2$. The *df* in a chi-square distribution corresponds to the number of terms in the sum in (B.41). The concept of degrees of freedom will play an important role in our statistical and econometric analyses.

The pdf for chi-square distributions with varying degrees of freedom is given in Figure B.9; we will not need the formula for this pdf, and so we do not reproduce it here. From equation (B.41), it is clear that a chi-square random variable is always nonnegative, and that, unlike the normal distribution, the chi-square distribution is not symmetric about any point. It can be shown that if $X \sim \chi_n^2$, then the expected value of X is n [the number of terms in (B.41)], and the variance of X is $2n$.

B-5e The *t* Distribution

The t distribution is the workhorse in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal and a chi-square random variable.

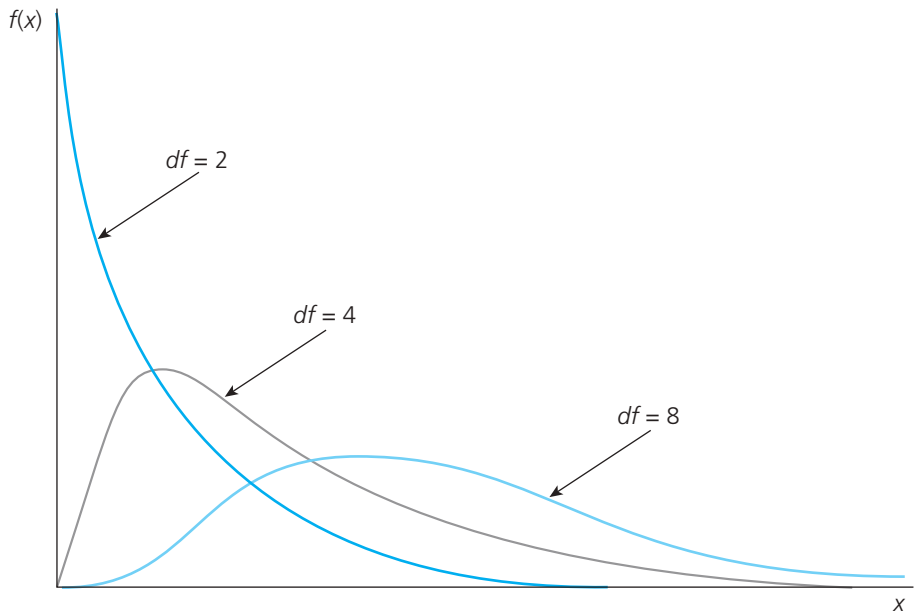
Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Further, assume that Z and X are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{X/n}} \qquad [B.42]$$

has a ***t* distribution** with n degrees of freedom. We will denote this by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi-square random variable in the denominator of (B.42).

The pdf of the t distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails. The expected value of a t distributed random variable is zero (strictly speaking, the expected value exists only for $n > 1$),

FIGURE B.9 The chi-square distribution with various degrees of freedom.



and the variance is $n/(n-2)$ for $n > 2$. (The variance does not exist for $n \leq 2$ because the distribution is so spread out.) The pdf of the t distribution is plotted in Figure B.10 for various degrees of freedom. As the degrees of freedom gets large, the t distribution approaches the standard normal distribution.

B-5f The *F* Distribution

Another important distribution for statistics and econometrics is the F distribution. In particular, the F distribution will be used for testing hypotheses in the context of multiple regression analysis.

To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \qquad [B.43]$$

has an ***F* distribution** with (k_1, k_2) degrees of freedom. We denote this as $F \sim F_{k_1, k_2}$. The pdf of the F distribution with different degrees of freedom is given in Figure B.11.

The order of the degrees of freedom in F_{k_1, k_2} is critical. The integer k_1 is called the *numerator degrees of freedom* because it is associated with the chi-square variable in the numerator. Likewise, the integer k_2 is called the *denominator degrees of freedom* because it is associated with the chi-square variable in the denominator. This can be a little tricky because (B.43) can also be written as $(X_1k_2)/(X_2k_1)$, so that k_1 appears in the denominator. Just remember that the numerator *df* is the integer associated with the chi-square variable in the numerator of (B.43), and similarly for the denominator *df*.

FIGURE B.10 The *t* distribution with various degrees of freedom.

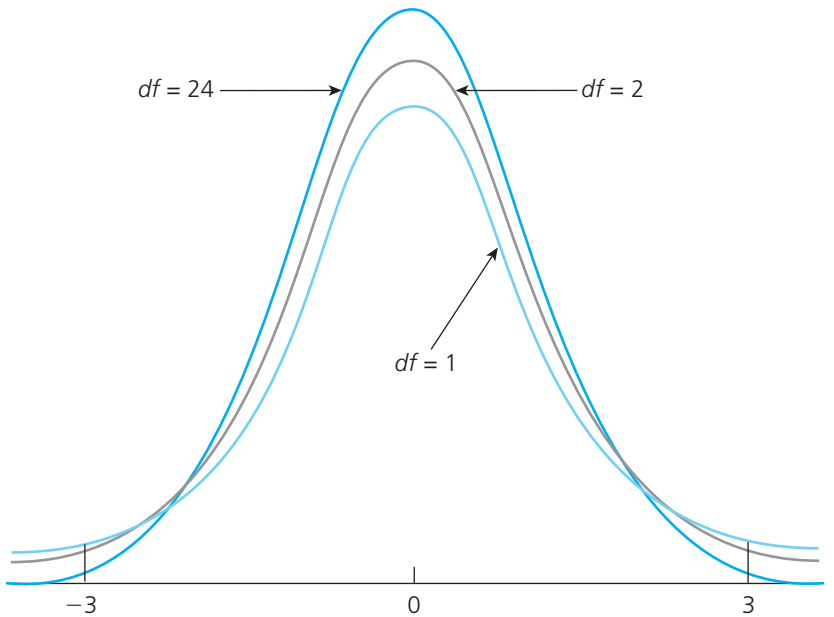
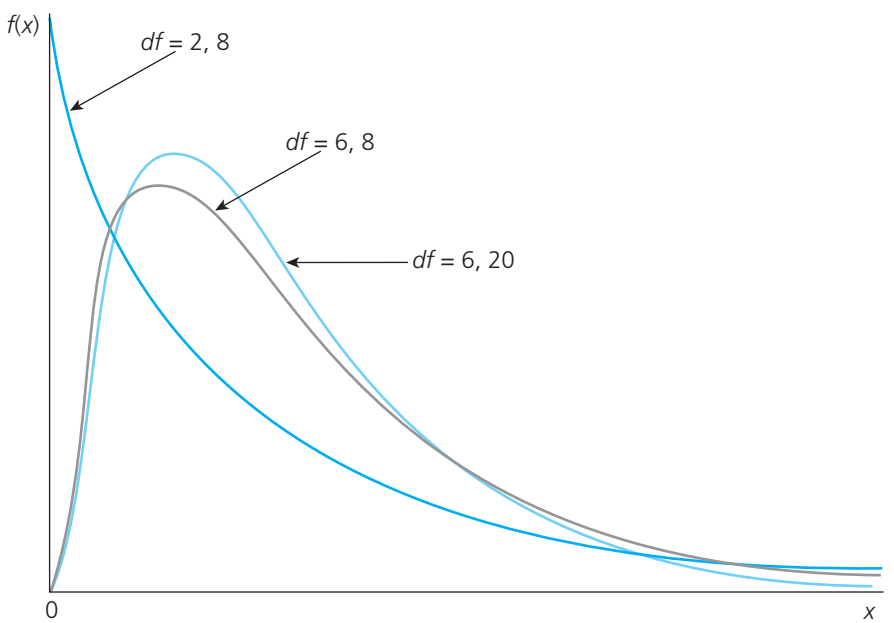


FIGURE B.11 The F_{k_1, k_2} distribution for various degrees of freedom, k_1 and k_2.



Summary

In this Math Refresher, we have reviewed the probability concepts that are needed in econometrics. Most of the concepts should be familiar from your introductory course in probability and statistics. Some of the more advanced topics, such as features of conditional expectations, do not need to be mastered now—there is time for that when these concepts arise in the context of regression analysis in Part 1.

In an introductory statistics course, the focus is on calculating means, variances, covariances, and so on for particular distributions. In Part 1, we will not need such calculations: we mostly rely on the *properties* of expectations, variances, and so on that have been stated in this Math Refresher.

Key Terms

Bernoulli (or Binary) Random Variable
Binomial Distribution
Chi-Square Distribution
Conditional Distribution
Conditional Expectation
Continuous Random Variable
Correlation Coefficient
Covariance
Cumulative Distribution Function (cdf)
Degrees of Freedom
Discrete Random Variable
Expected Value
Experiment
F Distribution
Independent Random Variables
Joint Distribution
Kurtosis
Law of Iterated Expectations
Median
Normal Distribution
Pairwise Uncorrelated Random Variables
Probability Density Function (pdf)
Random Variable
Skewness
Standard Deviation
Standard Normal Distribution
Standardized Random Variable
Symmetric Distribution
t Distribution
Uncorrelated Random Variables
Variance

Problems

1 Suppose that a high school student is preparing to take the SAT exam. Explain why his or her eventual SAT score is properly viewed as a random variable.

2 Let X be a random variable distributed as Normal(5,4). Find the probabilities of the following events:
(i) $P(X \leq 6)$.
(ii) $P(X > 4)$.
(iii) $P(|X - 5| > 1)$.

3 Much is made of the fact that certain mutual funds outperform the market year after year (that is, the return from holding shares in the mutual fund is higher than the return from holding a portfolio such as the S&P 500). For concreteness, consider a 10-year period and let the population be the 4,170 mutual funds reported in *The Wall Street Journal* on January 1, 1995. By saying that performance relative to the market is random, we mean that each fund has a 50–50 chance of outperforming the market in any year and that performance is independent from year to year.
(i) If performance relative to the market is truly random, what is the probability that any particular fund outperforms the market in all 10 years?
(ii) Of the 4,170 mutual funds, what is the expected number of funds that will outperform the market in all 10 years?
(iii) Find the probability that *at least* one fund out of 4,170 funds outperforms the market in all 10 years. What do you make of your answer?
(iv) If you have a statistical package that computes binomial probabilities, find the probability that at least five funds outperform the market in all 10 years.

4 For a randomly selected county in the United States, let X represent the proportion of adults over age 65 who are employed, or the elderly employment rate. Then, X is restricted to a value between zero and one. Suppose that the cumulative distribution function for X is given by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least .6 (60%).

5 Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that about 20% of the adult population believed Simpson was innocent (after much of the physical evidence in the case had been revealed to the public). Ignore the fact that this 20% is an estimate based on a subsample from the population; for illustration, take it as the true percentage of people who thought Simpson was innocent prior to jury selection. Assume that the 12 jurors were selected randomly and independently from the population (although this turned out not to be true).

(i) Find the probability that the jury had at least one member who believed in Simpson's innocence prior to jury selection. [*Hint*: Define the Binomial(12,.20) random variable X to be the number of jurors believing in Simpson's innocence.]

(ii) Find the probability that the jury had at least two members who believed in Simpson's innocence. [*Hint*: $P(X \geq 2) = 1 - P(X \leq 1)$ and $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

6 (Requires calculus) Let X denote the prison sentence, in years, for people convicted of auto theft in a particular state in the United States. Suppose that the pdf of X is given by

$$f(x) = (1/9)x^2, 0 < x < 3.$$

Use integration to find the expected prison sentence.

7 If a basketball player is a 74% free throw shooter, then, on average, how many free throws will he or she make in a game with eight free throw attempts?

8 Suppose that a college student is taking three courses: a two-credit course, a three-credit course, and a four-credit course. The expected grade in the two-credit course is 3.5, while the expected grade in the three- and four-credit courses is 3.0. What is the expected overall grade point average for the semester? (Remember that each course grade is weighted by its share of the total number of units.)

9 Let X denote the annual salary of university professors in the United States, measured in thousands of dollars. Suppose that the average salary is 52.3, with a standard deviation of 14.6. Find the mean and standard deviation when salary is measured in dollars.

10 Suppose that at a large university, college grade point average, *GPA*, and SAT score, *SAT*, are related by the conditional expectation $E(GPA|SAT) = .70 + .002\,SAT$.

(i) Find the expected *GPA* when $SAT = 800$. Find $E(GPA|SAT = 1{,}400)$. Comment on the difference.

(ii) If the average *SAT* in the university is 1,100, what is the average *GPA*? (*Hint*: Use Property CE.4.)

(iii) If a student's SAT score is 1,100, does this mean he or she will have the GPA found in part (ii)? Explain.

11 (i) Let X be a random variable taking on the values -1 and 1, each with probability 1/2. Find $E(X)$ and $E(X^2)$.

(ii) Now let X be a random variable taking on the values 1 and 2, each with probability 1/2. Find $E(X)$ and $E(1/X)$.

(iii) Conclude from parts (i) and (ii) that, in general,

$$E[g(X)] \neq g[E(X)]$$

for a nonlinear function $g(\cdot)$.

(iv) Given the definition of the F random variable in equation (B.43), show that

$$E(F) = E\left[\frac{1}{(X_2/k_2)}\right].$$

Can you conclude that $E(F) = 1$?

12 The *geometric distribution* can be used to model the number of trials before a certain event occurs. For example, we might flip a coin repeatedly until the first head appears. If the coin is fair, the probability of getting a head on each flip is 0. 5. Furthermore, we may realistically assume that the trials are independent. The flip on which the first head occurs can be represented by a random variable, X.

For the general geometric distribution, we maintain the assumption of independent trails—which, admittedly, is sometimes too strong—but allow the probability of the event occurring on any trial to be θ for any $0 < \theta < 1$. We assume that this probability is the same from trial to trial. In the coin-flipping example, allowing the coin to be biased toward, say, heads, would mean $\theta > 0.5$. Another example would be an unemployed worker repeatedly interviewing for jobs until the first job offer. Then θ is the probability of receiving an offer during any particular interview. To follow the geometric distribution, we would assume θ is the same for all interviews and that the outcomes are independent across interviews. Both assumptions may be too strong.

One way to characterize the geometric distribution is to define a sequence of Bernoulli (binary) variables, say $W_1, W_2, W_3, \ldots$. If $W_k = 1$ then the event occurs on trial k; if $W_k = 0$, it does not occur. Assume that the W_k are independent across k with the Bernoulli(θ) distribution, so that $P(W_k = 1) = \theta$.

(i) Let X denote the trial upon which the first event occurs. The possible values of X are $\{1, 2, 3, \ldots\}$. Show that for any positive integer k,

$$P(X = k) = (1 - \theta)^{k-1}\theta.$$

[Hint: If $X = k$, you must observe $k - 1$ "failures" (zeros) followed by a "success" (one).]

(ii) Use the formula for a geometric sum to show that

$$P(X \leq k) = 1 - (1 - \theta)^{k}, k = 1, 2, \ldots.$$

(iii) Suppose you have observed 29 failures in a row. If $\theta = 0.04$, what is the probability of observing a success on the 30^{th} trial?

(iv) In the setup of part (iii), before conducting any of the trials, what is the probability that the first success occurs before the 30^{th} trial?

(v) Reconcile your answers from parts (iii) and (iv).

13 In March of 1985, the NCAA men's basketball tournament increased its field of teams to 64. Since that time, each year of the tournament involves four games pitting a #1 seed against a #16 seed. The #1 seeds are purportedly awarded to the four most deserving teams. The #16 teams are generally viewed as the weakest four teams in the field. In answering this question, we will make some simplifying assumptions to make the calculations easier.

(i) Assume that the probability of a #16 seed beating a #1 seed is ρ, where $0 < \rho < 1$. (In practice, ρ varies by matchup, but we will assume it is the same across all matchups and years.) Assume that the outcomes of #1 vs #16 games are independent of one another. Show that the probability that at least one #16 seed wins in a particular year is $1 - (1 - \rho)^4$. Evaluate this probability when $\rho = 0.02$. [Hint: You might define four binary variables, say Z_1, Z_2, Z_3, and Z_4,where $Z_i = 1$ if game i is won by the #16 seed. Then first compute $P(Z_1 = 0, Z_2 = 0, Z_3 = 0, Z_4 = 0)$.]

(ii) Let X be the number of years before a #16 beats a #1 seed in the tournament. Assuming independence across years—a very reasonable assumption—explain why X has a geometric distribution and that the probability of "success" on a given trial is $\theta = 1 - (1 - \rho)^4$.

(iii) In the 2017 NCCA Men's Tournament, #16 seed University of Maryland, Baltimore County defeated #1 seed University of Virginia. It took 33 years for such an upset to occur. Suppose $\rho = 0.02$. Find $P(X \leq 33)$. Interpret this probability using the perspective of a basketball observer in February 1985.

(iv) Using $\rho = 0.02$, in February 2018 what was the probability that a #16 seed would defeat a #1 seed in the March 2018 tournament? (It had not happened in the previous 32 years.) Why does this differ so much from your answer in part (iii)?

(v) Derive the general formula

$$P(X \leq k) = 1 - (1 - \rho)^{4k}.$$

Math Refresher C

Fundamentals of Mathematical Statistics

C-1 Populations, Parameters, and Random Sampling

Statistical inference involves learning something about a population given the availability of a sample from that population. By **population**, we mean any well-defined group of subjects, which could be individuals, firms, cities, or many other possibilities. By "learning," we can mean several things, which are broadly divided into the categories of *estimation* and *hypothesis testing*.

A couple of examples may help you understand these terms. In the population of all working adults in the United States, labor economists are interested in learning about the return to education, as measured by the average percentage increase in earnings given another year of education. It would be impractical and costly to obtain information on earnings and education for the entire working population in the United States, but we can obtain data on a subset of the population. Using the data collected, a labor economist may report that his or her best estimate of the return to another year of education is 7.5%. This is an example of a *point estimate*. Or, she or he may report a range, such as "the return to education is between 5.6% and 9.4%." This is an example of an *interval estimate*.

An urban economist might want to know whether neighborhood crime watch programs are associated with lower crime rates. After comparing crime rates of neighborhoods with and without such programs in a sample from the population, he or she can draw one of two conclusions: neighborhood watch programs do affect crime, or they do not. This example falls under the rubric of hypothesis testing.

The first step in statistical inference is to identify the population of interest. This may seem obvious, but it is important to be very specific. Once we have identified the population, we can specify a model for the population relationship of interest. Such models involve probability distributions or features of probability distributions, and these depend on unknown parameters. Parameters are simply constants that determine the directions and strengths of relationships among variables. In the labor economics example just presented, the parameter of interest is the return to education in the population.

C-1a Sampling

For reviewing statistical inference, we focus on the simplest possible setting. Let Y be a random variable representing a population with a probability density function $f(y; \theta)$, which depends on the single parameter θ. The probability density function (pdf) of Y is assumed to be known except for the value of θ; different values of θ imply different population distributions, and therefore we are interested in the value of θ. If we can obtain certain kinds of samples from the population, then we can learn something about θ. The easiest sampling scheme to deal with is random sampling.

Random Sampling. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with a common probability density function $f(y; \theta)$, then $\{Y_1, \ldots, Y_n\}$ is said to be a **random sample** from $f(y; \theta)$ [or a random sample from the population represented by $f(y; \theta)$].

When $\{Y_1, \ldots, Y_n\}$ is a random sample from the density $f(y; \theta)$, we also say that the Y_i are *independent, identically distributed* (or *i.i.d.*) random variables from $f(y; \theta)$. In some cases, we will not need to entirely specify what the common distribution is.

The random nature of $Y_1, Y_2, \ldots, Y_n$ in the definition of random sampling reflects the fact that many different outcomes are possible before the sampling is actually carried out. For example, if family income is obtained for a sample of $n = 100$ families in the United States, the incomes we observe will usually differ for each different sample of 100 families. Once a sample is obtained, we have a set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, which constitute the data that we work with. Whether or not it is appropriate to assume the sample came from a random sampling scheme requires knowledge about the actual sampling process.

Random samples from a Bernoulli distribution are often used to illustrate statistical concepts, and they also arise in empirical applications. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Bernoulli(θ), so that $P(Y_i = 1) = \theta$ and $P(Y_i = 0) = 1 - \theta$, then $\{Y_1, Y_2, \ldots, Y_n\}$ constitutes a random sample from the Bernoulli(θ) distribution. As an illustration, consider the airline reservation example carried along in Math Refresher B. Each Y_i denotes whether customer i shows up for his or her reservation; $Y_i = 1$ if passenger i shows up, and $Y_i = 0$ otherwise. Here, θ is the probability that a randomly drawn person from the population of all people who make airline reservations shows up for his or her reservation.

For many other applications, random samples can be assumed to be drawn from a normal distribution. If $\{Y_1, \ldots, Y_n\}$ is a random sample from the Normal(μ, σ^2) population, then the population is characterized by two parameters, the mean μ and the variance σ^2. Primary interest usually lies in μ, but σ^2 is of interest in its own right because making inferences about μ often requires learning about σ^2.

C-2 Finite Sample Properties of Estimators

In this section, we study what are called finite sample properties of estimators. The term "finite sample" comes from the fact that the properties hold for a sample of any size, no matter how large or small. Sometimes, these are called small sample properties. In Section C-3, we cover "asymptotic properties," which have to do with the behavior of estimators as the sample size grows without bound.

C-2a Estimators and Estimates

To study properties of estimators, we must define what we mean by an estimator. Given a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ drawn from a population distribution that depends on an unknown parameter θ, an **estimator** of θ is a rule that assigns each possible outcome of the sample a value of θ. The rule is specified before any sampling is carried out; in particular, the rule is the same regardless of the data actually obtained.

As an example of an estimator, let $\{Y_1, \ldots, Y_n\}$ be a random sample from a population with mean μ. A natural estimator of μ is the average of the random sample:

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i. \qquad \text{[C.1]}$$

$\bar{Y}$ is called the **sample average** but, unlike in Math Refresher A where we defined the sample average of a set of numbers as a descriptive statistic, $\bar{Y}$ is now viewed as an estimator. Given any outcome of the random variables $Y_1, \ldots, Y_n$, we use the same rule to estimate μ: we simply average them. For actual data outcomes $\{y_1, \ldots, y_n\}$, the **estimate** is just the average in the sample: $\bar{y} = (y_1 + y_2 + \cdots + y_n)/n$.

EXAMPLE C.1 **City Unemployment Rates**

Suppose we obtain the following sample of unemployment rates for 10 cities in the United States:

City	Unemployment Rate
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6
8	3.5
9	5.8
10	7.5

Our estimate of the average city unemployment rate in the United States is $\bar{y} = 6.0$. Each sample generally results in a different estimate. But the *rule* for obtaining the estimate is the same, regardless of which cities appear in the sample, or how many.

More generally, an estimator W of a parameter θ can be expressed as an abstract mathematical formula:

$$W = h(Y_1, Y_2, \ldots, Y_n), \qquad \text{[C.2]}$$

for some known function h of the random variables $Y_1, Y_2, \ldots, Y_n$. As with the special case of the sample average, W is a random variable because it depends on the random sample: as we obtain different random samples from the population, the value of W can change. When a particular set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, is plugged into the function h, we obtain an *estimate* of θ, denoted $w = h(y_1, \ldots, y_n)$. Sometimes, W is called a point estimator and w a point estimate to distinguish these from *interval* estimators and estimates, which we will come to in Section C-5.

For evaluating estimation procedures, we study various properties of the probability distribution of the random variable W. The distribution of an estimator is often called its **sampling distribution**, because this distribution describes the likelihood of various outcomes of W across different random samples. Because there are unlimited rules for combining data to estimate parameters, we need some sensible criteria for choosing among estimators, or at least for eliminating some estimators from consideration. Therefore, we must leave the realm of descriptive statistics, where we compute things such as the sample average to simply summarize a body of data. In mathematical statistics, we study the sampling distributions of estimators.

C-2b Unbiasedness

In principle, the entire sampling distribution of W can be obtained given the probability distribution of Y_i and the function h. It is usually easier to focus on a few features of the distribution of W in evaluating it as an estimator of θ. The first important property of an estimator involves its expected value.

Unbiased Estimator. An estimator, W of θ, is an **unbiased estimator** if

$$\mathrm{E}(W) = \theta, \qquad \text{[C.3]}$$

for all possible values of θ.

If an estimator is unbiased, then its probability distribution has an expected value equal to the parameter it is supposed to be estimating. Unbiasedness does *not* mean that the estimate we get with any particular sample is equal to θ, or even very close to θ. Rather, if we could *indefinitely* draw random samples on Y from the population, compute an estimate each time, and then average these estimates over all random samples, we would obtain θ. This thought experiment is abstract because, in most applications, we just have one random sample to work with.

For an estimator that is not unbiased, we define its **bias** as follows.

Bias of an Estimator. If W is a **biased estimator** of θ, its bias is defined

$$\text{Bias}(W) \equiv \text{E}(W) - \theta. \qquad \text{[C.4]}$$

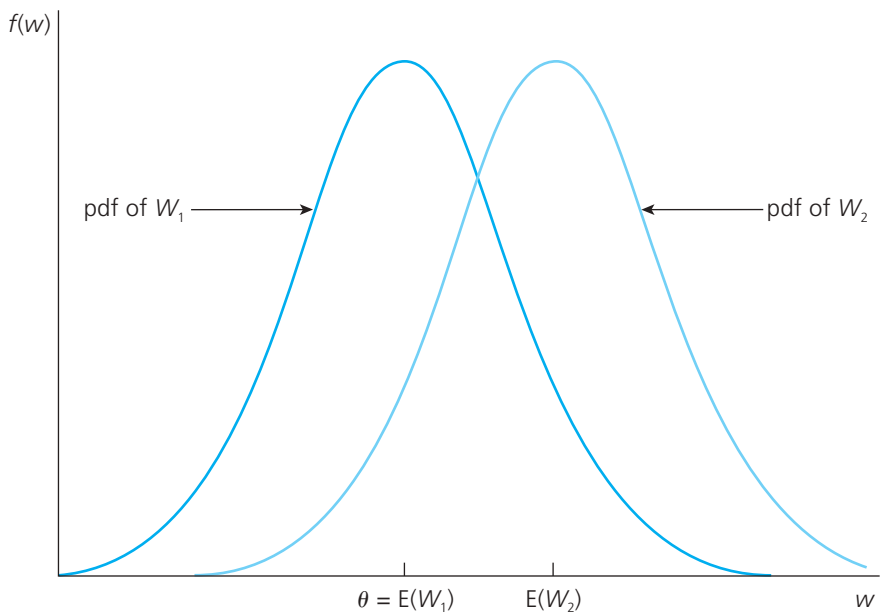
Figure C.1 shows two estimators; the first one is unbiased, and the second one has a positive bias.

The unbiasedness of an estimator and the size of any possible bias depend on the distribution of Y and on the function h. The distribution of Y is usually beyond our control (although we often choose a *model* for this distribution): it may be determined by nature or social forces. But the choice of the rule h is ours, and if we want an unbiased estimator, then we must choose h accordingly.

Some estimators can be shown to be unbiased quite generally. We now show that the sample average $\overline{Y}$ is an unbiased estimator of the population mean μ, regardless of the underlying population distribution. We use the properties of expected values (E.1 and E.2) that we covered in Section B-3:

$$\text{E}(\overline{Y}) = \text{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\text{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n} \text{E}(Y_i)\right)$$
$$= (1/n)\left(\sum_{i=1}^{n} \mu\right) = (1/n)(n\mu) = \mu.$$

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



For hypothesis testing, we will need to estimate the variance σ^2 from a population with mean μ. Letting $\{Y_1, \ldots, Y_n\}$ denote the random sample from the population with $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2$, define the estimator as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2, \qquad \text{[C.5]}$$

which is usually called the **sample variance**. It can be shown that S^2 is unbiased for σ^2: $\mathrm{E}(S^2) = \sigma^2$. The division by $n - 1$, rather than n, accounts for the fact that the mean μ is estimated rather than known. If μ were known, an unbiased estimator of σ^2 would be $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$, but μ is rarely known in practice.

Although unbiasedness has a certain appeal as a property for an estimator—indeed, its antonym, "biased," has decidedly negative connotations—it is not without its problems. One weakness of unbiasedness is that some reasonable, and even some very good, estimators are not unbiased. We will see an example shortly.

Another important weakness of unbiasedness is that unbiased estimators exist that are actually quite poor estimators. Consider estimating the mean μ from a population. Rather than using the sample average $\overline{Y}$ to estimate μ, suppose that, after collecting a sample of size n, we discard all of the observations except the first. That is, our estimator of μ is simply $W \equiv Y_1$. This estimator is unbiased because $\mathrm{E}(Y_1) = \mu$. Hopefully, you sense that ignoring all but the first observation is not a prudent approach to estimation: it throws out most of the information in the sample. For example, with $n = 100$, we obtain 100 outcomes of the random variable Y, but then we use only the first of these to estimate $\mathrm{E}(Y)$.

C-2c The Sampling Variance of Estimators

The example at the end of the previous subsection shows that we need additional criteria to evaluate estimators. Unbiasedness only ensures that the sampling distribution of an estimator has a mean value equal to the parameter it is supposed to be estimating. This is fine, but we also need to know how spread out the distribution of an estimator is. An estimator can be equal to θ, on average, but it can also be very far away with large probability. In Figure C.2, W_1 and W_2 are both unbiased estimators of θ. But the distribution of W_1 is more tightly centered about θ: the probability that W_1 is greater than any given distance from θ is less than the probability that W_2 is greater than that same distance from θ. Using W_1 as our estimator means that it is less likely that we will obtain a random sample that yields an estimate very far from θ.

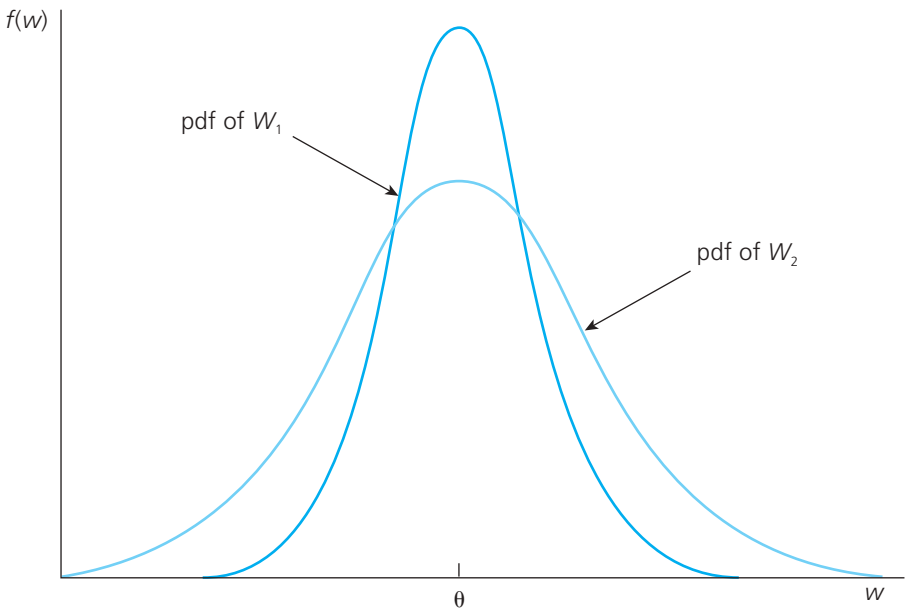
To summarize the situation shown in Figure C.2, we rely on the variance (or standard deviation) of an estimator. Recall that this gives a single measure of the dispersion in the distribution. The variance of an estimator is often called its **sampling variance** because it is the variance associated with a sampling distribution. Remember, the sampling variance is not a random variable; it is a constant, but it might be unknown.

We now obtain the variance of the sample average for estimating the mean μ from a population:

$$\begin{aligned}\mathrm{Var}(\overline{Y}) &= \mathrm{Var}\left((1/n)\sum_{i=1}^{n}Y_i\right) = (1/n^2)\mathrm{Var}\left(\sum_{i=1}^{n}Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n}\mathrm{Var}(Y_i)\right)\\ &= (1/n^2)\left(\sum_{i=1}^{n}\sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n. \qquad \text{[C.6]}\end{aligned}$$

Notice how we used the properties of variance from Sections B-3 and B-4 (VAR.2 and VAR.4), as well as the independence of the Y_i. To summarize: If $\{Y_i: i = 1, 2, \ldots, n\}$ is a random sample from a population with mean μ and variance σ^2, then $\overline{Y}$ has the same mean as the population, but its sampling variance equals the population variance, σ^2, divided by the sample size.

FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



An important implication of $\text{Var}(\bar{Y}) = \sigma^2/n$ is that it can be made very close to zero by increasing the sample size n. This is a key feature of a reasonable estimator, and we return to it in Section C-3.

As suggested by Figure C.2, among unbiased estimators, we prefer the estimator with the smallest variance. This allows us to eliminate certain estimators from consideration. For a random sample from a population with mean μ and variance σ^2, we know that $\bar{Y}$ is unbiased and $\text{Var}(\bar{Y}) = \sigma^2/n$. What about the estimator Y_1, which is just the first observation drawn? Because Y_1 is a random draw from the population, $\text{Var}(Y_1) = \sigma^2$. Thus, the difference between $\text{Var}(Y_1)$ and $\text{Var}(\bar{Y})$ can be large even for small sample sizes. If $n = 10$, then $\text{Var}(Y_1)$ is 10 times as large as $\text{Var}(\bar{Y}) = \sigma^2/10$. This gives us a formal way of excluding Y_1 as an estimator of μ.

To emphasize this point, Table C.1 contains the outcome of a small simulation study. Using the statistical package Stata®, 20 random samples of size 10 were generated from a normal distribution, with $\mu = 2$ and $\sigma^2 = 1$; we are interested in estimating μ here. For each of the 20 random samples, we compute two estimates, y_1 and $\bar{y}$; these values are listed in Table C.1. As can be seen from the table, the values for y_1 are much more spread out than those for $\bar{y}$: y_1 ranges from -0.64 to 4.27, while $\bar{y}$ ranges only from 1.16 to 2.58. Further, in 16 out of 20 cases, $\bar{y}$ is closer than y_1 to $\mu = 2$. The average of y_1 across the simulations is about 1.89, while that for $\bar{y}$ is 1.96. The fact that these averages are close to 2 illustrates the unbiasedness of both estimators (and we could get these averages closer to 2 by doing more than 20 replications). But comparing just the average outcomes across random draws masks the fact that the sample average $\bar{Y}$ is far superior to Y_1 as an estimator of μ.

C-2d Efficiency

Comparing the variances of $\bar{Y}$ and Y_1 in the previous subsection is an example of a general approach to comparing different unbiased estimators.

Relative Efficiency. If W_1 and W_2 are two unbiased estimators of θ, W_1 is efficient relative to W_2 when $\text{Var}(W_1) \leq \text{Var}(W_2)$ for all θ, with strict inequality for at least one value of θ.

TABLE C.1 Simulation of Estimators for a Normal(μ, 1) Distribution with $\mu = 2$

Replication	y_1	$\bar{y}$
1	−0.64	1.98
2	1.06	1.43
3	4.27	1.65
4	1.03	1.88
5	3.16	2.34
6	2.77	2.58
7	1.68	1.58
8	2.98	2.23
9	2.25	1.96
10	2.04	2.11
11	0.95	2.15
12	1.36	1.93
13	2.62	2.02
14	2.97	2.10
15	1.93	2.18
16	1.14	2.10
17	2.08	1.94
18	1.52	2.21
19	1.33	1.16
20	1.21	1.75

Earlier, we showed that, for estimating the population mean μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ for any value of σ^2 whenever $n > 1$. Thus, $\bar{Y}$ is efficient relative to Y_1 for estimating μ. We cannot always choose between unbiased estimators based on the smallest variance criterion: given two unbiased estimators of θ, one can have smaller variance from some values of θ, while the other can have smaller variance for other values of θ.

If we restrict our attention to a certain class of estimators, we can show that the sample average has the smallest variance. Problem C.2 asks you to show that $\bar{Y}$ has the smallest variance among all unbiased estimators that are also linear functions of $Y_1, Y_2, \ldots, Y_n$. The assumptions are that the Y_i have common mean and variance, and that they are pairwise uncorrelated.

If we do not restrict our attention to unbiased estimators, then comparing variances is meaningless. For example, when estimating the population mean μ, we can use a trivial estimator that is equal to zero, regardless of the sample that we draw. Naturally, the variance of this estimator is zero (because it is the same value for every random sample). But the bias of this estimator is $-\mu$, so it is a very poor estimator when $|\mu|$ is large.

One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If W is an estimator of θ, then the MSE of W is defined as $\text{MSE}(W) = \text{E}[(W - \theta)^2]$. The MSE measures how far, on average, the estimator is away from θ. It can be shown that $\text{MSE}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$, so that MSE($W$) depends on the variance and bias (if any is present). This allows us to compare two estimators when one or both are biased.

C-3 Asymptotic or Large Sample Properties of Estimators

In Section C-2, we encountered the estimator Y_1 for the population mean μ, and we saw that, even though it is unbiased, it is a poor estimator because its variance can be much larger than that of the sample mean. One notable feature of Y_1 is that it has the same variance for any sample size. It seems reasonable to require any estimation procedure to improve as the sample size increases. For estimating a population mean μ, $\overline{Y}$ improves in the sense that its variance gets smaller as n gets larger; Y_1 does not improve in this sense.

We can rule out certain silly estimators by studying the *asymptotic* or *large sample* properties of estimators. In addition, we can say something positive about estimators that are not unbiased and whose variances are not easily found.

Asymptotic analysis involves approximating the features of the sampling distribution of an estimator. These approximations depend on the size of the sample. Unfortunately, we are necessarily limited in what we can say about how "large" a sample size is needed for asymptotic analysis to be appropriate; this depends on the underlying population distribution. But large sample approximations have been known to work well for sample sizes as small as $n = 20$.

C-3a Consistency

The first asymptotic property of estimators concerns how far the estimator is likely to be from the parameter it is supposed to be estimating as we let the sample size increase indefinitely.

Consistency. Let W_n be an estimator of θ based on a sample $Y_1, Y_2, \ldots, Y_n$ of size n. Then, W_n is a **consistent estimator** of θ if for every $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty. \qquad \text{[C.7]}$$

If W_n is not consistent for θ, then we say it is **inconsistent**.

When W_n is consistent, we also say that θ is the **probability limit** of W_n, written as $\text{plim}(W_n) = \theta$.

Unlike unbiasedness—which is a feature of an estimator for a given sample size—consistency involves the behavior of the sampling distribution of the estimator as the sample size n gets large. To emphasize this, we have indexed the estimator by the sample size in stating this definition, and we will continue with this convention throughout this section.

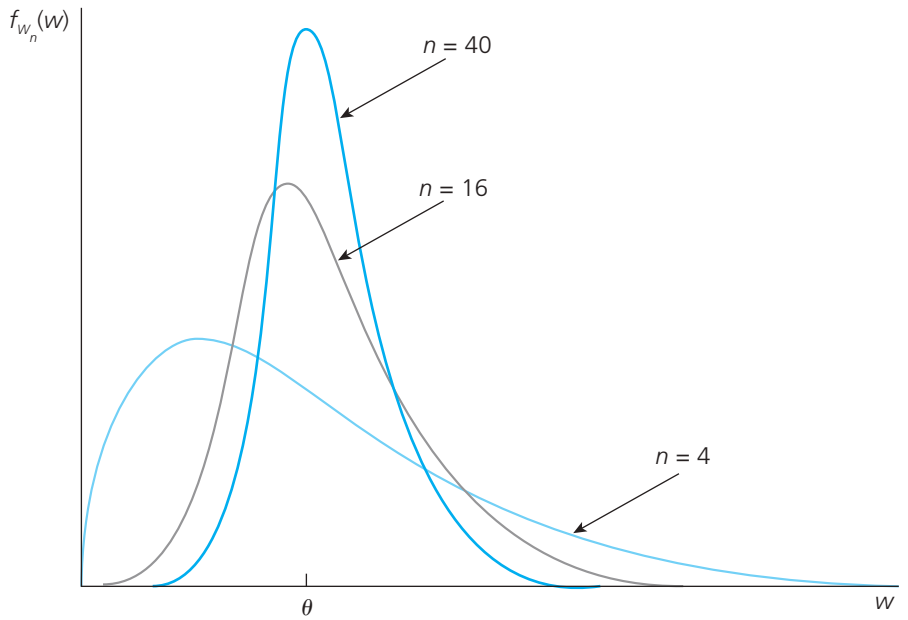
Equation (C.7) looks technical, and it can be rather difficult to establish based on fundamental probability principles. By contrast, interpreting (C.7) is straightforward. It means that the distribution of W_n becomes more and more concentrated about θ, which roughly means that for larger sample sizes, W_n is less and less likely to be very far from θ. This tendency is illustrated in Figure C.3.

If an estimator is not consistent, then it does not help us to learn about θ, even with an unlimited amount of data. For this reason, consistency is a minimal requirement of an estimator used in statistics or econometrics. We will encounter estimators that are consistent under certain assumptions and inconsistent when those assumptions fail. When estimators are inconsistent, we can usually find their probability limits, and it will be important to know how far these probability limits are from θ.

As we noted earlier, unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows *are* consistent. This can be stated formally: If W_n is an unbiased estimator of θ and $\text{Var}(W_n) \to 0$ as $n \to \infty$, then $\text{plim}(W_n) = \theta$. Unbiased estimators that use the entire data sample will usually have a variance that shrinks to zero as the sample size grows, thereby being consistent.

A good example of a consistent estimator is the average of a random sample drawn from a population with mean μ and variance σ^2. We have already shown that the sample average is unbiased for μ.

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



In Equation (C.6), we derived $\text{Var}(\bar{Y}_n) = \sigma^2/n$ for any sample size n. Therefore, $\text{Var}(\bar{Y}_n) \to 0$ as $n \to \infty$, so $\bar{Y}_n$ is a consistent estimator of μ (in addition to being unbiased).

The conclusion that $\bar{Y}_n$ is consistent for μ holds even if $\text{Var}(\bar{Y}_n)$ does not exist. This classic result is known as the **law of large numbers (LLN)**.

Law of Large Numbers. Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}(\bar{Y}_n) = \mu. \qquad \text{[C.8]}$$

The law of large numbers means that, if we are interested in estimating the population average μ, we can get arbitrarily close to μ by choosing a sufficiently large sample. This fundamental result can be combined with basic properties of plims to show that fairly complicated estimators are consistent.

Property PLIM.1: Let θ be a parameter and define a new parameter, $\gamma = g(\theta)$, for some *continuous function* $g(\theta)$. Suppose that $\text{plim}(W_n) = \theta$. Define an estimator of γ by $G_n = g(W_n)$. Then,

$$\text{plim}(G_n) = \gamma. \qquad \text{[C.9]}$$

This is often stated as

$$\text{plim}\, g(W_n) = g(\text{plim}\, W_n) \qquad \text{[C.10]}$$

for a continuous function $g(\theta)$.

The assumption that $g(\theta)$ is continuous is a technical requirement that has often been described nontechnically as "a function that can be graphed without lifting your pencil from the paper." Because all the functions we encounter in this text are continuous, we do not provide a formal definition of a continuous function. Examples of continuous functions are $g(\theta) = a + b\theta$ for constants a and b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, and many variants on these. We will not need to mention the continuity assumption again.

As an important example of a consistent but biased estimator, consider estimating the standard deviation, σ, from a population with mean μ and variance σ^2. We already claimed that the sample variance $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$ is unbiased for σ^2. Using the law of large numbers and some algebra, S_n^2 can also be shown to be consistent for σ^2. The natural estimator of $\sigma = \sqrt{\sigma^2}$ is $S_n = \sqrt{S_n^2}$ (where the square root is always the positive square root). S_n, which is called the **sample standard deviation**, is *not* an unbiased estimator because the expected value of the square root is *not* the square root of the expected value (see Section B-3). Nevertheless, by PLIM.1, $\text{plim}\, S_n = \sqrt{\text{plim}\, S_n^2} = \sqrt{\sigma^2} = \sigma$, so S_n is a consistent estimator of σ.

Here are some other useful properties of the probability limit:

Property PLIM.2: If $\text{plim}(T_n) = \alpha$ and $\text{plim}(U_n) = \beta$, then

(i) $\text{plim}(T_n + U_n) = \alpha + \beta$;
(ii) $\text{plim}(T_n U_n) = \alpha\beta$;
(iii) $\text{plim}(T_n/U_n) = \alpha/\beta$, provided $\beta \neq 0$.

These three facts about probability limits allow us to combine consistent estimators in a variety of ways to get other consistent estimators. For example, let $\{Y_1, \ldots, Y_n\}$ be a random sample of size n on annual earnings from the population of workers with a high school education and denote the population mean by μ_Y. Let $\{Z_1, \ldots, Z_n\}$ be a random sample on annual earnings from the population of workers with a college education and denote the population mean by μ_Z. We wish to estimate the percentage difference in annual earnings between the two groups, which is $\gamma = 100\cdot(\mu_Z - \mu_Y)/\mu_Y$. (This is the percentage by which average earnings for college graduates differs from average earnings for high school graduates.) Because $\bar{Y}_n$ is consistent for μ_Y and $\bar{Z}_n$ is consistent for μ_Z, it follows from PLIM.1 and part (iii) of PLIM.2 that

$$G_n \equiv 100\cdot(\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

is a consistent estimator of γ. G_n is just the percentage difference between $\bar{Z}_n$ and $\bar{Y}_n$ in the sample, so it is a natural estimator. G_n is not an unbiased estimator of γ, but it is still a good estimator except possibly when n is small.

C-3b Asymptotic Normality

Consistency is a property of point estimators. Although it does tell us that the distribution of the estimator is collapsing around the parameter as the sample size gets large, it tells us essentially nothing about the *shape* of that distribution for a given sample size. For constructing interval estimators and testing hypotheses, we need a way to approximate the distribution of our estimators. Most econometric estimators have distributions that are well approximated by a normal distribution for large samples, which motivates the following definition.

Asymptotic Normality. Let $\{Z_n: n = 1, 2, \ldots\}$ be a sequence of random variables, such that for all numbers z,

$$P(Z_n \leq z) \to \Phi(z) \text{ as } n \to \infty, \qquad \text{[C.11]}$$

where $\Phi(z)$ is the standard normal cumulative distribution function. Then, Z_n is said to have an *asymptotic standard normal distribution*. In this case, we often write $Z_n \overset{a}{\sim} \text{Normal}(0, 1)$. (The "$a$" above the tilde stands for "asymptotically" or "approximately.")

Property (C.11) means that the cumulative distribution function for Z_n gets closer and closer to the cdf of the standard normal distribution as the sample size n gets large. When **asymptotic normality** holds, for large n we have the approximation $P(Z_n \leq z) \approx \Phi(z)$. Thus, probabilities concerning Z_n can be approximated by standard normal probabilities.

The **central limit theorem (CLT)** is one of the most powerful results in probability and statistics. It states that the average from a random sample for *any* population (with finite variance), when standardized, has an asymptotic standard normal distribution.

Central Limit Theorem. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample with mean μ and variance σ^2. Then,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma} \qquad [C.12]$$

has an asymptotic standard normal distribution.

The variable Z_n in (C.12) is the standardized version of $\bar{Y}_n$: we have subtracted off $\mathrm{E}(\bar{Y}_n) = \mu$ and divided by $\mathrm{sd}(\bar{Y}_n) = \sigma/\sqrt{n}$. Thus, regardless of the population distribution of Y, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution. Remarkably, the entire distribution of Z_n gets arbitrarily close to the standard normal distribution as n gets large.

The second equality in equation (C.12) expresses the standardized variable as $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, which shows that we must multiply the difference between the sample mean and the population mean by the square root of the sample size in order to obtain a useful limiting distribution. Without the multiplication by $\sqrt{n}$, we would just have $(\bar{Y}_n - \mu)/\sigma$, which converges in probability to zero. In other words, the distribution of $(\bar{Y}_n - \mu)/\sigma$ simply collapses to a single point as $n \to \infty$, which we know cannot be a good approximation to the distribution of $(\bar{Y}_n - \mu)/\sigma$ for reasonable sample sizes. Multiplying by $\sqrt{n}$ ensures that the variance of Z_n remains constant. Practically, we often treat $\bar{Y}_n$ as being approximately normally distributed with mean μ and variance σ^2/n, and this gives us the correct statistical procedures because it leads to the standardized variable in equation (C.12).

Most estimators encountered in statistics and econometrics can be written as functions of sample averages, in which case we can apply the law of large numbers and the central limit theorem. When two consistent estimators have asymptotic normal distributions, we choose the estimator with the smallest asymptotic variance.

In addition to the standardized sample average in (C.12), many other statistics that depend on sample averages turn out to be asymptotically normal. An important one is obtained by replacing σ with its consistent estimator S_n in equation (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \qquad [C.13]$$

also has an approximate standard normal distribution for large n. The exact (finite sample) distributions of (C.12) and (C.13) are definitely not the same, but the difference is often small enough to be ignored for large n.

Throughout this section, each estimator has been subscripted by n to emphasize the nature of asymptotic or large sample analysis. Continuing this convention clutters the notation without providing additional insight, once the fundamentals of asymptotic analysis are understood. Henceforth, we drop the n subscript and rely on you to remember that estimators depend on the sample size, and properties such as consistency and asymptotic normality refer to the growth of the sample size without bound.

C-4 General Approaches to Parameter Estimation

Until this point, we have used the sample average to illustrate the finite and large sample properties of estimators. It is natural to ask: Are there general approaches to estimation that produce estimators with good properties, such as unbiasedness, consistency, and efficiency?

The answer is yes. A detailed treatment of various approaches to estimation is beyond the scope of this text; here, we provide only an informal discussion. A thorough discussion is given in Larsen and Marx (1986, Chapter 5).

C-4a Method of Moments

Given a parameter θ appearing in a population distribution, there are usually many ways to obtain unbiased and consistent estimators of θ. Trying all different possibilities and comparing them on the basis of the criteria in Sections C-2 and C-3 is not practical. Fortunately, some methods have been shown to have good general properties, and, for the most part, the logic behind them is intuitively appealing.

In the previous sections, we have studied the sample average as an unbiased estimator of the population average and the sample variance as an unbiased estimator of the population variance. These estimators are examples of **method of moments** estimators. Generally, method of moments estimation proceeds as follows. The parameter θ is shown to be related to some expected value in the distribution of Y, usually $E(Y)$ or $E(Y^2)$ (although more exotic choices are sometimes used). Suppose, for example, that the parameter of interest, θ, is related to the population mean as $\theta = g(\mu)$ for some function g. Because the sample average $\bar{Y}$ is an unbiased and consistent estimator of μ, it is natural to replace μ with $\bar{Y}$, which gives us the estimator $g(\bar{Y})$ of θ. The estimator $g(\bar{Y})$ is consistent for θ, and if $g(\mu)$ is a linear function of μ, then $g(\bar{Y})$ is unbiased as well. What we have done is replace the population moment, μ, with its sample counterpart, $\bar{Y}$. This is where the name "method of moments" comes from.

We cover two additional method of moments estimators that will be useful for our discussion of regression analysis. Recall that the covariance between two random variables X and Y is defined as $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. The method of moments suggests estimating σ_{XY} by $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. This is a consistent estimator of σ_{XY}, but it turns out to be biased for essentially the same reason that the sample variance is biased if n, rather than $n - 1$, is used as the divisor. The **sample covariance** is defined as

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \qquad \text{[C.14]}$$

It can be shown that this is an unbiased estimator of σ_{XY}. (Replacing n with $n - 1$ makes no difference as the sample size grows indefinitely, so this estimator is still consistent.)

As we discussed in Section B-4, the covariance between two variables is often difficult to interpret. Usually, we are more interested in correlation. Because the population correlation is $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, the method of moments suggests estimating ρ_{XY} as

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \qquad \text{[C.15]}$$

which is called the **sample correlation coefficient** (or *sample correlation* for short). Notice that we have canceled the division by $n - 1$ in the sample covariance and the sample standard deviations. In fact, we could divide each of these by n, and we would arrive at the same final formula.

It can be shown that the sample correlation coefficient is always in the interval $[-1,1]$, as it should be. Because S_{XY}, S_X, and S_Y are consistent for the corresponding population parameter, R_{XY} is a consistent estimator of the population correlation, ρ_{XY}. However, R_{XY} is a biased estimator for two reasons. First, S_X and S_Y are biased estimators of σ_X and σ_Y, respectively. Second, R_{XY} is a ratio of estimators, so it would not be unbiased, even if S_X and S_Y were. For our purposes, this is not important, although the fact that no unbiased estimator of ρ_{XY} exists is a classical result in mathematical statistics.

C-4b Maximum Likelihood

Another general approach to estimation is the method of *maximum likelihood*, a topic covered in many introductory statistics courses. A brief summary in the simplest case will suffice here. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample from the population distribution $f(y; \theta)$. Because of the random

sampling assumption, the joint distribution of $\{Y_1, Y_2, \ldots, Y_n\}$ is simply the product of the densities: $f(y_1; \theta)f(y_2; \theta) \cdots f(y_n; \theta)$. In the discrete case, this is $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$. Now, define the *likelihood function* as

$$L(\theta; Y_1, Y_2, \ldots, Y_n) = f(Y_1; \theta)f(Y_2; \theta) \cdot \cdots \cdot f(Y_n; \theta),$$

which is a random variable because it depends on the outcome of the random sample $\{Y_1, Y_2, \ldots, Y_n\}$. The **maximum likelihood estimator** of θ, call it W, is the value of θ that maximizes the likelihood function. (This is why we write L as a function of θ, followed by the random sample.) Clearly, this value depends on the random sample. The maximum likelihood principle says that, out of all the possible values for θ, the value that makes the likelihood of the observed data largest should be chosen. Intuitively, this is a reasonable approach to estimating θ.

Usually, it is more convenient to work with the **log-likelihood function**, which is obtained by taking the natural log of the likelihood function:

$$\mathcal{L}(\theta) = \log[L(\theta; Y_1, Y_2, \ldots, Y_n)] = \sum_{i=1}^{n} \log[f(Y_i; \theta)] = \sum_{i=1}^{n} \ell(\theta; X_i), \qquad \text{[C.16]}$$

where we use the fact that the log of the product is the sum of the logs. The function $\ell(\theta; X_i) = \log[f(Y_i; \theta)]$ is the log-likelihood function for random draw i. Because (C.16) is the sum of independent, identically distributed random variables, analyzing estimators that come from (C.16) is relatively easy.

Maximum likelihood estimation (MLE) is usually consistent and sometimes unbiased. But so are many other estimators. The widespread appeal of MLE is that it is generally the most asymptotically efficient estimator when the population model $f(y; \theta)$ is correctly specified. In addition, the MLE is sometimes the **minimum variance unbiased estimator**; that is, it has the smallest variance among all unbiased estimators of θ. [See Larsen and Marx (1986, Chapter 5) for verification of these claims.]

In Chapter 17, we will need maximum likelihood to estimate the parameters of more advanced econometric models. In econometrics, we are almost always interested in the distribution of Y conditional on a set of explanatory variables, say, $X_1, X_2, \ldots, X_k$. Then, we replace the density in (C.16) with $f(Y_i|X_{i1}, \ldots, X_{ik}; \theta_1, \ldots, \theta_p)$, where this density is allowed to depend on p parameters, $\theta_1, \ldots, \theta_p$. Fortunately, for successful application of maximum likelihood methods, we do not need to delve much into the computational issues or the large-sample statistical theory. Wooldridge (2010, Chapter 13) covers the theory of MLE.

C-4c Least Squares

A third kind of estimator, and one that plays a major role throughout the text, is called a **least squares estimator**. We have already seen an example of least squares: the sample mean, $\overline{Y}$, is a least squares estimator of the population mean, μ. We already know $\overline{Y}$ is a method of moments estimator. What makes it a least squares estimator? It can be shown that the value of m that makes the sum of squared deviations

$$\sum_{i=1}^{n} (Y_i - m)^2$$

as small as possible is $m = \overline{Y}$. Showing this is not difficult, but we omit the algebra.

For some important distributions, including the normal and the Bernoulli, the sample average $\overline{Y}$ is also the maximum likelihood estimator of the population mean μ. Thus, the principles of least squares, method of moments, and maximum likelihood often result in the *same* estimator. In other cases, the estimators are similar but not identical.

C-5 Interval Estimation and Confidence Intervals

C-5a The Nature of Interval Estimation

A point estimate obtained from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but, by its nature, it provides no information about how close the estimate is "likely" to be to the population parameter. As an example, suppose a researcher reports, on the basis of a random sample of workers, that job training grants increase hourly wage by 6.4%. How are we to know whether or not this is close to the effect in the population of workers who could have been trained? Because we do not know the population value, we cannot know how close an estimate is for a particular sample. However, we can make statements involving probabilities, and this is where interval estimation comes in.

We already know one way of assessing the uncertainty in an estimator: find its sampling standard deviation. Reporting the standard deviation of the estimator, along with the point estimate, provides some information on the accuracy of our estimate. However, even if the problem of the standard deviation's dependence on unknown population parameters is ignored, reporting the standard deviation along with the point estimate makes no direct statement about where the population value is likely to lie in relation to the estimate. This limitation is overcome by constructing a **confidence interval**.

We illustrate the concept of a confidence interval with an example. Suppose the population has a Normal$(\mu, 1)$ distribution and let $\{Y_1, \ldots, Y_n\}$ be a random sample from this population. (We assume that the variance of the population is known and equal to unity for the sake of illustration; we then show what to do in the more realistic case that the variance is unknown.) The sample average, $\overline{Y}$, has a normal distribution with mean μ and variance $1/n$: $\overline{Y} \sim \text{Normal}(\mu, 1/n)$. From this, we can standardize $\overline{Y}$, and, because the standardized version of $\overline{Y}$ has a standard normal distribution, we have

$$\text{P}\left(-1.96 < \frac{\overline{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = .95.$$

The event in parentheses is identical to the event $\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96/\sqrt{n}$, so

$$\text{P}(\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96\sqrt{n}) = .95. \qquad \text{[C.17]}$$

Equation (C.17) is interesting because it tells us that the probability that the random interval $[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}]$ contains the population mean μ is .95, or 95%. This information allows us to construct an *interval estimate* of μ, which is obtained by plugging in the sample outcome of the average, $\bar{y}$. Thus,

$$[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}] \qquad \text{[C.18]}$$

is an example of an interval estimate of μ. It is also called a 95% confidence interval. A shorthand notation for this interval is $\bar{y} \pm 1.96/\sqrt{n}$.

The confidence interval in equation (C.18) is easy to compute, once the sample data $\{y_1, y_2, \ldots, y_n\}$ are observed; $\bar{y}$ is the only factor that depends on the data. For example, suppose that $n = 16$ and the average of the 16 data points is 7.3. Then, the 95% confidence interval for μ is $7.3 \pm 1.96/\sqrt{16} = 7.3 \pm .49$, which we can write in interval form as [6.81,7.79]. By construction, $\bar{y} = 7.3$ is in the center of this interval.

Unlike its computation, the meaning of a confidence interval is more difficult to understand. When we say that equation (C.18) is a 95% confidence interval for μ, we mean that the *random* interval

$$[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}] \qquad \text{[C.19]}$$

contains μ with probability .95. In other words, *before* the random sample is drawn, there is a 95% chance that (C.19) contains μ. Equation (C.19) is an example of an **interval estimator**. It is a random interval, because the endpoints change with different samples.

A confidence interval is often interpreted as follows: "The probability that μ is in the interval (C.18) is .95." This is incorrect. Once the sample has been observed and $\bar{y}$ has been computed, the limits of the confidence interval are simply numbers (6.81 and 7.79 in the example just given). The population parameter, μ, though unknown, is also just some number. Therefore, μ either is or is not in the interval (C.18) (and we will never know with certainty which is the case). Probability plays no role once the confidence interval is computed for the particular data at hand. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ.

To emphasize the meaning of a confidence interval, Table C.2 contains calculations for 20 random samples (or replications) from the Normal(2,1) distribution with sample size $n = 10$. For each of the 20 samples, $\bar{y}$ is obtained, and (C.18) is computed as $\bar{y} \pm 1.96/\sqrt{10} = \bar{y} \pm .62$ (each rounded to two decimals). As you can see, the interval changes with each random sample. Nineteen of the 20 intervals contain the population value of μ. Only for replication number 19 is μ not in the confidence interval. In other words, 95% of the samples result in a confidence interval that contains μ. This did not have to be the case with only 20 replications, but it worked out that way for this particular simulation.

TABLE C.2 Simulated Confidence Intervals from a Normal(μ, 1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

C-5b Confidence Intervals for the Mean from a Normally Distributed Population

The confidence interval derived in equation (C.18) helps illustrate how to construct and interpret confidence intervals. In practice, equation (C.18) is not very useful for the mean of a normal population because it assumes that the variance is known to be unity. It is easy to extend (C.18) to the case where the standard deviation σ is known to be any value: the 95% confidence interval is

$$[\bar{y} - 1.96\sigma/\sqrt{n}, \bar{y} + 1.96\sigma/\sqrt{n}]. \qquad [C.20]$$

Therefore, provided σ is known, a confidence interval for μ is readily constructed. To allow for unknown σ, we must use an estimate. Let

$$s = \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2} \qquad [C.21]$$

denote the sample standard deviation. Then, we obtain a confidence interval that depends entirely on the observed data by replacing σ in equation (C.20) with its estimate, s. Unfortunately, this does not preserve the 95% level of confidence because s depends on the particular sample. In other words, the random interval $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ no longer contains μ with probability .95 because the constant σ has been replaced with the random variable S.

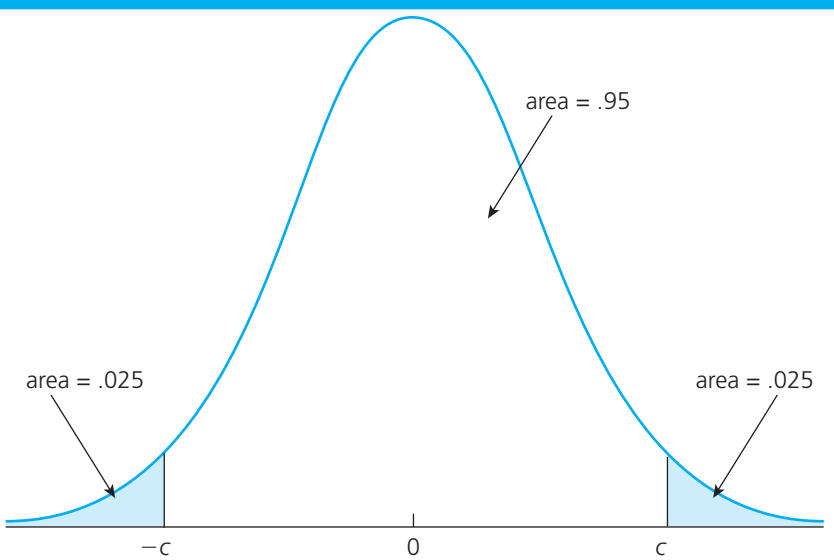
How should we proceed? Rather than using the standard normal distribution, we must rely on the t distribution. The t distribution arises from the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \qquad [C.22]$$

where $\bar{Y}$ is the sample average and S is the sample standard deviation of the random sample $\{Y_1, \ldots, Y_n\}$. We will not prove (C.22); a careful proof can be found in a variety of places [for example, Larsen and Marx (1986, Chapter 7)].

To construct a 95% confidence interval, let c denote the 97.5th percentile in the t_{n-1} distribution. In other words, c is the value such that 95% of the area in the t_{n-1} is between $-c$ and c: $P(-c < t_{n-1} < c) = .95$. (The value of c depends on the degrees of freedom $n - 1$, but we do not

FIGURE C.4 The 97.5th percentile, c, in a t distribution.



make this explicit.) The choice of c is illustrated in Figure C.4. Once c has been properly chosen, the random interval $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ contains μ with probability .95. For a particular sample, the 95% confidence interval is calculated as

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \qquad [C.23]$$

The values of c for various degrees of freedom can be obtained from Table G.2 in Statistical Tables. For example, if $n = 20$, so that the df is $n - 1 = 19$, then $c = 2.093$. Thus, the 95% confidence interval is $[\bar{y} \pm 2.093(s/\sqrt{20})]$, where $\bar{y}$ and s are the values obtained from the sample. Even if $s = \sigma$ (which is very unlikely), the confidence interval in (C.23) is wider than that in (C.20) because $c > 1.96$. For small degrees of freedom, (C.23) is much wider.

More generally, let c_α denote the $100(1 - \alpha)$ percentile in the t_{n-1} distribution. Then, a $100(1 - \alpha)\%$ confidence interval is obtained as

$$[\bar{y} - c_{\alpha/2} s/\sqrt{n}, \bar{y} + c_{\alpha/2} s/\sqrt{n}]. \qquad [C.24]$$

Obtaining $c_{\alpha/2}$ requires choosing α and knowing the degrees of freedom $n - 1$; then, Table G.2 can be used. For the most part, we will concentrate on 95% confidence intervals.

There is a simple way to remember how to construct a confidence interval for the mean of a normal distribution. Recall that $\text{sd}(\bar{Y}) = \sigma/\sqrt{n}$. Thus, $s/\sqrt{n}$ is the point estimate of $\text{sd}(\bar{Y})$. The associated random variable, $S/\sqrt{n}$, is sometimes called the **standard error** of $\bar{Y}$. Because what shows up in formulas is the point estimate $s/\sqrt{n}$, we define the standard error of $\bar{y}$ as $\text{se}(\bar{y}) = s/\sqrt{n}$. Then, (C.24) can be written in shorthand as

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{se}(\bar{y})]. \qquad [C.25]$$

This equation shows why the notion of the standard error of an estimate plays an important role in econometrics.

EXAMPLE C.2 Effect of Job Training Grants on Worker Productivity

Holzer, Block, Cheatham, and Knott (1993) studied the effects of job training grants on worker productivity by collecting information on "scrap rates" for a sample of Michigan manufacturing firms receiving job training grants in 1988. Table C.3 lists the scrap rates—measured as number of items per 100 produced that are not usable and therefore need to be scrapped—for 20 firms. Each of these firms received a job training grant in 1988; there were no grants awarded in 1987. We are interested in constructing a confidence interval for the change in the scrap rate from 1987 to 1988 for the population of all manufacturing firms that could have received grants.

We assume that the change in scrap rates has a normal distribution. Because $n = 20$, a 95% confidence interval for the mean change in scrap rates μ is $[\bar{y} \pm 2.093 \cdot \text{se}(\bar{y})]$, where $\text{se}(\bar{y}) = s/\sqrt{n}$. The value 2.093 is the 97.5^{th} percentile in a t_{19} distribution. For the particular sample values, $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$ (each rounded to two decimals), so the 95% confidence interval is $[-2.28, -.02]$. The value zero is excluded from this interval, so we conclude that, with 95% confidence, the average change in scrap rates in the population is not zero.

TABLE C.3 Scrap Rates for 20 Michigan Manufacturing Firms

Firm	1987	1988	Change
1	10	3	−7
2	1	1	0
3	6	5	−1
4	.45	.5	.05
5	1.25	1.54	.29
6	1.3	1.5	.2
7	1.06	.8	−.26
8	3	2	−1
9	8.18	.67	−7.51
10	1.67	1.17	−.5
11	.98	.51	−.47
12	1	.5	−.5
13	.45	.61	.16
14	5.03	6.7	1.67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	.28	.2	−.08
19	7	5	−2
20	3.97	3.83	−.14
Average	4.38	3.23	−1.15

At this point, Example C.2 is mostly illustrative because it has some potentially serious flaws as an econometric analysis. Most importantly, it assumes that any systematic reduction in scrap rates is due to the job training grants. But many things can happen over the course of the year to change worker productivity. From this analysis, we have no way of knowing whether the fall in average scrap rates is attributable to the job training grants or if, at least partly, some external force is responsible.

C-5c A Simple Rule of Thumb for a 95% Confidence Interval

The confidence interval in (C.25) can be computed for any sample size and any confidence level. As we saw in Section B-5, the t distribution approaches the standard normal distribution as the degrees of freedom gets large. In particular, for $\alpha = .05$, $c_{\alpha/2} \to 1.96$ as $n \to \infty$, although $c_{\alpha/2}$ is always greater than 1.96 for each n. A *rule of thumb* for an approximate 95% confidence interval is

$$[\bar{y} \pm 2 \cdot \text{se}(\bar{y})]. \qquad \text{[C.26]}$$

In other words, we obtain $\bar{y}$ and its standard error and then compute $\bar{y}$ plus or minus twice its standard error to obtain the confidence interval. This is slightly too wide for very large n, and it is too narrow for small n. As we can see from Example C.2, even for n as small as 20, (C.26) is in the ballpark for a 95% confidence interval for the mean from a normal distribution. This means we can get pretty close to a 95% confidence interval without having to refer to t tables.

C-5d Asymptotic Confidence Intervals for Nonnormal Populations

In some applications, the population is clearly nonnormal. A leading case is the Bernoulli distribution, where the random variable takes on only the values zero and one. In other cases, the nonnormal population has no standard distribution. This does not matter, provided the sample size is sufficiently large for the central limit theorem to give a good approximation for the distribution of the sample average $\bar{Y}$. For large n, an *approximate* 95% confidence interval is

$$[\bar{y} \pm 1.96 \cdot \text{se}(\bar{y})], \qquad \text{[C.27]}$$

where the value 1.96 is the 97.5$^{\text{th}}$ percentile in the standard normal distribution. Mechanically, computing an approximate confidence interval does not differ from the normal case. A slight difference is that the number multiplying the standard error comes from the standard normal distribution, rather than the t distribution, because we are using asymptotics. Because the t distribution approaches the standard normal as the df increases, equation (C.25) is also perfectly legitimate as an approximate 95% interval; some prefer this to (C.27) because the former is exact for normal populations.

EXAMPLE C.3 Race Discrimination in Hiring

The Urban Institute conducted a study in 1988 in Washington, D.C., to examine the extent of race discrimination in hiring. Five pairs of people interviewed for several jobs. In each pair, one person was black and the other person was white. They were given résumés indicating that they were virtually the same in terms of experience, education, and other factors that determine job qualification. The idea was to make individuals as similar as possible with the exception of race. Each person in a pair interviewed for the same job, and the researchers recorded which applicant received a job offer. This is an example of a *matched pairs analysis*, where each trial consists of data on two people (or two firms, two cities, and so on) that are thought to be similar in many respects but different in one important characteristic.

Let θ_B denote the probability that the black person is offered a job and let θ_W be the probability that the white person is offered a job. We are primarily interested in the difference, $\theta_B - \theta_W$. Let B_i denote a Bernoulli variable equal to one if the black person gets a job offer from employer i, and zero otherwise. Similarly, $W_i = 1$ if the white person gets a job offer from employer i, and zero otherwise. Pooling across the five pairs of people, there were a total of $n = 241$ trials (pairs of interviews with employers). Unbiased estimators of θ_B and θ_W are $\overline{B}$ and $\overline{W}$, the fractions of interviews for which blacks and whites were offered jobs, respectively.

To put this into the framework of computing a confidence interval for a population mean, define a new variable $Y_i = B_i - W_i$. Now, Y_i can take on three values: -1 if the black person did not get the job but the white person did, 0 if both people either did or did not get the job, and 1 if the black person got the job and the white person did not. Then, $\mu \equiv \text{E}(Y_i) = \text{E}(B_i) - \text{E}(W_i) = \theta_B - \theta_W$.

The distribution of Y_i is certainly not normal—it is discrete and takes on only three values. Nevertheless, an approximate confidence interval for $\theta_B - \theta_W$ can be obtained by using large sample methods.

The data from the Urban Institute audit study are in the file AUDIT. Using the 241 observed data points, $\overline{b} = .224$ and $\overline{w} = .357$, so $\bar{y} = .224 - .357 = -.133$. Thus, 22.4% of black applicants were offered jobs, while 35.7% of white applicants were offered jobs. This is *prima facie* evidence of discrimination against blacks, but we can learn much more by computing a confidence interval for μ. To compute an approximate 95% confidence interval, we need the sample standard deviation. This turns out to be $s = .482$ [using equation (C.21)]. Using (C.27), we obtain a 95% CI for $\mu = \theta_B - \theta_W$ as $-.133 \pm 1.96(.482/\sqrt{241}) = -.133 \pm .031 = [-.164, -.102]$. The approximate 99% CI is $-.133 \pm 2.58(.482/\sqrt{241}) = [-.213, -.053]$. Naturally, this contains a wider range of values than the 95% CI. But even the 99% CI does not contain the value zero. Thus, we are very confident that the population difference $\theta_B - \theta_W$ is not zero.

Before we turn to hypothesis testing, it is useful to review the various population and sample quantities that measure the spreads in the population distributions and the sampling distributions of the estimators. These quantities appear often in statistical analysis, and extensions of them are important for the regression analysis in the main text. The quantity σ is the (unknown) population standard deviation; it is a measure of the spread in the distribution of Y. When we divide σ by $\sqrt{n}$, we obtain the **sampling standard deviation** of $\overline{Y}$ (the sample average). While σ is a fixed feature of the population, $\text{sd}(\overline{Y}) = \sigma/\sqrt{n}$ shrinks to zero as $n \to \infty$: our estimator of μ gets more and more precise as the sample size grows.

The estimate of σ for a particular sample, s, is called the sample standard deviation because it is obtained from the sample. (We also call the underlying random variable, S, which changes across different samples, the sample standard deviation.) Like $\bar{y}$ as an estimate of μ, s is our "best guess" at σ given the sample at hand. The quantity $s/\sqrt{n}$ is what we call the standard error of $\bar{y}$, and it is our best estimate of $\sigma/\sqrt{n}$. Confidence intervals for the population parameter μ depend directly on $\text{se}(\bar{y}) = s/\sqrt{n}$. Because this standard error shrinks to zero as the sample size grows, a larger sample size generally means a smaller confidence interval. Thus, we see clearly that one benefit of more data is that they result in narrower confidence intervals. The notion of the standard error of an estimate, which in the vast majority of cases shrinks to zero at the rate $1/\sqrt{n}$, plays a fundamental role in hypothesis testing (as we will see in the next section) and for confidence intervals and testing in the context of multiple regression (as discussed in Chapter 4).

C-6 Hypothesis Testing

So far, we have reviewed how to evaluate point estimators, and we have seen—in the case of a population mean—how to construct and interpret confidence intervals. But sometimes the question we are interested in has a definite yes or no answer. Here are some examples: (1) Does a job training program effectively increase average worker productivity? (see Example C.2); (2) Are blacks discriminated against in hiring? (see Example C.3); (3) Do stiffer state drunk driving laws reduce the number of drunk driving arrests? Devising methods for answering such questions, using a sample of data, is known as hypothesis testing.

C-6a Fundamentals of Hypothesis Testing

To illustrate the issues involved with hypothesis testing, consider an election example. Suppose there are two candidates in an election, Candidates A and B. Candidate A is reported to have received 42% of the popular vote, while Candidate B received 58%. These are supposed to represent the true percentages in the voting population, and we treat them as such.

Candidate A is convinced that more people must have voted for him, so he would like to investigate whether the election was rigged. Knowing something about statistics, Candidate A hires a consulting agency to randomly sample 100 voters to record whether or not each person voted for him. Suppose that, for the sample collected, 53 people voted for Candidate A. This sample estimate of 53% clearly exceeds the reported population value of 42%. Should Candidate A conclude that the election was indeed a fraud?

While it appears that the votes for Candidate A were undercounted, we cannot be certain. Even if only 42% of the population voted for Candidate A, it is possible that, in a sample of 100, we observe 53 people who did vote for Candidate A. The question is: How *strong* is the sample evidence against the officially reported percentage of 42%?

One way to proceed is to set up a **hypothesis test**. Let θ denote the true proportion of the population voting for Candidate A. The hypothesis that the reported results are accurate can be stated as

$$H_0: \theta = .42 \qquad \text{[C.28]}$$

This is an example of a **null hypothesis**. We always denote the null hypothesis by H_0. In hypothesis testing, the null hypothesis plays a role similar to that of a defendant on trial in many judicial systems: just as a defendant is presumed to be innocent until proven guilty, the null hypothesis is presumed to be true until the data strongly suggest otherwise. In the current example, Candidate A must present fairly strong evidence against (C.28) in order to win a recount.

The **alternative hypothesis** in the election example is that the true proportion voting for Candidate A in the election is greater than .42:

$$H_1: \theta > .42. \qquad \text{[C.29]}$$

In order to conclude that H_0 is false and that H_1 is true, we must have evidence "beyond reasonable doubt" against H_0. How many votes out of 100 would be needed before we feel the evidence is strongly against H_0? Most would agree that observing 43 votes out of a sample of 100 is not enough to overturn the original election results; such an outcome is well within the expected sampling variation. On the other hand, we do not need to observe 100 votes for Candidate A to cast doubt on H_0. Whether 53 out of 100 is enough to reject H_0 is much less clear. The answer depends on how we quantify "beyond reasonable doubt."

Before we turn to the issue of quantifying uncertainty in hypothesis testing, we should head off some possible confusion. You may have noticed that the hypotheses in equations (C.28) and (C.29) do not exhaust all possibilities: it could be that θ is less than .42. For the application at hand, we are not particularly interested in that possibility; it has nothing to do with overturning the results of the election. Therefore, we can just state at the outset that we are ignoring alternatives θ with $\theta < .42$. Nevertheless, some authors prefer to state null and alternative hypotheses so that they are exhaustive, in which case our null hypothesis should be $H_0: \theta \leq .42$. Stated in this way, the null hypothesis is a *composite* null hypothesis because it allows for more than one value under H_0. [By contrast, equation (C.28) is an example of a *simple* null hypothesis.] For these kinds of examples, it does not matter whether we state the null as in (C.28) or as a composite null: the most difficult value to reject if $\theta \leq .42$ is $\theta = .42$. (That is, if we reject the value $\theta = .42$, against $\theta > .42$, then logically we must reject any value less than .42.) Therefore, our testing procedure based on (C.28) leads to the same test as if $H_0: \theta \leq .42$. In this text, we always state a null hypothesis as a simple null hypothesis.

In hypothesis testing, we can make two kinds of mistakes. First, we can reject the null hypothesis when it is in fact true. This is called a **Type I error**. In the election example, a Type I error occurs if we reject H_0 when the true proportion of people voting for Candidate A is in fact .42. The second kind of error is failing to reject H_0 when it is actually false. This is called a **Type II error**. In the election example, a Type II error occurs if $\theta > .42$ but we fail to reject H_0.

After we have made the decision of whether or not to reject the null hypothesis, we have either decided correctly or we have committed an error. We will never know with certainty whether an error was committed. However, we can compute the *probability* of making either a Type I or a Type II error. Hypothesis testing rules are constructed to make the probability of committing a Type I error fairly small. Generally, we define the **significance level** (or simply the *level*) of a test as the probability of a Type I error; it is typically denoted by α. Symbolically, we have

$$\alpha = \mathrm{P}(\text{Reject } H_0 | H_0). \qquad \text{[C.30]}$$

The right-hand side is read as: "The probability of rejecting H_0 given that H_0 is true."

Classical hypothesis testing requires that we initially specify a significance level for a test. When we specify a value for α, we are essentially quantifying our tolerance for a Type I error. Common values for α are .10, .05, and .01. If $\alpha = .05$, then the researcher is willing to falsely reject H_0 5% of the time, in order to detect deviations from H_0.

Once we have chosen the significance level, we would then like to minimize the probability of a Type II error. Alternatively, we would like to maximize the **power of a test** against all relevant alternatives. The power of a test is just one minus the probability of a Type II error. Mathematically,

$$\pi(\theta) = \mathrm{P}(\text{Reject } H_0 | \theta) = 1 - \mathrm{P}(\text{Type II} | \theta),$$

where θ denotes the actual value of the parameter. Naturally, we would like the power to equal unity whenever the null hypothesis is false. But this is impossible to achieve while keeping the significance level small. Instead, we choose our tests to maximize the power for a given significance level.

C-6b Testing Hypotheses about the Mean in a Normal Population

In order to test a null hypothesis against an alternative, we need to choose a test statistic (or statistic, for short) and a critical value. The choices for the statistic and critical value are based on convenience and on the desire to maximize power given a significance level for the test. In this subsection, we review how to test hypotheses for the mean of a normal population.

A **test statistic**, denoted T, is some function of the random sample. When we compute the statistic for a particular outcome, we obtain an outcome of the test statistic, which we will denote by t.

Given a test statistic, we can define a rejection rule that determines when H_0 is rejected in favor of H_1. In this text, all rejection rules are based on comparing the value of a test statistic, t, to a **critical value**, c. The values of t that result in rejection of the null hypothesis are collectively known as the **rejection region**. To determine the critical value, we must first decide on a significance level of the test. Then, given α, the critical value associated with α is determined by the distribution of T, *assuming* that H_0 is true. We will write this critical value as c, suppressing the fact that it depends on α.

Testing hypotheses about the mean μ from a Normal(μ, σ^2) population is straightforward. The null hypothesis is stated as

$$H_0: \mu = \mu_0, \tag{C.31}$$

where μ_0 is a value that we specify. In the majority of applications, $\mu_0 = 0$, but the general case is no more difficult.

The rejection rule we choose depends on the nature of the alternative hypothesis. The three alternatives of interest are

$$H_1: \mu > \mu_0, \tag{C.32}$$

$$H_1: \mu < \mu_0, \tag{C.33}$$

and

$$H_1: \mu \neq \mu_0. \tag{C.34}$$

Equation (C.32) gives a **one-sided alternative**, as does (C.33). When the alternative hypothesis is (C.32), the null is effectively $H_0: \mu \leq \mu_0$, because we reject H_0 only when $\mu > \mu_0$. This is appropriate when we are interested in the value of μ only when μ is at least as large as μ_0. Equation (C.34) is a **two-sided alternative**. This is appropriate when we are interested in any departure from the null hypothesis.

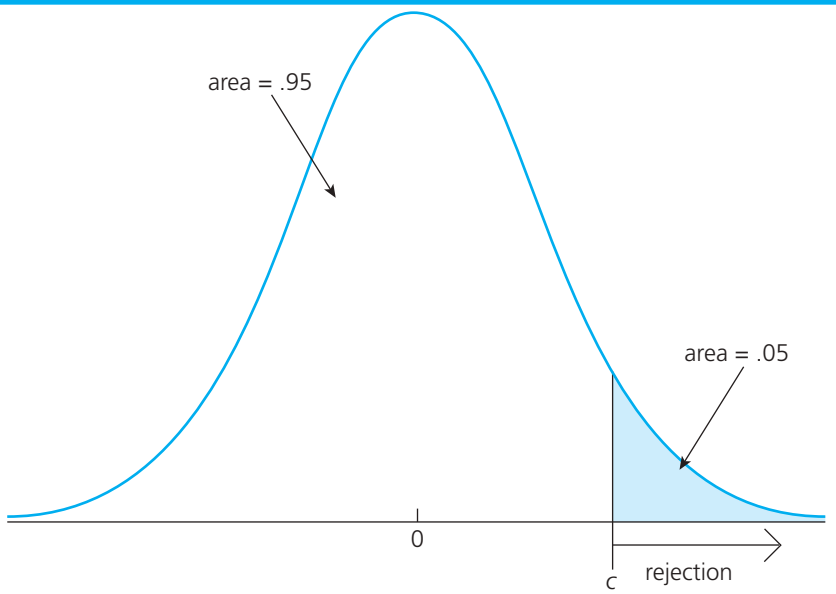
Consider first the alternative in (C.32). Intuitively, we should reject H_0 in favor of H_1 when the value of the sample average, $\bar{y}$, is "sufficiently" greater than μ_0. But how should we determine when $\bar{y}$ is large enough for H_0 to be rejected at the chosen significance level? This requires knowing the probability of rejecting the null hypothesis when it is true. Rather than working directly with $\bar{y}$, we use its standardized version, where σ is replaced with the sample standard deviation, s:

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{se}(\bar{y}), \tag{C.35}$$

where $\text{se}(\bar{y}) = s/\sqrt{n}$ is the standard error of $\bar{y}$. Given the sample of data, it is easy to obtain t. We work with t because, under the null hypothesis, the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

FIGURE C.5 Rejection region for a 5% significance level test against the one-sided alternative $\mu > \mu_0$.



has a t_{n-1} distribution. Now, suppose we have settled on a 5% significance level. Then, the critical value c is chosen so that $P(T > c|H_0) = .05$; that is, the probability of a Type I error is 5%. Once we have found c, the rejection rule is

$$t > c, \qquad [C.36]$$

where c is the $100(1 - \alpha)$ percentile in a t_{n-1} distribution; as a percent, the significance level is $100 \cdot \alpha\%$. This is an example of a **one-tailed test** because the rejection region is in one tail of the t distribution. For a 5% significance level, c is the 95th percentile in the t_{n-1} distribution; this is illustrated in Figure C.5. A different significance level leads to a different critical value.

The statistic in equation (C.35) is often called the ***t* statistic** for testing H_0: $\mu = \mu_0$. The t statistic measures the distance from $\bar{y}$ to μ_0 *relative* to the standard error of $\bar{y}$, $se(\bar{y})$.

EXAMPLE C.4 Effect of Enterprise Zones on Business Investments

In the population of cities granted enterprise zones in a particular state [see Papke (1994) for Indiana], let Y denote the percentage change in investment from the year before to the year after a city became an enterprise zone. Assume that Y has a Normal(μ, σ^2) distribution. The null hypothesis that enterprise zones have no effect on business investment is H_0: $\mu = 0$; the alternative that they have a positive effect is H_1: $\mu > 0$. (We assume that they do not have a negative effect.) Suppose that we wish to test H_0 at the 5% level. The test statistic in this case is

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{se(\bar{y})}. \qquad [C.37]$$

Suppose that we have a sample of 36 cities that are granted enterprise zones. Then, the critical value is $c = 1.69$ (see Table G.2), and we reject H_0 in favor of H_1 if $t > 1.69$. Suppose that the sample yields $\bar{y} = 8.2$ and $s = 23.9$. Then, $t \approx 2.06$, and H_0 is therefore rejected at the 5% level. Thus, we conclude

that, at the 5% significance level, enterprise zones have an effect on average investment. The 1% critical value is 2.44, so H_0 is not rejected at the 1% level. The same caveat holds here as in Example C.2: we have not controlled for other factors that might affect investment in cities over time, so we cannot claim that the effect is causal.

The rejection rule is similar for the one-sided alternative (C.33). A test with a significance level of $100 \cdot \alpha\%$ rejects H_0 against (C.33) whenever

$$t < -c; \qquad [C.38]$$

in other words, we are looking for negative values of the t statistic—which implies $\bar{y} < \mu_0$—that are sufficiently far from zero to reject H_0.

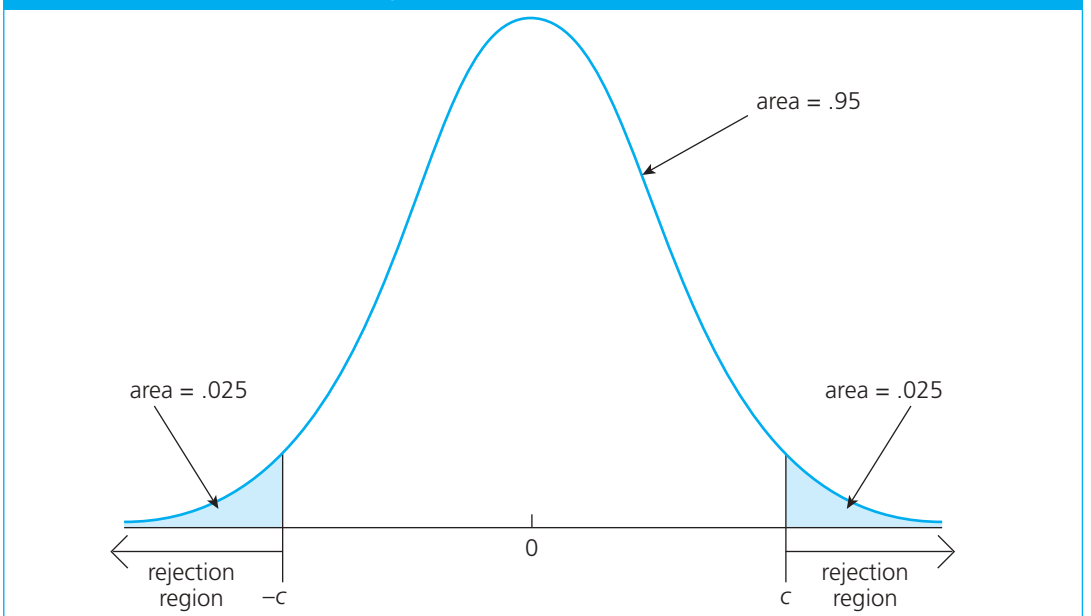
For two-sided alternatives, we must be careful to choose the critical value so that the significance level of the test is still α. If H_1 is given by H_1: $\mu \neq \mu_0$, then we reject H_0 if $\bar{y}$ is far from μ_0 in *absolute value*: a $\bar{y}$ much larger or much smaller than μ_0 provides evidence against H_0 in favor of H_1. A $100 \cdot \alpha\%$ level test is obtained from the rejection rule

$$|t| > c, \qquad [C.39]$$

where $|t|$ is the absolute value of the t statistic in (C.35). This gives a **two-tailed test**. We must now be careful in choosing the critical value: c is the $100(1 - \alpha/2)$ percentile in the t_{n-1} distribution. For example, if $\alpha = .05$, then the critical value is the 97.5^{th} percentile in the t_{n-1} distribution. This ensures that H_0 is rejected only 5% of the time when it is true (see Figure C.6). For example, if $n = 22$, then the critical value is $c = 2.08$, the 97.5^{th} percentile in a t_{21} distribution (see Table G.2). The absolute value of the t statistic must exceed 2.08 in order to reject H_0 against H_1 at the 5% level.

It is important to know the proper language of hypothesis testing. Sometimes, the appropriate phrase "we fail to reject H_0 in favor of H_1 at the 5% significance level" is replaced with "we accept H_0 at the 5% significance level." The latter wording is incorrect. With the same set of data, there are

FIGURE C.6 Rejection region for a 5% significance level test against the two-sided alternative H_1: $\mu \neq \mu_0$.



usually many hypotheses that cannot be rejected. In the earlier election example, it would be logically inconsistent to say that H_0: $\theta = .42$ and H_0: $\theta = .43$ are both "accepted," because only one of these can be true. But it is entirely possible that neither of these hypotheses is rejected. For this reason, we always say "fail to reject H_0" rather than "accept H_0."

C-6c Asymptotic Tests for Nonnormal Populations

If the sample size is large enough to invoke the central limit theorem (see Section C-3), the mechanics of hypothesis testing for population means are the *same* whether or not the population distribution is normal. The theoretical justification comes from the fact that, under the null hypothesis,

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Therefore, with large n, we can compare the t statistic in (C.35) with the critical values from a standard normal distribution. Because the t_{n-1} distribution converges to the standard normal distribution as n gets large, the t and standard normal critical values will be very close for extremely large n. Because asymptotic theory is based on n increasing without bound, it cannot tell us whether the standard normal or t critical values are better. For moderate values of n, say, between 30 and 60, it is traditional to use the t distribution because we know this is correct for normal populations. For $n > 120$, the choice between the t and standard normal distributions is largely irrelevant because the critical values are practically the same.

Because the critical values chosen using either the standard normal or t distribution are only approximately valid for nonnormal populations, our chosen significance levels are also only approximate; thus, for nonnormal populations, our significance levels are really *asymptotic* significance levels. Thus, if we choose a 5% significance level, but our population is nonnormal, then the actual significance level will be larger or smaller than 5% (and we cannot know which is the case). When the sample size is large, the actual significance level will be very close to 5%. Practically speaking, the distinction is not important, so we will now drop the qualifier "asymptotic."

EXAMPLE C.5 Race Discrimination in Hiring

In the Urban Institute study of discrimination in hiring (see Example C.3) using the data in AUDIT, we are primarily interested in testing H_0: $\mu = 0$ against H_1: $\mu < 0$ where $\mu = \theta_B - \theta_W$ is the difference in probabilities that blacks and whites receive job offers. Recall that μ is the population mean of the variable $Y = B - W$, where B and W are binary indicators. Using the $n = 241$ paired comparisons in the data file AUDIT, we obtained $\bar{y} = -.133$ and $\text{se}(\bar{y}) = .482/\sqrt{241} \approx .031$. The t statistic for testing H_0: $\mu = 0$ is $t = -.133/.031 \approx -4.29$. You will remember from Math Refresher B that the standard normal distribution is, for practical purposes, indistinguishable from the t distribution with 240 degrees of freedom. The value -4.29 is so far out in the left tail of the distribution that we reject H_0 at any reasonable significance level. In fact, the .005 (one-half of a percent) critical value (for the one-sided test) is about -2.58. A t value of -4.29 is *very* strong evidence against H_0 in favor of H_1. Hence, we conclude that there is discrimination in hiring.

C-6d Computing and Using *p*-Values

The traditional requirement of choosing a significance level ahead of time means that different researchers, using the same data and same procedure to test the same hypothesis, could wind up with different conclusions. Reporting the significance level at which we are carrying out the test solves this problem to some degree, but it does not completely remove the problem.

To provide more information, we can ask the following question: What is the *largest* significance level at which we could carry out the test and still fail to reject the null hypothesis? This value is known as the ***p*-value** of a test (sometimes called the *prob-value*). Compared with choosing a significance level ahead of time and obtaining a critical value, computing a p-value is somewhat more difficult. But with the advent of quick and inexpensive computing, p-values are now fairly easy to obtain.

As an illustration, consider the problem of testing H_0: $\mu = 0$ in a Normal(μ, σ^2) population. Our test statistic in this case is $T = \sqrt{n}\cdot\overline{Y}/S$, and we assume that n is large enough to treat T as having a standard normal distribution under H_0. Suppose that the observed value of T for our sample is $t = 1.52$. (Note how we have skipped the step of choosing a significance level.) Now that we have seen the value t, we can find the largest significance level at which we would fail to reject H_0. This is the significance level associated with using t as our critical value. Because our test statistic T has a standard normal distribution under H_0, we have

$$p\text{-value} = P(T > 1.52|H_0) = 1 - \Phi(1.52) = .065, \qquad \text{[C.40]}$$

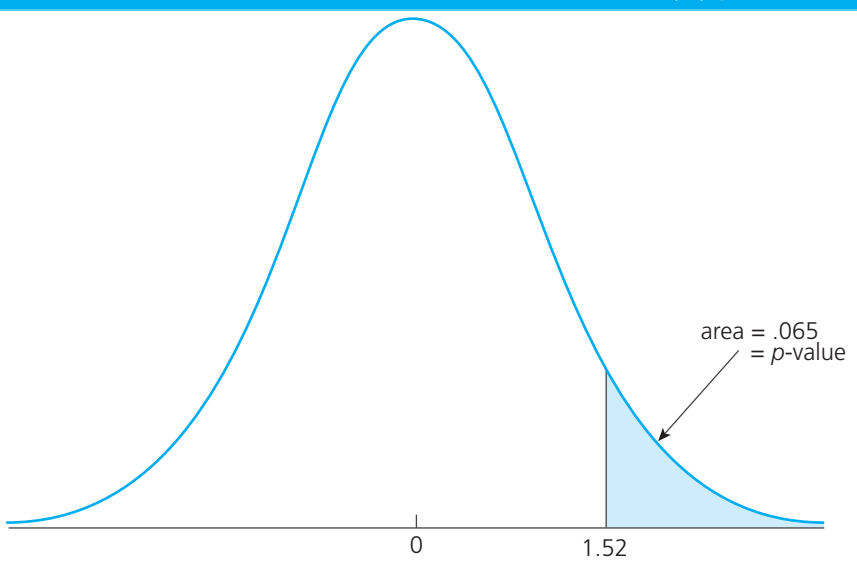
where $\Phi(\cdot)$ denotes the standard normal cdf. In other words, the p-value in this example is simply the area to the right of 1.52, the observed value of the test statistic, in a standard normal distribution. See Figure C.7 for illustration.

Because the p-value $= .065$, the largest significance level at which we can carry out this test and fail to reject is 6.5%. If we carry out the test at a level below 6.5% (such as at 5%), we fail to reject H_0. If we carry out the test at a level larger than 6.5% (such as 10%), we reject H_0. With the p-value at hand, we can carry out the test at any level.

The p-value in this example has another useful interpretation: it is the probability that we observe a value of T as large as 1.52 when the null hypothesis is true. If the null hypothesis is actually true, we would observe a value of T as large as 1.52 due to chance only 6.5% of the time. Whether this is small enough to reject H_0 depends on our tolerance for a Type I error. The p-value has a similar interpretation in all other cases, as we will see.

Generally, small p-values are evidence *against* H_0, because they indicate that the outcome of the data occurs with small probability if H_0 is true. In the previous example, if t had been a larger value, say, $t = 2.85$, then the p-value would be $1 - \Phi(2.85) \approx .002$. This means that, if the null hypothesis were true, we would observe a value of T as large as 2.85 with probability .002. How do we

FIGURE C.7 The *p*-value when $t = 1.52$ for the one-sided alternative $\mu\cdot\mu_0$.



interpret this? Either we obtained a very unusual sample or the null hypothesis is false. Unless we have a *very* small tolerance for Type I error, we would reject the null hypothesis. On the other hand, a large p-value is weak evidence against H_0. If we had gotten $t = .47$ in the previous example, then the p-value $= 1 - \Phi(.47) = .32$. Observing a value of T larger than .47 happens with probability .32, even when H_0 is true; this is large enough so that there is insufficient doubt about H_0, unless we have a very high tolerance for Type I error.

For hypothesis testing about a population mean using the t distribution, we need detailed tables in order to compute p-values. Table G.2 only allows us to put bounds on p-values. Fortunately, many statistics and econometrics packages now compute p-values routinely, and they also provide calculation of cdfs for the t and other distributions used for computing p-values.

EXAMPLE C.6 **Effect of Job Training Grants on Worker Productivity**

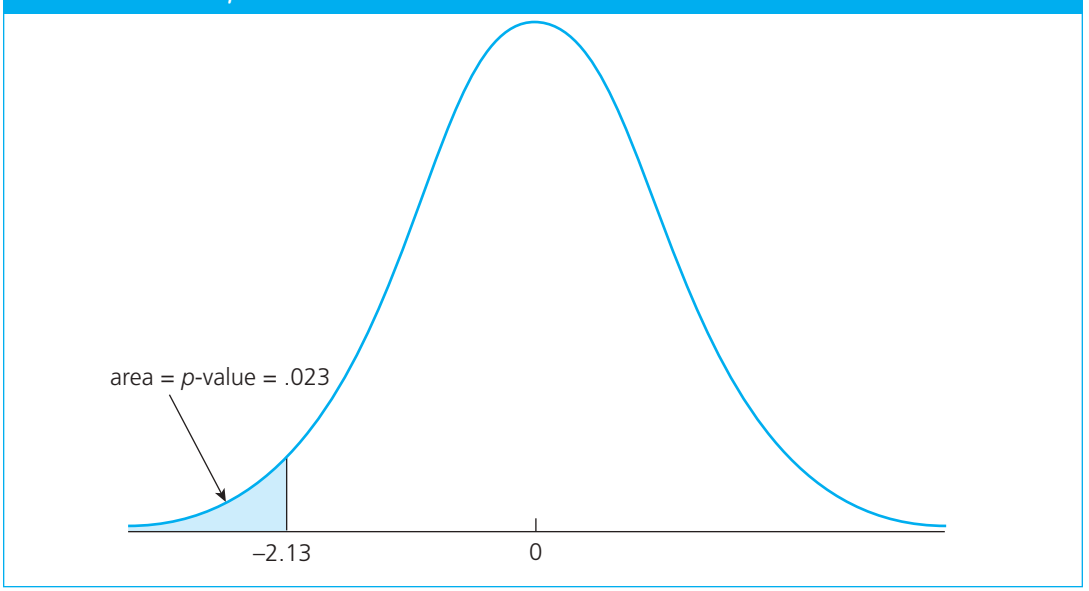
Consider again the Holzer et al. (1993) data in Example C.2. From a policy perspective, there are two questions of interest. First, what is our best estimate of the mean change in scrap rates, μ? We have already obtained this for the sample of 20 firms listed in Table C.3: the sample average of the change in scrap rates is -1.15. Relative to the initial average scrap rate in 1987, this represents a fall in the scrap rate of about 26.3% $(-1.15/4.38 \approx -.263)$, which is a nontrivial effect.

We would also like to know whether the sample provides strong evidence for an effect in the population of manufacturing firms that could have received grants. The null hypothesis is H_0: $\mu = 0$, and we test this against H_1: $\mu < 0$, where μ is the average change in scrap rates. Under the null, the job training grants have no effect on average scrap rates. The alternative states that there is an effect. We do not care about the alternative $\mu > 0$, so the null hypothesis is effectively H_0: $\mu \geq 0$.

Because $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$, $t = -1.15/.54 = -2.13$. This is below the 5% critical value of -1.73 (from a t_{19} distribution) but above the 1% critical value, -2.54. The p-value in this case is computed as

$$p\text{-value} = \text{P}(T_{19} < -2.13), \qquad \text{[C.41]}$$

FIGURE C.8 **The p-value when $t = -2.13$ with 19 degrees of freedom for the one-sided alternative $\mu < 0$.**



where T_{19} represents a t distributed random variable with 19 degrees of freedom. The inequality is reversed from (C.40) because the alternative has the form in (C.33). The probability in (C.41) is the area to the left of -2.13 in a t_{19} distribution (see Figure C.8).

Using Table G.2, the most we can say is that the p-value is between .025 and .01, but it is closer to .025 (because the 97.5th percentile is about 2.09). Using a statistical package, such as Stata®, we can compute the exact p-value. It turns out to be about .023, which is reasonable evidence against H_0. This is certainly enough evidence to reject the null hypothesis that the training grants had no effect at the 2.5% significance level (and therefore at the 5% level).

Computing a p-value for a two-sided test is similar, but we must account for the two-sided nature of the rejection rule. For t testing about population means, the p-value is computed as

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \qquad \text{[C.42]}$$

where t is the value of the test statistic and T_{n-1} is a t random variable. (For large n, replace T_{n-1} with a standard normal random variable.) Thus, compute the absolute value of the t statistic, find the area to the right of this value in a t_{n-1} distribution, and multiply the area by two.

For nonnormal populations, the exact p-value can be difficult to obtain. Nevertheless, we can find *asymptotic* p-values by using the same calculations. These p-values are valid for large sample sizes. For n larger than, say, 120, we might as well use the standard normal distribution. Table G.1 is detailed enough to get accurate p-values, but we can also use a statistics or econometrics program.

EXAMPLE C.7 **Race Discrimination in Hiring**

Using the matched pairs data from the Urban Institute in the AUDIT data file ($n = 241$), we obtained $t = -4.29$. If Z is a standard normal random variable, $P(Z < -4.29)$ is, for practical purposes, zero. In other words, the (asymptotic) p-value for this example is essentially zero. This is very strong evidence against H_0.

Summary of How to Use *p*-Values:

(i) Choose a test statistic T and decide on the nature of the alternative. This determines whether the rejection rule is $t > c$, $t < -c$, or $|t| > c$.

(ii) Use the observed value of the t statistic as the critical value and compute the corresponding significance level of the test. This is the p-value. If the rejection rule is of the form $t > c$, then $p\text{-value} = P(T > t)$. If the rejection rule is $t < -c$, then $p\text{-value} = P(T < t)$; if the rejection rule is $|t| > c$, then $p\text{-value} = P(|T| > |t|)$.

(iii) If a significance level α has been chosen, then we reject H_0 at the $100{\cdot}\alpha\%$ level if $p\text{-value} < \alpha$. If $p\text{-value} \geq \alpha$, then we fail to reject H_0 at the $100{\cdot}\alpha\%$ level. Therefore, it is a small p-value that leads to rejection of the null hypothesis.

C-6e The Relationship between Confidence Intervals and Hypothesis Testing

Because constructing confidence intervals and hypothesis tests both involve probability statements, it is natural to think that they are somehow linked. It turns out that they are. After a confidence interval has been constructed, we can carry out a variety of hypothesis tests.

The confidence intervals we have discussed are all two-sided by nature. (In this text, we will have no need to construct one-sided confidence intervals.) Thus, confidence intervals can be used to

test against *two-sided* alternatives. In the case of a population mean, the null is given by (C.31), and the alternative is (C.34). Suppose we have constructed a 95% confidence interval for μ. Then, if the hypothesized value of μ under H_0, μ_0, is not in the confidence interval, then H_0: $\mu = \mu_0$ is rejected against H_1: $\mu \neq \mu_0$ at the 5% level. If μ_0 lies in this interval, then we fail to reject H_0 at the 5% level. Notice how any value for μ_0 can be tested once a confidence interval is constructed, and because a confidence interval contains more than one value, there are many null hypotheses that will not be rejected.

EXAMPLE C.8 Training Grants and Worker Productivity

In the Holzer et al. example, we constructed a 95% confidence interval for the mean change in scrap rate μ as $[-2.28, -.02]$. Because zero is excluded from this interval, we reject H_0: $\mu = 0$ against H_1: $\mu \neq 0$ at the 5% level. This 95% confidence interval also means that we fail to reject H_0: $\mu = -2$ at the 5% level. In fact, there is a continuum of null hypotheses that are not rejected given this confidence interval.

C-6f Practical versus Statistical Significance

In the examples covered so far, we have produced three kinds of evidence concerning population parameters: point estimates, confidence intervals, and hypothesis tests. These tools for learning about population parameters are equally important. There is an understandable tendency for students to focus on confidence intervals and hypothesis tests because these are things to which we can attach confidence or significance levels. But in any study, we must also interpret the *magnitudes* of point estimates.

The sign and magnitude of $\bar{y}$ determine its **practical significance** and allow us to discuss the direction of an intervention or policy effect, and whether the estimated effect is "large" or "small." On the other hand, **statistical significance** of $\bar{y}$ depends on the magnitude of its t statistic. For testing H_0: $\mu = 0$, the t statistic is simply $t = \bar{y}/\text{se}(\bar{y})$. In other words, statistical significance depends on the ratio of $\bar{y}$ to its standard error. Consequently, a t statistic can be large because $\bar{y}$ is large or $\text{se}(\bar{y})$ is small. In applications, it is important to discuss both practical and statistical significance, being aware that an estimate can be statistically significant without being especially large in a practical sense. Whether an estimate is practically important depends on the context as well as on one's judgment, so there are no set rules for determining practical significance.

EXAMPLE C.9 Effect of Freeway Width on Commute Time

Let Y denote the change in commute time, measured in minutes, for commuters in a metropolitan area from before a freeway was widened to after the freeway was widened. Assume that $Y \sim \text{Normal}(\mu,\sigma^2)$. The null hypothesis that the widening did not reduce average commute time is H_0: $\mu = 0$; the alternative that it reduced average commute time is H_1: $\mu < 0$. Suppose a random sample of commuters of size $n = 900$ is obtained to determine the effectiveness of the freeway project. The average change in commute time is computed to be $\bar{y} = -3.6$, and the sample standard deviation is $s = 32.7$; thus, $\text{se}(\bar{y}) = 32.7/\sqrt{900} = 1.09$. The t statistic is $t = -3.6/1.09 \approx -3.30$, which is very statistically significant; the p-value is about .0005. Thus, we conclude that the freeway widening had a statistically significant effect on average commute time.

If the outcome of the hypothesis test is all that were reported from the study, it would be misleading. Reporting only statistical significance masks the fact that the estimated reduction in average commute time, 3.6 minutes, seems pretty meager, although this depends to some extent on what the average commute time was prior to widening the freeway. To be up front, we should report the point estimate of -3.6, along with the significance test.

Finding point estimates that are statistically significant without being practically significant can occur when we are working with large samples. To discuss why this happens, it is useful to have the following definition.

Test Consistency. A **consistent test** rejects H_0 with probability approaching one as the sample size grows whenever H_1 is true.

Another way to say that a test is consistent is that, as the sample size tends to infinity, the power of the test gets closer and closer to unity whenever H_1 is true. All of the tests we cover in this text have this property. In the case of testing hypotheses about a population mean, test consistency follows because the variance of $\overline{Y}$ converges to zero as the sample size gets large. The t statistic for testing H_0: $\mu = 0$ is $T = \overline{Y}/(S/\sqrt{n})$. Because $\text{plim}(\overline{Y}) = \mu$ and $\text{plim}(S) = \sigma$, it follows that if, say, $\mu > 0$, then T gets larger and larger (with high probability) as $n \to \infty$. In other words, no matter how close m is to zero, we can be almost certain to reject H_0: $\mu = 0$ given a large enough sample size. This says nothing about whether μ is large in a practical sense.

C-7 Remarks on Notation

In our review of probability and statistics here and in Math Refresher B, we have been careful to use standard conventions to denote random variables, estimators, and test statistics. For example, we have used W to indicate an estimator (random variable) and w to denote a particular estimate (outcome of the random variable W). Distinguishing between an estimator and an estimate is important for understanding various concepts in estimation and hypothesis testing. However, making this distinction quickly becomes a burden in econometric analysis because the models are more complicated: many random variables and parameters will be involved, and being true to the usual conventions from probability and statistics requires many extra symbols.

In the main text, we use a simpler convention that is widely used in econometrics. If θ is a population parameter, the notation $\hat{\theta}$ ("theta hat") will be used to denote both an estimator and an estimate of θ. This notation is useful in that it provides a simple way of attaching an estimator to the population parameter it is supposed to be estimating. Thus, if the population parameter is β, then $\hat{\beta}$ denotes an estimator or estimate of β; if the parameter is σ^2, $\hat{\sigma}^2$ is an estimator or estimate of σ^2; and so on. Sometimes, we will discuss two estimators of the same parameter, in which case we will need a different notation, such as $\tilde{\theta}$ ("theta tilde").

Although dropping the conventions from probability and statistics to indicate estimators, random variables, and test statistics puts additional responsibility on you, it is not a big deal once the difference between an estimator and an estimate is understood. If we are discussing *statistical* properties of $\hat{\theta}$—such as deriving whether or not it is unbiased or consistent—then we are necessarily viewing $\hat{\theta}$ as an estimator. On the other hand, if we write something like $\hat{\theta} = 1.73$, then we are clearly denoting a point estimate from a given sample of data. The confusion that can arise by using $\hat{\theta}$ to denote both should be minimal once you have a good understanding of probability and statistics.

Summary

We have discussed topics from mathematical statistics that are heavily relied upon in econometric analysis. The notion of an estimator, which is simply a rule for combining data to estimate a population parameter, is fundamental. We have covered various properties of estimators. The most important small sample properties are unbiasedness and efficiency, the latter of which depends on comparing variances when estimators are unbiased. Large sample properties concern the sequence of estimators

obtained as the sample size grows, and they are also depended upon in econometrics. Any useful estimator is consistent. The central limit theorem implies that, in large samples, the sampling distribution of most estimators is approximately normal.

The sampling distribution of an estimator can be used to construct confidence intervals. We saw this for estimating the mean from a normal distribution and for computing approximate confidence intervals in nonnormal cases. Classical hypothesis testing, which requires specifying a null hypothesis, an alternative hypothesis, and a significance level, is carried out by comparing a test statistic to a critical value. Alternatively, a *p*-value can be computed that allows us to carry out a test at any significance level.

Key Terms

Alternative Hypothesis
Asymptotic Normality
Bias
Biased Estimator
Central Limit Theorem (CLT)
Confidence Interval
Consistent Estimator
Consistent Test
Critical Value
Estimate
Estimator
Hypothesis Test
Inconsistent
Interval Estimator
Law of Large Numbers (LLN)
Least Squares Estimator
Log-Likelihood Function
Maximum Likelihood Estimator
Mean Squared Error (MSE)
Method of Moments
Minimum Variance Unbiased Estimator
Null Hypothesis
One-Sided Alternative
One-Tailed Test
Population
Power of a Test
Practical Significance
Probability Limit
p-Value
Random Sample
Rejection Region
Sample Average
Sample Correlation Coefficient
Sample Covariance
Sample Standard Deviation
Sample Variance
Sampling Distribution
Sampling Standard Deviation
Sampling Variance
Significance Level
Standard Error
Statistical Significance
t Statistic
Test Statistic
Two-Sided Alternative
Two-Tailed Test
Type I Error
Type II Error
Unbiased Estimator

Problems

1 Let Y_1, Y_2, Y_3, and Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ denote the average of these four random variables.

(i) What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

(ii) Now, consider a different estimator of μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

This is an example of a *weighted* average of the Y_i. Show that W is also an unbiased estimator of μ. Find the variance of W.

(iii) Based on your answers to parts (i) and (ii), which estimator of m do you prefer, $\bar{Y}$ or W?

2 This is a more general version of Problem C.1. Let $Y_1, Y_2, \ldots, Y_n$ be n pairwise uncorrelated random variables with common mean m and common variance σ^2. Let $\bar{Y}$ denote the sample average.

(i) Define the class of *linear estimators* of μ by

$$W_a = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n,$$

where the a_i are constants. What restriction on the a_i is needed for W_a to be an unbiased estimator of μ?

(ii) Find $\text{Var}(W_a)$.

(iii) For any numbers $a_1, a_2, \ldots, a_n$, the following inequality holds: $(a_1 + a_2 + \cdots + a_n)^2/n \leq a_1^2 + a_2^2 + \cdots + a_n^2$. Use this, along with parts (i) and (ii), to show that $\text{Var}(W_a) \geq \text{Var}(\bar{Y})$ whenever W_a is unbiased, so that $\bar{Y}$ is the *best linear unbiased estimator*. [*Hint*: What does the inequality become when the a_i satisfy the restriction from part (i)?]

3 Let $\bar{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider two alternative estimators of μ: $W_1 = [(n-1)/n]\bar{Y}$ and $W_2 = \bar{Y}/2$.

(i) Show that W_1 and W_2 are both biased estimators of μ and find the biases. What happens to the biases as $n \to \infty$? Comment on any important differences in bias for the two estimators as the sample size gets large.

(ii) Find the probability limits of W_1 and W_2. {*Hint*: Use Properties PLIM.1 and PLIM.2; for W_1, note that plim $[(n-1)/n] = 1$.} Which estimator is consistent?

(iii) Find $\text{Var}(W_1)$ and $\text{Var}(W_2)$.

(iv) Argue that W_1 is a better estimator than $\bar{Y}$ if μ is "close" to zero. (Consider both bias and variance.)

4 For positive random variables X and Y, suppose the expected value of Y given X is $\text{E}(Y|X) = \theta X$. The unknown parameter θ shows how the expected value of Y changes with X.

(i) Define the random variable $Z = Y/X$. Show that $\text{E}(Z) = \theta$. [*Hint*: Use Property CE.2 in Math Refresher B along with the law of iterated expectations, Property CE.4 (also in Math Refresher B). In particular, first show that $\text{E}(Z|X) = \theta$ and then use CE.4.]

(ii) Use part (i) to prove that the estimator $W_1 = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ is unbiased for θ, where $\{(X_i, Y_i): i = 1, 2, \ldots, n\}$ is a random sample.

(iii) Explain why the estimator $W_2 = \bar{Y}/\bar{X}$, where the overbars denote sample averages, is not the same as W_1. Nevertheless, show that W_2 is also unbiased for θ.

(iv) The following table contains data on corn yields for several counties in Iowa. The USDA predicts the number of hectares of corn in each county based on satellite photos. Researchers count the number of "pixels" of corn in the satellite picture (as opposed to, for example, the number of pixels of soybeans or of uncultivated land) and use these to predict the actual number of hectares. To develop a prediction equation to be used for counties in general, the USDA surveyed farmers in selected counties to obtain corn yields in hectares. Let Y_i = corn yield in county i and let X_i = number of corn pixels in the satellite picture for county i. There are $n = 17$ observations for eight counties. Use this sample to compute the estimates of θ devised in parts (ii) and (iii). Are the estimates similar?

Plot	Corn Yield	Corn Pixels
1	165.76	374
2	96.32	209
3	76.08	253
4	185.35	432
5	116.43	367
6	162.08	361
7	152.04	288
8	161.75	369
9	92.88	206
10	149.94	316
11	64.75	145
12	127.07	355
13	133.55	295
14	77.70	223
15	206.39	459
16	108.33	290
17	118.17	307

5 Let Y denote a Bernoulli(θ) random variable with $0 < \theta < 1$. Suppose we are interested in estimating the *odds ratio*, $\gamma = \theta/(1 - \theta)$, which is the probability of success over the probability of failure. Given a random sample $\{Y_1, \ldots, Y_n\}$, we know that an unbiased and consistent estimator of θ is $\overline{Y}$, the proportion of successes in n trials. A natural estimator of γ is $G = \overline{Y}/(1 - \overline{Y})$, the proportion of successes over the proportion of failures in the sample.

(i) Why is G not an unbiased estimator of γ?

(ii) Use PLIM.2 (iii) to show that G is a consistent estimator of γ.

6 You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a Normal(μ, σ^2) distribution.

(i) The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

(ii) The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

(iii) Now, suppose your sample size is $n = 900$ and you obtain the estimates $\bar{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? At the 1% level?

(iv) Would you say that the estimated fall in consumption is large in magnitude? Comment on the practical versus statistical significance of this estimate.

(v) What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

7 The new management at a bakery claims that workers are now more productive than they were under old management, which is why wages have "generally increased." Let W_i^b be Worker i's wage under the old management and let W_i^a be Worker i's wage after the change. The difference is $D_i \equiv W_i^a - W_i^b$. Assume that the D_i are a random sample from a Normal (μ, σ^2) distribution.

(i) Using the following data on 15 workers, construct an exact 95% confidence interval for μ.

(ii) Formally state the null hypothesis that there has been no change in average wages. In particular, what is $E(D_i)$ under H_0? If you are hired to examine the validity of the new management's claim, what is the relevant alternative hypothesis in terms of $\mu = E(D_i)$?

(iii) Test the null hypothesis from part (ii) against the stated alternative at the 5% and 1% levels.

(iv) Obtain the p-value for the test in part (iii).

Worker	Wage Before	Wage After
1	8.30	9.25
2	9.40	9.00
3	9.00	9.25
4	10.50	10.00
5	11.40	12.00
6	8.75	9.50
7	10.00	10.25
8	9.50	9.50
9	10.80	11.50
10	12.55	13.10
11	12.00	11.50
12	8.65	9.00
13	7.75	7.75
14	11.25	11.50
15	12.65	13.00

8 The *New York Times* (2/5/90) reported three-point shooting performance for the top 10 three-point shooters in the NBA. The following table summarizes these data:

Player	FGA-FGM
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Note: FGA = field goals attempted and FGM = field goals made.

For a given player, the outcome of a particular shot can be modeled as a Bernoulli (zero-one) variable: if Y_i is the outcome of shot i, then $Y_i = 1$ if the shot is made, and $Y_i = 0$ if the shot is missed. Let θ denote the probability of making any particular three-point shot attempt. The natural estimator of θ is $\overline{Y} = FGM/FGA$.

(i) Estimate θ for Mark Price.

(ii) Find the standard deviation of the estimator $\overline{Y}$ in terms of θ and the number of shot attempts, n.

(iii) The asymptotic distribution of $(\overline{Y} - \theta)/\text{se}(\overline{Y})$ is standard normal, where $\text{se}(\overline{Y}) = \sqrt{\overline{Y}(1 - \overline{Y})/n}$. Use this fact to test H_0: $\theta = .5$ against H_1: $\theta < .5$ for Mark Price. Use a 1% significance level.

9 Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. A human rights group suspects foul play and hires you to test the validity of the dictator's claim. You have a budget that allows you to randomly sample 200 voters from the country.

(i) Let X be the number of yes votes obtained from a random sample of 200 out of the entire voting population. What is the expected value of X if, in fact, 65% of all voters supported the dictator?

(ii) What is the standard deviation of X, again assuming that the true fraction voting yes in the plebiscite is .65?

(iii) Now, you collect your sample of 200, and you find that 115 people actually voted yes. Use the CLT to approximate the probability that you would find 115 or fewer yes votes from a random sample of 200 if, in fact, 65% of the entire population voted yes.

(iv) How would you explain the relevance of the number in part (iii) to someone who does not have training in statistics?

10 Before a strike prematurely ended the 1994 major league baseball season, Tony Gwynn of the San Diego Padres had 165 hits in 419 at bats, for a .394 batting average. There was discussion about whether Gwynn was a potential .400 hitter that year. This issue can be couched in terms of Gwynn's probability of getting a hit on a particular at bat, call it θ. Let Y_i be the Bernoulli(θ) indicator equal to unity if Gwynn gets a hit during his i^{th} at bat, and zero otherwise. Then, $Y_1, Y_2, \ldots, Y_n$ is a random sample from a Bernoulli(θ) distribution, where θ is the probability of success, and $n = 419$.

Our best point estimate of θ is Gwynn's batting average, which is just the proportion of successes: $\bar{y} = .394$. Using the fact that $\text{se}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, construct an approximate 95% confidence interval for θ, using the standard normal distribution. Would you say there is strong evidence against Gwynn's being a potential .400 hitter? Explain.

11 Suppose that between their first and second years in college, 400 students are randomly selected and given a university grant to purchase a new computer. For student i, y_i denotes the change in GPA from the first year to the second year. If the average change is $\bar{y} = .132$ with standard deviation $s = 1.27$, is the average change in GPAs statistically greater than zero?

12 (Requires Calculus) A count random variable, say Y, takes on nonnegative integer values, $\{0, 1, 2, \ldots\}$. The most common distribution for a count variable is the *Poisson*(θ) distribution, where the parameter θ is the expected value: $\theta = \text{E}(Y)$. The probability density function is

$$f(y; \theta) = \exp(-\theta)\theta^y/y!, \; y = 0, 1, 2, \ldots$$
$$= 0 \text{ otherwise}$$

It can be shown that $\text{Var}(Y) = \theta$, so that the mean and variance are the same.

(i) For a random draw Y_i from the population, find the log-likelihood function $\ell(\theta; Y_i) = \log[f(Y_i; \theta)]$. What is the log likelihood for a random sample of size n, say $\mathcal{L}_n(\theta)$? [Hint: Look at equation (C.16).]

(ii) Using the notational convention in Section C.7, find the first order condition for the MLE, $\hat{\theta}$, and show that $\hat{\theta} = \bar{Y}$, the sample average.

(iii) Why is $\hat{\theta}$ unbiased?

(iv) Find $\text{Var}(\bar{Y})$ as a function of θ and n.

(v) Why is $\bar{Y}$ consistent?

(vi) Do the unbiasedness and consistency of the MLE in this case depend on whether the Poisson distribution is correct? Explain.

(vii) What is the distribution of

$$\frac{\sqrt{n}(\bar{Y} - \theta)}{\sqrt{\theta}}$$

as $n \to \infty$? Explain.

(viii) If $\text{E}(Y) = \theta$ but $\text{Var}(Y) = v(\theta) > 0$—so that the Poisson distribution may fail—modify the random variable in (vii) so that it has a limiting distribution that does not depend on θ.

Advanced Treatment D

Summary of Matrix Algebra

This Advanced Treatment summarizes the matrix algebra concepts, including the algebra of probability, needed for the study of multiple linear regression models using matrices in Advanced Treatment E. None of this material is used in the main text.

D-1 Basic Definitions

Definition D.1 (Matrix). A **matrix** is a rectangular array of numbers. More precisely, an $m \times n$ matrix has m rows and n columns. The positive integer m is called the *row dimension*, and n is called the *column dimension*.

We use uppercase boldface letters to denote matrices. We can write an $m \times n$ matrix generically as

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix},$$

where a_{ij} represents the element in the i^{th} row and the j^{th} column. For example, a_{25} stands for the number in the second row and the fifth column of $\mathbf{A}$. A specific example of a 2×3 matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \qquad [D.1]$$

where $a_{13} = 7$. The shorthand $\mathbf{A} = [a_{ij}]$ is often used to define matrix operations.

Definition D.2 (Square Matrix). A **square matrix** has the same number of rows and columns. The dimension of a square matrix is its number of rows and columns.

Definition D.3 (Vectors)

(i) A $1 \times m$ matrix is called a **row vector** (of dimension m) and can be written as $\mathbf{x} \equiv (x_1, x_2, \dots, x_m)$.

(ii) An $n \times 1$ matrix is called a **column vector** and can be written as

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Definition D.4 (Diagonal Matrix). A square matrix $\mathbf{A}$ is a **diagonal matrix** when all of its off-diagonal elements are zero, that is, $a_{ij} = 0$ for all $i \neq j$. We can always write a diagonal matrix as

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

Definition D.5 (Identity and Zero Matrices)

(i) The $n \times n$ **identity matrix**, denoted $\mathbf{I}$, or sometimes $\mathbf{I}_n$ to emphasize its dimension, is the diagonal matrix with unity (one) in each diagonal position, and zero elsewhere:

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

(ii) The $m \times n$ **zero matrix**, denoted $\mathbf{0}$, is the $m \times n$ matrix with zero for all entries. This need not be a square matrix.

D-2 Matrix Operations

D-2a Matrix Addition

Two matrices $\mathbf{A}$ and $\mathbf{B}$, each having dimension $m \times n$, can be added element by element: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. More precisely,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Matrices of different dimensions cannot be added.

D-2b Scalar Multiplication

Given any real number γ (often called a scalar), **scalar multiplication** is defined as $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, or

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \dots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \dots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \dots & \gamma a_{mn} \end{bmatrix}.$$

For example, if $\gamma = 2$ and $\mathbf{A}$ is the matrix in equation (D.1), then

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

D-2c Matrix Multiplication

To multiply matrix $\mathbf{A}$ by matrix $\mathbf{B}$ to form the product $\mathbf{AB}$, the *column* dimension of $\mathbf{A}$ must equal the *row* dimension of $\mathbf{B}$. Therefore, let $\mathbf{A}$ be an $m \times n$ matrix and let $\mathbf{B}$ be an $n \times p$ matrix. Then, **matrix multiplication** is defined as

$$\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik} b_{kj} \right].$$

In other words, the $(i, j)^{\text{th}}$ element of the new matrix $\mathbf{AB}$ is obtained by multiplying each element in the i^{th} row of $\mathbf{A}$ by the corresponding element in the j^{th} column of $\mathbf{B}$ and adding these n products together. A schematic may help make this process more transparent:

$$\begin{array}{ccccc} & \mathbf{A} & & \mathbf{B} & \mathbf{AB} \\ i^{\text{th}}\text{ row} \rightarrow & \begin{bmatrix} a_{i1} a_{i2} a_{i3} \ldots a_{in} \end{bmatrix} & \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix} & = & \begin{bmatrix} \sum_{k=1}^{n} a_{ik} b_{kj} \end{bmatrix}, \\ & & \uparrow & & \uparrow \\ & & j^{\text{th}}\text{ column} & & (i, j)^{\text{th}}\text{ element} \end{array}$$

where, by the definition of the summation operator in Math Refresher A,

$$\sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

We can also multiply a matrix and a vector. If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{y}$ is an $m \times 1$ vector, then $\mathbf{Ay}$ is an $n \times 1$ vector. If $\mathbf{x}$ is a $1 \times n$ vector, then $\mathbf{xA}$ is a $1 \times m$ vector.

Matrix addition, scalar multiplication, and matrix multiplication can be combined in various ways, and these operations satisfy several rules that are familiar from basic operations on numbers. In the following list of properties, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are matrices with appropriate dimensions for applying each operation, and α and β are real numbers. Most of these properties are easy to illustrate from the definitions.

Properties of Matrix Operations. (1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; (2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; (3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; (4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; (5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; (6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; (7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; (8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; (9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; (10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; (11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; (12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; (13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$; and (14) $\mathbf{AB} \neq \mathbf{BA}$, even when both products are defined.

The last property deserves further comment. If $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times p$, then $\mathbf{AB}$ is defined, but $\mathbf{BA}$ is defined only if $n = p$ (the row dimension of $\mathbf{A}$ equals the column dimension of $\mathbf{B}$). If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$, then $\mathbf{AB}$ and $\mathbf{BA}$ are both defined, but they are not usually the same; in fact, they have different dimensions, unless $\mathbf{A}$ and $\mathbf{B}$ are both square matrices. Even when $\mathbf{A}$ and $\mathbf{B}$ are both square, $\mathbf{AB} \neq \mathbf{BA}$, except under special circumstances.

D-2d Transpose

Definition D.6 (Transpose). Let $\mathbf{A} = [a_{ij}]$ be an $m \times n$ matrix. The **transpose** of $\mathbf{A}$, denoted $\mathbf{A}'$ (called $\mathbf{A}$ *prime*), is the $n \times m$ matrix obtained by interchanging the rows and columns of $\mathbf{A}$. We can write this as $\mathbf{A}' \equiv [a_{ji}]$.

For example,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

Properties of Transpose. (1) $(\mathbf{A}')' = \mathbf{A}$; (2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ for any scalar α; (3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; (4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times k$; (5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$, where $\mathbf{x}$ is an $n \times 1$ vector; and (6) If $\mathbf{A}$ is an $n \times k$ matrix with *rows* given by the $1 \times k$ vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, so that we can write

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix},$$

then $\mathbf{A}' = (\mathbf{a}_1' \mathbf{a}_2' \ldots \mathbf{a}_n')$.

Definition D.7 (Symmetric Matrix). A square matrix $\mathbf{A}$ is a **symmetric matrix** if, and only if, $\mathbf{A}' = \mathbf{A}$.

If $\mathbf{X}$ is any $n \times k$ matrix, then $\mathbf{X}'\mathbf{X}$ is always defined and is a symmetric matrix, as can be seen by applying the first and fourth transpose properties (see Problem 3).

D-2e Partitioned Matrix Multiplication

Let $\mathbf{A}$ be an $n \times k$ matrix with rows given by the $1 \times k$ vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, and let $\mathbf{B}$ be an $n \times m$ matrix with rows given by $1 \times m$ vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Then,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \mathbf{a}_i' \mathbf{b}_i,$$

where for each i, $\mathbf{a}_i'\mathbf{b}_i$ is a $k \times m$ matrix. Therefore, $\mathbf{A}'\mathbf{B}$ can be written as the sum of n matrices, each of which is $k \times m$. As a special case, we have

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{a}_i,$$

where $\mathbf{a}_i'\mathbf{a}_i$ is a $k \times k$ matrix for all i.

A more general form of partitioned matrix multiplication holds when we have matrices $\mathbf{A}$ $(m \times n)$ and $\mathbf{B}$ $(n \times p)$ written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ is $m_1 \times n_1$, $\mathbf{A}_{12}$ is $m_1 \times n_2$, $\mathbf{A}_{21}$ is $m_2 \times n_1$, $\mathbf{A}_{22}$ is $m_2 \times n_2$, $\mathbf{B}_{11}$ is $n_1 \times p_1$, $\mathbf{B}_{12}$ is $n_1 \times p_2$, $\mathbf{B}_{21}$ is $n_2 \times p_1$, and $\mathbf{B}_{22}$ is $n_2 \times p_2$. Naturally, $m_1 + m_2 = m$, $n_1 + n_2 = n$, and $p_1 + p_2 = p$.

When we form the product **AB,** the expression looks just like when the entries are scalars:

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}.$$

Note that each of the matrix multiplications that form the partition on the right is well defined because the column and row dimensions are compatible for multiplication.

D-2f Trace

The trace of a matrix is a very simple operation defined only for *square* matrices.

Definition D.8 (Trace). For any $n \times n$ matrix $\mathbf{A}$, the **trace of a matrix A**, denoted $\text{tr}(\mathbf{A})$, is the sum of its diagonal elements. Mathematically,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Properties of Trace. (1) $\text{tr}(\mathbf{I}_n) = n$; (2) $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$; (3) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$; (4) $\text{tr}(\alpha\mathbf{A}) = \alpha\text{tr}(\mathbf{A})$, for any scalar α; and (5) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$.

D-2g Inverse

The notion of a matrix inverse is very important for square matrices.

Definition D.9 (Inverse). An $n \times n$ matrix $\mathbf{A}$ has an **inverse**, denoted $\mathbf{A}^{-1}$, provided that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. In this case, $\mathbf{A}$ is said to be *invertible* or *nonsingular.* Otherwise, it is said to be *noninvertible* or *singular*.

Properties of Inverse. (1) If an inverse exists, it is unique; (2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, if $\alpha \neq 0$ and $\mathbf{A}$ is invertible; (3) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, if $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$ and invertible; and (4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

We will not be concerned with the mechanics of calculating the inverse of a matrix. Any matrix algebra text contains detailed examples of such calculations.

D-3 Linear Independence and Rank of a Matrix

For a set of vectors having the same dimension, it is important to know whether one vector can be expressed as a linear combination of the remaining vectors.

Definition D.10 (Linear Independence). Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ be a set of $n \times 1$ vectors. These are **linearly independent vectors** if, and only if,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_r\mathbf{x}_r = \mathbf{0} \qquad [D.2]$$

implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. If (D.2) holds for a set of scalars that are not all zero, then $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is *linearly dependent*.

The statement that $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is linearly dependent is equivalent to saying that at least one vector in this set can be written as a linear combination of the others.

Definition D.11 (Rank)

(i) Let $\mathbf{A}$ be an $n \times m$ matrix. The **rank of a matrix A**, denoted rank($\mathbf{A}$), is the maximum number of linearly independent columns of $\mathbf{A}$.

(ii) If $\mathbf{A}$ is $n \times m$ and $\text{rank}(\mathbf{A}) = m$, then $\mathbf{A}$ has *full column rank*.

If $\mathbf{A}$ is $n \times m$, its rank can be at most m. A matrix has full column rank if its columns form a linearly independent set. For example, the 3×2 matrix

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

can have at most rank two. In fact, its rank is only one because the second column is three times the first column.

Properties of Rank. (1) $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A})$; (2) If $\mathbf{A}$ is $n \times k$, then $\text{rank}(\mathbf{A}) \leq \min(n, k)$; and (3) If $\mathbf{A}$ is $k \times k$ and $\text{rank}(\mathbf{A}) = k$, then $\mathbf{A}$ is invertible.

D-4 Quadratic Forms and Positive Definite Matrices

Definition D.12 (Quadratic Form). Let $\mathbf{A}$ be an $n \times n$ *symmetric* matrix. The **quadratic form** associated with the matrix $\mathbf{A}$ is the real-valued function defined for all $n \times 1$ vectors $\mathbf{x}$:

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n} a_{ij}x_ix_j.$$

Definition D.13 (Positive Definite and Positive Semi-Definite)

(i) A symmetric matrix $\mathbf{A}$ is said to be **positive definite (p.d.)** if

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ for all } n \times 1 \text{ vectors } \mathbf{x} \text{ except } \mathbf{x} = \mathbf{0}.$$

(ii) A symmetric matrix $\mathbf{A}$ is **positive semi-definite (p.s.d.)** if

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ for all } n \times 1 \text{ vectors.}$$

If a matrix is positive definite or positive semi-definite, it is automatically assumed to be symmetric.

Properties of Positive Definite and Positive Semi-Definite Matrices. (1) A p.d. matrix has diagonal elements that are strictly positive, while a p.s.d. matrix has nonnegative diagonal elements; (2) If $\mathbf{A}$ is p.d., then $\mathbf{A}^{-1}$ exists and is p.d.; (3) If $\mathbf{X}$ is $n \times k$, then $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are p.s.d.; and (4) If $\mathbf{X}$ is $n \times k$ and $\text{rank}(\mathbf{X}) = k$, then $\mathbf{X}'\mathbf{X}$ is p.d. (and therefore nonsingular).

D-5 Idempotent Matrices

Definition D.14 (Idempotent Matrix). Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Then $\mathbf{A}$ is said to be an **idempotent matrix** if, and only if, $\mathbf{AA} = \mathbf{A}$.

For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is an idempotent matrix, as direct multiplication verifies.

Properties of Idempotent Matrices. Let $\mathbf{A}$ be an $n \times n$ idempotent matrix. (1) $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$, and (2) $\mathbf{A}$ is positive semi-definite.

We can construct idempotent matrices very generally. Let $\mathbf{X}$ be an $n \times k$ matrix with $\text{rank}(\mathbf{X}) = k$. Define

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Then $\mathbf{P}$ and $\mathbf{M}$ are symmetric, idempotent matrices with $\text{rank}(\mathbf{P}) = k$ and $\text{rank}(\mathbf{M}) = n - k$. The ranks are most easily obtained by using Property 1: $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (from Property 5 for trace) $= \text{tr}(\mathbf{I}_k) = k$ (by Property 1 for trace). It easily follows that $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

D-6 Differentiation of Linear and Quadratic Forms

For a given $n \times 1$ vector $\mathbf{a}$, consider the linear function defined by

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x},$$

for all $n \times 1$ vectors $\mathbf{x}$. The derivative of f with respect to $\mathbf{x}$ is the $1 \times n$ vector of partial derivatives, which is simply

$$\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'.$$

For an $n \times n$ symmetric matrix $\mathbf{A}$, define the quadratic form

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

Then,

$$\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A},$$

which is a $1 \times n$ vector.

D-7 Moments and Distributions of Random Vectors

In order to derive the expected value and variance of the OLS estimators using matrices, we need to define the expected value and variance of a **random vector**. As its name suggests, a random vector is simply a vector of random variables. We also need to define the multivariate normal distribution. These concepts are simply extensions of those covered in Math Refresher B.

D-7a Expected Value

Definition D.15 (Expected Value)

(i) If $\mathbf{y}$ is an $n \times 1$ random vector, the **expected value** of $\mathbf{y}$, denoted $\mathrm{E}(\mathbf{y})$, is the vector of expected values: $\mathrm{E}(\mathbf{y}) = [\mathrm{E}(y_1), \mathrm{E}(y_2), \ldots, \mathrm{E}(y_n)]'$.

(ii) If $\mathbf{Z}$ is an $n \times m$ random matrix, $\mathrm{E}(\mathbf{Z})$ is the $n \times m$ matrix of expected values: $\mathrm{E}(\mathbf{Z}) = [\mathrm{E}(z_{ij})]$.

Properties of Expected Value. (1) If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an $n \times 1$ vector, where both are nonrandom, then $\mathrm{E}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}\mathrm{E}(\mathbf{y}) + \mathbf{b}$; and (2) If $\mathbf{A}$ is $p \times n$ and $\mathbf{B}$ is $m \times k$, where both are nonrandom, then $\mathrm{E}(\mathbf{AZB}) = \mathbf{A}\mathrm{E}(\mathbf{Z})\mathbf{B}$.

D-7b Variance-Covariance Matrix

Definition D.16 (Variance-Covariance Matrix). If $\mathbf{y}$ is an $n \times 1$ random vector, its **variance-covariance matrix**, denoted $\mathrm{Var}(\mathbf{y})$, is defined as

$$\mathrm{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix},$$

where $\sigma_j^2 = \mathrm{Var}(y_j)$ and $\sigma_{ij} = \mathrm{Cov}(y_i, y_j)$. In other words, the variance-covariance matrix has the variances of each element of $\mathbf{y}$ down its diagonal, with covariance terms in the off diagonals. Because $\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_j, y_i)$, it immediately follows that a variance-covariance matrix is symmetric.

Properties of Variance. (1) If $\mathbf{a}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'[\mathrm{Var}(\mathbf{y})]\mathbf{a} \geq 0$; (2) If $\mathrm{Var}(\mathbf{a}'\mathbf{y}) > 0$ for all $\mathbf{a} \neq \mathbf{0}$, $\mathrm{Var}(\mathbf{y})$ is positive definite; (3) $\mathrm{Var}(\mathbf{y}) = \mathrm{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$, where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$; (4) If the elements of $\mathbf{y}$ are uncorrelated, $\mathrm{Var}(\mathbf{y})$ is a diagonal matrix. If, in addition, $\mathrm{Var}(y_j) = \sigma^2$ for $j = 1, 2, \ldots, n$, then $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$; and (5) If $\mathbf{A}$ is an $m \times n$ nonrandom matrix and $\mathbf{b}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}[\mathrm{Var}(\mathbf{y})]\mathbf{A}'$.

D-7c Multivariate Normal Distribution

The normal distribution for a random variable was discussed at some length in Math Refresher B. We need to extend the normal distribution to random vectors. We will not provide an expression for the probability distribution function, as we do not need it. It is important to know that a multivariate normal random vector is completely characterized by its mean and its variance-covariance matrix. Therefore, if $\mathbf{y}$ is an $n \times 1$ multivariate normal random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, we write $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We now state several useful properties of the **multivariate normal distribution**.

Properties of the Multivariate Normal Distribution. (1) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then each element of $\mathbf{y}$ is normally distributed; (2) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then y_i and y_j, any two elements of $\mathbf{y}$, are independent if, and only if, they are uncorrelated, that is, $\sigma_{ij} = 0$; (3) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Ay} + \mathbf{b} \sim \mathrm{Normal}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, where $\mathbf{A}$ and $\mathbf{b}$ are nonrandom; (4) If $\mathbf{y} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$,

then, for nonrandom matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{Ay}$ and $\mathbf{By}$ are independent if, and only if, $\mathbf{A\Sigma B}' = \mathbf{0}$. In particular, if $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$, then $\mathbf{AB}' = \mathbf{0}$ is necessary and sufficient for independence of $\mathbf{Ay}$ and $\mathbf{By}$; (5) If $\mathbf{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$, $\mathbf{A}$ is a $k \times n$ nonrandom matrix, and $\mathbf{B}$ is an $n \times n$ symmetric, idempotent matrix, then $\mathbf{Ay}$ and $\mathbf{y}'\mathbf{By}$ are independent if, and only if, $\mathbf{AB} = \mathbf{0}$; and (6) If $\mathbf{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are nonrandom symmetric, idempotent matrices, then $\mathbf{y}'\mathbf{Ay}$ and $\mathbf{y}'\mathbf{By}$ are independent if, and only if, $\mathbf{AB} = \mathbf{0}$.

D-7d Chi-Square Distribution

In Math Refresher B, we defined a **chi-square random variable** as the sum of *squared* independent standard normal random variables. In vector notation, if $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, then $\mathbf{u}'\mathbf{u} \sim \chi^2_n$.

Properties of the Chi-Square Distribution. (1) If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix with $\text{rank}(\mathbf{A}) = q$, then $\mathbf{u}'\mathbf{Au} \sim \chi^2_q$; (2) If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ symmetric, idempotent matrices such that $\mathbf{AB} = \mathbf{0}$, then $\mathbf{u}'\mathbf{Au}$ and $\mathbf{u}'\mathbf{Bu}$ are independent, chi-square random variables; and (3) If $\mathbf{z} \sim \text{Normal}(\mathbf{0},\mathbf{C})$, where $\mathbf{C}$ is an $m \times m$ nonsingular matrix, then $\mathbf{z}'\mathbf{C}^{-1}\mathbf{z} \sim \chi^2_m$.

D-7e *t* Distribution

We also defined the ***t* distribution** in Math Refresher B. Now we add an important property.

Property of the t Distribution. If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, $\mathbf{c}$ is an $n \times 1$ nonrandom vector, $\mathbf{A}$ is a nonrandom $n \times n$ symmetric, idempotent matrix with rank q, and $\mathbf{Ac} = \mathbf{0}$, then $\{\mathbf{c}'\mathbf{u}/(\mathbf{c}'\mathbf{c})^{1/2}\}/(\mathbf{u}'\mathbf{Au}/q)^{1/2} \sim t_q$.

D-7f *F* Distribution

Recall that an ***F* random variable** is obtained by taking two *independent* chi-square random variables and finding the ratio of each, standardized by degrees of freedom.

Property of the F Distribution. If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ nonrandom symmetric, idempotent matrices with $\text{rank}(\mathbf{A}) = k_1$, $\text{rank}(\mathbf{B}) = k_2$, and $\mathbf{AB} = \mathbf{0}$, then $(\mathbf{u}'\mathbf{Au}/k_1)/(\mathbf{u}'\mathbf{Bu}/k_2) \sim F_{k_1,k_2}$.

Summary

This Advanced Treatment contains a condensed form of the background information needed to study the classical linear model using matrices. Although the material here is self-contained, it is primarily intended as a review for readers who are familiar with matrix algebra and multivariate statistics, and it will be used extensively in Advanced Treatment E.

Key Terms

Chi-Square Random Variable
Column Vector
Diagonal Matrix
Expected Value
F Random Variable
Idempotent Matrix
Identity Matrix
Inverse
Linearly Independent Vectors
Matrix
Matrix Multiplication
Multivariate Normal Distribution
Positive Definite (p.d.)
Positive Semi-Definite (p.s.d.)
Quadratic Form

Random Vector
Rank of a Matrix
Row Vector
Scalar Multiplication
Square Matrix
Symmetric Matrix
t Distribution
Trace of a Matrix
Transpose
Variance-Covariance Matrix
Zero Matrix

Problems

1 (i) Find the product **AB** using

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

(ii) Does **BA** exist?

2 If **A** and **B** are $n \times n$ diagonal matrices, show that $\mathbf{AB} = \mathbf{BA}$.

3 Let **X** be any $n \times k$ matrix. Show that $\mathbf{X}'\mathbf{X}$ is a symmetric matrix.

4 (i) Use the properties of trace to argue that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$ for any $n \times m$ matrix **A**.

(ii) For $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, verify that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$.

5 (i) Use the definition of inverse to prove the following: if **A** and **B** are $n \times n$ nonsingular matrices, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

(ii) If **A**, **B**, and **C** are all $n \times n$ nonsingular matrices, find $(\mathbf{ABC})^{-1}$ in terms of $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$, and $\mathbf{C}^{-1}$.

6 (i) Show that if **A** is an $n \times n$ symmetric, positive semi-definite matrix, then **A** must have nonnegative diagonal elements.

(ii) Show that if **A** is an $n \times n$ symmetric, positive definite matrix, then **A** must have strictly positive diagonal elements.

(iii) Write down a 2×2 symmetric matrix with strictly positive diagonal elements that is *not* positive definite.

7 Let **A** be an $n \times n$ symmetric, positive definite matrix. Show that if **P** is any $n \times n$ nonsingular matrix, then $\mathbf{P}'\mathbf{AP}$ is positive definite.

8 Prove Property 5 of variances for vectors, using Property 3.

9 Let **a** be an $n \times 1$ nonrandom vector and let **u** be an $n \times 1$ random vector with $\text{E}(\mathbf{u}\mathbf{u}') = \mathbf{I}_n$. Show that $\text{E}[\text{tr}(\mathbf{a}\mathbf{u}\mathbf{u}'\mathbf{a}')] = \sum_{i=1}^{n} a_i^2$.

10 Take as given the properties of the chi-square distribution listed in the text. Show how those properties, along with the definition of an *F* random variable, imply the stated property of the *F* distribution (concerning ratios of quadratic forms).

11 Let **X** be an $n \times k$ matrix partitioned as

$$\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2),$$

where $\mathbf{X}_1$ is $n \times k_1$ and $\mathbf{X}_2$ is $n \times k_2$.

(i) Show that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}.$$

What are the dimensions of each of the matrices?

(ii) Let $\mathbf{b}$ be a $k \times 1$ vector, partitioned as

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

where $\mathbf{b}_1$ is $k_1 \times 1$ and $\mathbf{b}_2$ is $k_2 \times 1$. Show that

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_1'\mathbf{X}_2)\mathbf{b}_2 \\ (\mathbf{X}_2'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_2'\mathbf{X}_2)\mathbf{b}_2 \end{pmatrix}.$$

12 (i) Let $\mathbf{A}$ be an $n \times n$ symmetric matrix such that $\mathbf{A}$ and $\mathbf{I}_n - \mathbf{A}$ are both positive semi-definite. Show that $0 \leq a_{ii} \leq 1$ for $i = 1, \ldots, n$, where a_{ii} is the i^{th} diagonal element of $\mathbf{A}$.

(ii) Prove that if $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix then it must be positive semi-definite.

(iii) Prove that the only $n \times n$ symmetric, idempotent matrix that is also invertible is $\mathbf{I}_n$.

Advanced Treatment E

The Linear Regression Model in Matrix Form

This Advanced Treatment derives various results for ordinary least squares estimation of the multiple linear regression model using matrix notation and matrix algebra (see Advanced Treatment D for a summary). The material presented here is much more advanced than that in the text.

E-1 The Model and Ordinary Least Squares Estimation

Throughout this Advanced Treatment, we use the t subscript to index observations and an n to denote the sample size. It is useful to write the multiple linear regression model with k parameters as follows:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, t = 1, 2, \ldots, n, \qquad [E.1]$$

where y_t is the dependent variable for observation t and $x_{tj}, j = 1, 2, \ldots, k$, are the independent variables. As usual, β_0 is the intercept and $\beta_1, \ldots, \beta_k$ denote the slope parameters.

For each t, define a $1 \times (k + 1)$ vector, $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ be the $(k + 1) \times 1$ vector of all parameters. Then, we can write (E.1) as

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t, t = 1, 2, \ldots, n. \qquad [E.2]$$

[Some authors prefer to define $\mathbf{x}_t$ as a column vector, in which case $\mathbf{x}_t$ is replaced with $\mathbf{x}_t'$ in (E.2). Mathematically, it makes more sense to define it as a row vector.] We can write (E.2) in full matrix notation by appropriately defining data vectors and matrices. Let $\mathbf{y}$ denote the $n \times 1$ vector of observations on y: the t^{th} element of $\mathbf{y}$ is y_t. Let $\mathbf{X}$ be the $n \times (k + 1)$ vector of observations on the explanatory variables. In other words, the t^{th} row of $\mathbf{X}$ consists of the vector $\mathbf{x}_t$. Written out in detail,

$$\underset{n \times (k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}.$$

Finally, let $\mathbf{u}$ be the $n \times 1$ vector of unobservable errors or disturbances. Then, we can write (E.2) for all n observations in **matrix notation**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \qquad \text{[E.3]}$$

Remember, because $\mathbf{X}$ is $n \times (k + 1)$ and $\boldsymbol{\beta}$ is $(k + 1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ is $n \times 1$.

Estimation of $\boldsymbol{\beta}$ proceeds by minimizing the sum of squared residuals, as in Section 3-2. Define the sum of squared residuals function for any possible $(k + 1) \times 1$ parameter vector $\mathbf{b}$ as

$$\text{SSR}(\mathbf{b}) \equiv \sum_{t=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2.$$

The $(k + 1) \times 1$ vector of ordinary least squares estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$, minimizes SSR($\mathbf{b}$) over all possible $(k + 1) \times 1$ vectors $\mathbf{b}$. This is a problem in multivariable calculus. For $\hat{\boldsymbol{\beta}}$ to minimize the sum of squared residuals, it must solve the **first order condition**

$$\partial \text{SSR}(\hat{\boldsymbol{\beta}})/\partial \mathbf{b} \equiv 0. \qquad \text{[E.4]}$$

Using the fact that the derivative of $(y_t - \mathbf{x}_t\mathbf{b})^2$ with respect to $\mathbf{b}$ is the $1 \times (k + 1)$ vector $-2(y_t - \mathbf{x}_t\mathbf{b})\mathbf{x}_t$, (E.4) is equivalent to

$$\sum_{t=1}^{n} \mathbf{x}_t'(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}) \equiv \mathbf{0}. \qquad \text{[E.5]}$$

(We have divided by −2 and taken the transpose.) We can write this first order condition as

$$\begin{aligned}
&\sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0 \\
&\sum_{t=1}^{n} x_{t1}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0 \\
&\vdots \\
&\sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0,
\end{aligned}$$

which is identical to the first order conditions in equation (3.13). We want to write these in matrix form to make them easier to manipulate. Using the formula for partitioned multiplication in Advanced Treatment D, we see that (E.5) is equivalent to

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \qquad \text{[E.6]}$$

or

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \qquad \text{[E.7]}$$

It can be shown that (E.7) always has at least one solution. Multiple solutions do not help us, as we are looking for a unique set of OLS estimates given our data set. Assuming that the $(k + 1) \times (k + 1)$ symmetric matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, we can premultiply both sides of (E.7) by $(\mathbf{X}'\mathbf{X})^{-1}$ to solve for the OLS estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad \text{[E.8]}$$

This is the critical formula for matrix analysis of the multiple linear regression model. The assumption that $\mathbf{X}'\mathbf{X}$ is invertible is equivalent to the assumption that rank($\mathbf{X}$) $= (k + 1)$, which means that the columns of $\mathbf{X}$ must be linearly independent. This is the matrix version of MLR.3 in Chapter 3.

Before we continue, (E.8) warrants a word of warning. It is tempting to simplify the formula for $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}.$$

The flaw in this reasoning is that $\mathbf{X}$ is usually not a square matrix, so it cannot be inverted. In other words, we cannot write $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ unless $n = (k + 1)$, a case that virtually never arises in practice.

The $n \times 1$ vectors of OLS fitted values and residuals are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ respectively.}$$

From (E.6) and the definition of $\hat{\mathbf{u}}$, we can see that the first order condition for $\hat{\boldsymbol{\beta}}$ is the same as

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \qquad \text{[E.9]}$$

Because the first column of $\mathbf{X}$ consists entirely of ones, (E.9) implies that the OLS residuals always sum to zero when an intercept is included in the equation and that the sample covariance between each independent variable and the OLS residuals is zero. (We discussed both of these properties in Chapter 3.)

The sum of squared residuals can be written as

$$\text{SSR} = \sum_{t=1}^{n} \hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \qquad \text{[E.10]}$$

All of the algebraic properties from Chapter 3 can be derived using matrix algebra. For example, we can show that the total sum of squares is equal to the explained sum of squares plus the sum of squared residuals [see (3.27)]. The use of matrices does not provide a simpler proof than summation notation, so we do not provide another derivation.

The matrix approach to multiple regression can be used as the basis for a geometrical interpretation of regression. This involves mathematical concepts that are even more advanced than those we covered in Advanced Treatment D. [See Goldberger (1991) or Greene (1997).]

E-1a The Frisch-Waugh Theorem

In Section 3-2, we described a "partialling out" interpretation of the ordinary least squares estimates. We can establish the partialling out interpretation very generally using matrix notation. Partition the $n \times (k + 1)$ matrix $\mathbf{X}$ as

$$\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2),$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and includes the intercept—although that is not required for the result to hold—and $\mathbf{X}_2$ is $n \times k_2$. We still assume that $\mathbf{X}$ has rank $k + 1$, which means $\mathbf{X}_1$ has rank $k_1 + 1$ and $\mathbf{X}_2$ has rank k_2.

Consider the OLS estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ from the (long) regression

$$\mathbf{y} \text{ on } \mathbf{X}_1, \mathbf{X}_2.$$

As we know, the multiple regression coefficients on $\mathbf{X}_2$, $\hat{\boldsymbol{\beta}}_2$, generally differs from $\tilde{\boldsymbol{\beta}}_2$ from the regression $\mathbf{y}$ on $\mathbf{X}_2$. One way to describe the difference is to understand that we can obtain $\hat{\boldsymbol{\beta}}_2$ from a shorter regression, but first we must "partial out" $\mathbf{X}_1$ from $\mathbf{X}_2$. Consider the following two-step method:

(i) Regress (each column of) $\mathbf{X}_2$ on $\mathbf{X}_1$ and obtain the matrix of residuals, say $\ddot{\mathbf{X}}_2$. We can write $\ddot{\mathbf{X}}_2$ as

$$\ddot{\mathbf{X}}_2 = [\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2,$$

where $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ are $n \times n$ symmetric, idempotent matrices.

(ii) Regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ and call the $k_2 \times 1$ vector of coefficient $\ddot{\boldsymbol{\beta}}_2$.

The **Frisch-Waugh (FW) theorem** states that

$$\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2.$$

Importantly, the FW theorem generally says nothing about equality of the estimates from the long regression, $\hat{\boldsymbol{\beta}}_2$, and those from the short regression, $\tilde{\boldsymbol{\beta}}_2$. Usually $\hat{\boldsymbol{\beta}}_2 \neq \tilde{\boldsymbol{\beta}}_2$. However, if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ then $\ddot{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$, in which case $\ddot{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$; then $\hat{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$ follows from FW. It is also worth noting that we obtain $\hat{\boldsymbol{\beta}}_2$ if we also partial $\mathbf{X}_1$ out of $\mathbf{y}$. In other words, let $\ddot{\mathbf{y}}$ be the residuals from regressing $\mathbf{y}$ on $\mathbf{X}_1$, so that

$$\ddot{\mathbf{y}} = \mathbf{M}_1\mathbf{y}.$$

Then $\hat{\boldsymbol{\beta}}_2$ is obtained from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$. It is important to understand that it is not enough to only partial out $\mathbf{X}_1$ from $\mathbf{y}$. The important step is partialling out $\mathbf{X}_1$ from $\mathbf{X}_2$. Problem 6 at the end of this chapter asks you to derive the FW theorem and to investigate some related issues.

Another useful algebraic result is that when we regress $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ and save the residuals, say $\ddot{\mathbf{u}}$, these are identical to the OLS residuals from the original (long) regression:

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \ddot{\mathbf{u}} = \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2,$$

where we have used the FW result $\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2$. We do not obtain the original OLS residuals if we regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ (but we do obtain $\hat{\boldsymbol{\beta}}_2$).

Before the advent of powerful computers, the Frisch-Waugh result was sometimes used as a computational device. Today, the result is more of theoretical interest, and it is very helpful in understanding the mechanics of OLS. For example, recall that in Chapter 10 we used the FW theorem to establish that adding a time trend to a multiple regression is algebraically equivalent to first linearly detrending all of the explanatory variables before running the regression. The FW theorem also can be used in Chapter 14 to establish that the fixed effects estimator, which we introduced as being obtained from OLS on time-demeaned data, can also be obtained from the (long) dummy variable regression.

E-2 Finite Sample Properties of OLS

Deriving the expected value and variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ is facilitated by matrix algebra, but we must show some care in stating the assumptions.

Assumption E.1 **Linear in Parameters**

The model can be written as in (E.3), where $\mathbf{y}$ is an observed $n \times 1$ vector, $\mathbf{X}$ is an $n \times (k + 1)$ observed matrix, and $\mathbf{u}$ is an $n \times 1$ vector of unobserved errors or disturbances.

Assumption E.2 **No Perfect Collinearity**

The matrix $\mathbf{X}$ has rank $(k + 1)$.

This is a careful statement of the assumption that rules out linear dependencies among the explanatory variables. Under Assumption E.2, $\mathbf{X}'\mathbf{X}$ is nonsingular, so $\hat{\boldsymbol{\beta}}$ is unique and can be written as in (E.8).

Assumption E.3 **Zero Conditional Mean**

Conditional on the entire matrix $\mathbf{X}$, each error u_t has zero mean: $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots, n$.

In vector form, Assumption E.3 can be written as

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \qquad [E.11]$$

This assumption is implied by MLR.4 under the random sampling assumption, MLR.2. In time series applications, Assumption E.3 imposes strict exogeneity on the explanatory variables, something discussed at length in Chapter 10. This rules out explanatory variables whose future values are correlated with u_t; in particular, it eliminates lagged dependent variables. Under Assumption E.3, we can condition on the x_{tj} when we compute the expected value of $\hat{\boldsymbol{\beta}}$.

THEOREM E.1

UNBIASEDNESS OF OLS

Under Assumptions E.1, E.2, and E.3, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

***PROOF*:** Use Assumptions E.1 and E.2 and simple algebra to write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}, \qquad [E.12]\end{aligned}$$

where we use the fact that $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. Taking the expectation conditional on $\mathbf{X}$ gives

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta},\end{aligned}$$

because $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ under Assumption E.3. This argument clearly does not depend on the value of $\boldsymbol{\beta}$, so we have shown that $\hat{\boldsymbol{\beta}}$ is unbiased.

To obtain the simplest form of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, we impose the assumptions of homoskedasticity and no serial correlation.

Assumption E.4 (Homoskedasticity)

Conditional on $\mathbf{X}$, the variances are constant:

$$\text{Var}(u_t|\mathbf{X}) = \sigma^2, t = 1, \ldots, n. \square$$

As we discussed throughout the text, especially in Chapters 8 and 12, heteroskedasticity—which is failure of E.4—can never be ruled out for any of the data structures (cross section, time series, panel).

Assumption E.5 (No Serial Correlation)

Conditional on $\mathbf{X}$, the errors are uncorrelated for all $t \neq s$:

$$\text{Cov}(u_t,u_s|\mathbf{X}) = 0, \text{ all } t \neq s. \square$$

Assumption E.5 is automatically satisfied under random sampling, which is why it does not appear until Chapter 10. With time series applications, Assumption E.5 means that the errors or innovations are uncorrelated across time. As we discussed in Chapters 10, 11, and 12, Assumption E.5 can be unrealistic, particularly in models that do not include lags of y_t. (Including, say, y_{t-1} in $\mathbf{x}_t$ is ruled out by Assumption E.3.)

We can combine Assumptions E.4 and E.5 into a simple expression using matrix notation:

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}_n. \qquad [E.13]$$

Under this assumption, the $n \times n$ variance-covariance matrix $\text{Var}(\mathbf{u}|\mathbf{X})$ depends only on a single parameter, σ^2, and we often say that $\mathbf{u}$ has a **scalar variance-covariance matrix**. (The "scalar" is σ^2.)

Assumptions E.1 through E.5 comprise the **Gauss-Markov assumptions.** The statements of the assumptions unify the conditions we used for cross-sectional analysis in Chapter 3 and time series analysis in Chapter 10.

Using the concise expression in (E.13), we can derive the **variance-covariance matrix of the OLS estimator** under the Gauss-Markov Assumptions.

THEOREM E.2

VARIANCE-COVARIANCE MATRIX OF THE OLS ESTIMATOR

Under Assumptions E.1 through E.5,

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad [E.14]$$

PROOF: From the last formula in equation (E.12), we have

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Now, we use equation (E.13) to get

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Expression (E.14) means that the variance of $\hat{\beta}_j$ (conditional on $\mathbf{X}$) is obtained by multiplying σ^2 by the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. For the slope coefficients, we gave an interpretable formula in equation (3.51). Equation (E.14) also tells us how to obtain the covariance between any two OLS estimates: multiply σ^2 by the appropriate off-diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. In Chapter 4, we showed how to avoid explicitly finding covariances for obtaining confidence intervals and hypothesis tests by appropriately rewriting the model.

The Gauss-Markov Theorem, in its full generality, can be proven.

THEOREM E.3

GAUSS-MARKOV THEOREM

Under Assumptions E.1 through E.5, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator.

PROOF: Any other linear estimator of $\boldsymbol{\beta}$ can be written as

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \qquad [E.15]$$

where $\mathbf{A}$ is an $n \times (k+1)$ matrix. In order for $\tilde{\boldsymbol{\beta}}$ to be unbiased conditional on $\mathbf{X}$, $\mathbf{A}$ can consist of nonrandom numbers and functions of $\mathbf{X}$. (For example, $\mathbf{A}$ cannot be a function of $\mathbf{y}$.) To see what further restrictions on $\mathbf{A}$ are needed, write

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \qquad [E.16]$$

Then,

$$\begin{aligned}\text{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \text{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\text{E}(\mathbf{u}|\mathbf{X}) \text{ because } \mathbf{A} \text{ is a function of } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ because } \text{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}.\end{aligned}$$

For $\tilde{\boldsymbol{\beta}}$ to be an unbiased estimator of $\boldsymbol{\beta}$, it must be true that $E(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ for all $(k+1) \times 1$ vectors $\boldsymbol{\beta}$, that is,

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ for all } (k+1) \times 1 \text{ vectors } \boldsymbol{\beta}. \qquad [E.17]$$

Because $\mathbf{A}'\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix, (E.17) holds if, and only if, $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. Equations (E.15) and (E.17) characterize the class of linear, unbiased estimators for $\boldsymbol{\beta}$.

Next, from (E.16), we have

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

by equation (E.13). Therefore,

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ because } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A},\end{aligned}$$

where $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Because $\mathbf{M}$ is symmetric and idempotent, $\mathbf{A}'\mathbf{M}\mathbf{A}$ is positive semi-definite for any $n \times (k+1)$ matrix $\mathbf{A}$. This establishes that the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. Why is this important? Let $\mathbf{c}$ be any $(k+1) \times 1$ vector and consider the linear combination $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \cdots + c_k\beta_k$, which is a scalar. The unbiased estimators of $\mathbf{c}'\boldsymbol{\beta}$ are $\mathbf{c}'\hat{\boldsymbol{\beta}}$ and $\mathbf{c}'\tilde{\boldsymbol{\beta}}$. But

$$\text{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]\mathbf{c} \geq 0,$$

because $[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]$ is p.s.d. Therefore, when it is used for estimating any linear combination of $\boldsymbol{\beta}$, OLS yields the smallest variance. In particular, $\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}_j\mathbf{X})$ for any other linear, unbiased estimator of β_j.

The unbiased estimator of the error variance σ^2 can be written as

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-k-1),$$

which is the same as equation (3.56).

THEOREM E.4

UNBIASEDNESS OF $\hat{\sigma}^2$

Under Assumptions E.1 through E.5, $\hat{\sigma}^2$ is unbiased for σ^2: $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ for all $\sigma^2 > 0$.

PROOF: Write $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{M}\mathbf{X} = \mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}.$$

Because $\mathbf{u}'\mathbf{M}\mathbf{u}$ is a scalar, it equals its trace. Therefore,

$$\begin{aligned}E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X}) &= E[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})|\mathbf{X}] = E[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}] \\ &= \text{tr}[E(\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}E(\mathbf{u}\mathbf{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n-k-1).\end{aligned}$$

The last equality follows from $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \text{tr}(\mathbf{I}_{k+1}) = n - (k+1) = n - k - 1$. Therefore,

$$E(\hat{\sigma}^2|\mathbf{X}) = E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X})/(n-k-1) = \sigma^2.$$

E-3 Statistical Inference

When we add the final classical linear model assumption, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, which leads to the t and F distributions for the standard test statistics covered in Chapter 4.

Assumption E.6 **Normality of Errors**

Conditional on $\mathbf{X}$, the u_t are independent and identically distributed as Normal$(0,\sigma^2)$. Equivalently, $\mathbf{u}$ given $\mathbf{X}$ is distributed as multivariate normal with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}_n$: $u \sim \text{Normal}(0,\sigma^2\mathbf{I}_n)$.

Assumption E.6 implies Assumptions E.3, E.4, and E.5, but it is much stronger because it assumes that each u_t has a Normal$(0, \sigma^2)$ distribution. As a technical point, Assumption E.6 implies that the u_t are actually independent across t rather than merely uncorrelated. From a practical perspective, this distinction is unimportant. Assumptions E.1 through E.6 are the **classical linear model (CLM) assumptions** expressed in matrix terms, and they are usually viewed as the Gauss-Markov assumptions plus normality of the errors.

THEOREM E.5 **NORMALITY OF $\hat{\beta}$**

Under the classical linear model Assumptions E.1 through E.6, $\hat{\boldsymbol{\beta}}$ conditional on $\mathbf{X}$ is distributed as multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem E.5 is the basis for statistical inference involving $\boldsymbol{\beta}$. In fact, along with the properties of the chi-square, t, and F distributions that we summarized in Advanced Treatment D, we can use Theorem E.5 to establish that t statistics have a t distribution under Assumptions E.1 through E.6 (under the null hypothesis) and likewise for F statistics. We illustrate with a proof for the t statistics.

THEOREM E.6 **DISTRIBUTION OF *t* STATISTIC**

Under Assumptions E.1 through E.6,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k-1}, j = 0, 1, \ldots, k.$$

***PROOF*:** The proof requires several steps; the following statements are initially conditional on $\mathbf{X}$. First, by Theorem E.5, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$, where $\text{sd}(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$, and c_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Next, under Assumptions E.1 through E.6, conditional on $\mathbf{X}$,

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}. \qquad \text{[E.18]}$$

This follows because $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma)$, where $\mathbf{M}$ is the $n \times n$ symmetric, idempotent matrix defined in Theorem E.3. But $\mathbf{u}/\sigma \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ by Assumption E.6. It follows from Property 1 for the chi-square distribution in Advanced Treatment D that $(\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma) \sim \chi^2_{n-k-1}$ (because $\mathbf{M}$ has rank $n - k - 1$).

We also need to show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. But $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, and $\hat{\sigma}^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Now, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$ because $\mathbf{X}'\mathbf{M} = \mathbf{0}$. It follows, from Property 5 of the

multivariate normal distribution in Advanced Treatment D, that $\hat{\boldsymbol{\beta}}$ and **Mu** are independent. Because $\hat{\sigma}^2$ is a function of **Mu**, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also independent.

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

which is the ratio of a standard normal random variable and the square root of a $\chi^2_{n-k-1}/(n-k-1)$ random variable. We just showed that these are independent, so, by definition of a t random variable, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has the t_{n-k-1} distribution. Because this distribution does not depend on **X**, it is the unconditional distribution of $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ as well.

From this theorem, we can plug in any hypothesized value for β_j and use the t statistic for testing hypotheses, as usual.

Under Assumptions E.1 through E.6, we can compute what is known as the *Cramer-Rao* lower bound for the variance-covariance matrix of unbiased estimators of $\boldsymbol{\beta}$ (again conditional on **X**) [see Greene (1997, Chapter 4)]. This can be shown to be $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which is exactly the variance-covariance matrix of the OLS estimator. This implies that $\hat{\boldsymbol{\beta}}$ is the **minimum variance unbiased estimator** of $\boldsymbol{\beta}$ (conditional on **X**): $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semi-definite for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$; we no longer have to restrict our attention to estimators linear in **y**.

It is easy to show that the OLS estimator is in fact the maximum likelihood estimator of $\boldsymbol{\beta}$ under Assumption E.6. For each t, the distribution of y_t given **X** is $\text{Normal}(\mathbf{x}_t\boldsymbol{\beta},\sigma^2)$. Because the y_t are independent conditional on **X**, the likelihood function for the sample is obtained from the product of the densities:

$$\prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2}\exp[-(y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)],$$

where Π denotes product. Maximizing this function with respect to $\boldsymbol{\beta}$ and σ^2 is the same as maximizing its natural logarithm:

$$\sum_{t=1}^{n} [-(1/2)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

For obtaining $\hat{\boldsymbol{\beta}}$, this is the same as minimizing $\sum_{t=1}^{n}(y_t - \mathbf{x}_t\boldsymbol{\beta})^2$—the division by $2\sigma^2$ does not affect the optimization—which is just the problem that OLS solves. The estimator of σ^2 that we have used, $\text{SSR}/(n-k)$, turns out not to be the MLE of σ^2; the MLE is SSR/n, which is a biased estimator. Because the unbiased estimator of σ^2 results in t and F statistics with exact t and F distributions under the null, it is always used instead of the MLE.

That the OLS estimator is the MLE under Assumption E.6 implies an interesting robustness property of the MLE based on the normal distribution. The reasoning is simple. We know that the OLS estimator is unbiased under Assumptions E.1 to E.3; normality of the errors is used nowhere in the proof, and neither are Assumptions E.4 and E.5. As the next section shows, the OLS estimator is also consistent without normality, provided the law of large numbers holds (as is widely true). These statistical properties of the OLS estimator imply that the MLE based on the normal log-likelihood function is robust to distributional specification: the distribution can be (almost) anything and yet we still obtain a consistent (and, under E.1 to E.3, unbiased) estimator. As discussed in Section 17-3, a maximum likelihood estimator obtained without assuming the distribution is correct is often called a **quasi-maximum likelihood estimator (QMLE).**

Generally, consistency of the MLE relies on having a correct distribution in order to conclude that it is consistent for the parameters. We have just seen that the normal distribution is a notable exception. There are some other distributions that share this property, including the Poisson distribution—as discussed in Section 17-3. Wooldridge (2010, Chapter 18) discusses some other useful examples.

E-4 Some Asymptotic Analysis

The matrix approach to the multiple regression model can also make derivations of asymptotic properties more concise. In fact, we can give general proofs of the claims in Chapter 11.

We begin by proving the consistency result of Theorem 11.1. Recall that these assumptions contain, as a special case, the assumptions for cross-sectional analysis under random sampling.

Proof of Theorem 11.1. As in Problem E.1 and using Assumption TS.1′ we write the OLS estimator as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'y_t\right) = \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'(\mathbf{x}_t\boldsymbol{\beta} + u_t)\right) \\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \qquad \text{[E.19]} \\ &= \boldsymbol{\beta} + \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right).\end{aligned}$$

Now, by the law of large numbers,

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ and } n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t \xrightarrow{p} \mathbf{0}, \qquad \text{[E.20]}$$

where $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$ is a $(k+1)\times(k+1)$ nonsingular matrix under Assumption TS.2′ and we have used the fact that $\mathrm{E}(\mathbf{x}_t'u_t) = 0$ under Assumption TS.3′. Now, we must use a matrix version of Property PLIM.1 in Math Refresher C. Namely, because $\mathbf{A}$ is nonsingular,

$$\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}. \qquad \text{[E.21]}$$

[Wooldridge (2010, Chapter 3) contains a discussion of these kinds of convergence results.] It now follows from (E.19), (E.20), and (E.21) that

$$\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.$$

This completes the proof.

Next, we sketch a proof of the asymptotic normality result in Theorem 11.2.

Proof of Theorem 11.2. From equation (E.19), we can write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \\ &= \mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) + o_p(1), \qquad \text{[E.22]}\end{aligned}$$

where the term "$o_p(1)$" is a remainder term that converges in probability to zero. This term is equal to $[(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t)^{-1} - \mathbf{A}^{-1}](n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. The term in brackets converges in probability to zero (by the same argument used in the proof of Theorem 11.1), while $(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$ is bounded in probability because it converges to a multivariate normal distribution by the central limit theorem. A well-known result in asymptotic theory is that the product of such terms converges in probability to zero. Further, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ inherits its asymptotic distribution from $\mathbf{A}^{-1}(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. See Wooldridge (2010, Chapter 3) for more details on the convergence results used in this proof.

By the central limit theorem, $n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t$ has an asymptotic normal distribution with mean zero and, say, $(k + 1) \times (k + 1)$ variance-covariance matrix $\mathbf{B}$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has an asymptotic multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. We now show that, under Assumptions TS.4′ and TS.5′, $\mathbf{B} = \sigma^2\mathbf{A}$. (The general expression is useful because it underlies heteroskedasticity-robust and serial correlation-robust standard errors for OLS, of the kind discussed in Chapter 12.) First, under Assumption TS.5′ $\mathbf{x}_t'u_t$ and $\mathbf{x}_s'u_s$ are uncorrelated for $t \neq s$. Why? Suppose $s < t$ for concreteness. Then, by the law of iterated expectations, $\mathrm{E}(\mathbf{x}_t'u_tu_s\mathbf{x}_s) = \mathrm{E}[\mathrm{E}(u_tu_s\mathbf{x}_t'\mathbf{x}_s)|\mathbf{x}_t,\mathbf{x}_s] = \mathrm{E}[0 \cdot \mathbf{x}_t'\mathbf{x}_s] = 0$. The zero covariances imply that the variance of the sum is the sum of the variances. But $\mathrm{Var}(\mathbf{x}_t'u_t) = \mathrm{E}(\mathbf{x}_t'u_tu_t\mathbf{x}_t) = \mathrm{E}(u_y^2\mathbf{x}_t'\mathbf{x}_t)$. By the law of iterated expectations, $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t|\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2|\mathbf{x}_t)\mathbf{x}_t'\mathbf{x}_t] = \mathrm{E}[\sigma^2\mathbf{x}_t'\mathbf{x}_t] = \sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathbf{A}$, where we use $\mathrm{E}(u_t^2|\mathbf{x}_t) = \sigma^2$ under Assumptions TS.3′ and TS.4′. This shows that $\mathbf{B} = \sigma^2\mathbf{A}$, and so, under Assumptions TS.1′ to TS.5′, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} \mathrm{Normal}(\mathbf{0},\sigma^2\mathbf{A}^{-1}). \qquad \text{[E.23]}$$

This completes the proof.

From equation (E.23), we treat $\hat{\boldsymbol{\beta}}$ as if it is approximately normally distributed with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2\mathrm{A}^{-1}/n$. The division by the sample size, n, is expected here: the approximation to the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ shrinks to zero at the rate $1/n$. When we replace σ^2 with its consistent estimator, $\hat{\sigma}^2 = \mathrm{SSR}/(n - k - 1)$, and replace A with its consistent estimator, $n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t = \mathbf{X}'\mathbf{X}/n$, we obtain an estimator for the asymptotic variance of $\hat{\boldsymbol{\beta}}$:

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.24]}$$

Notice how the two divisions by n cancel, and the right-hand side of (E.24) is just the usual way we estimate the variance matrix of the OLS estimator under the Gauss-Markov assumptions. To summarize, we have shown that, under Assumptions TS.1′ to TS.5′—which contain MLR.1 to MLR.5 as special cases—the usual standard errors and t statistics are asymptotically valid. It is perfectly legitimate to use the usual t distribution to obtain critical values and p-values for testing a single hypothesis. Interestingly, in the general setup of Chapter 11, assuming normality of the errors—say, u_t given $\mathbf{x}_t$, u_{t-1}, $\mathbf{x}_{t-1}$, . . . , u_1, $\mathbf{x}_1$ is distributed as Normal$(0, \sigma^2)$—does not necessarily help, as the t statistics would not generally have exact t statistics under this kind of normality assumption. When we do not assume strict exogeneity of the explanatory variables, exact distributional results are difficult, if not impossible, to obtain.

If we modify the argument above, we can derive a heteroskedasticity-robust, variance-covariance matrix. The key is that we must estimate $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$ separately because this matrix no longer equals $\sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$. But, if the $\hat{u}_t$ are the OLS residuals, a consistent estimator is

$$(n - k - 1)^{-1}\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t, \qquad \text{[E.25]}$$

where the division by $n - k - 1$ rather than n is a degrees of freedom adjustment that typically helps the finite sample properties of the estimator. When we use the expression in equation (E.25), we obtain

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) = [n/(n - k - 1)](\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t\right)(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.26]}$$

The square roots of the diagonal elements of this matrix are the same heteroskedasticity-robust standard errors we obtained in Section 8-2 for the pure cross-sectional case. A matrix extension of the serial correlation- (and heteroskedasticity-) robust standard errors we obtained in Section 12-5 is also available, but the matrix that must replace (E.25) is complicated because of the serial correlation. See, for example, Hamilton (1994, Section 10-5).

E-4a Wald Statistics for Testing Multiple Hypotheses

Similar arguments can be used to obtain the asymptotic distribution of the **Wald statistic** for testing multiple hypotheses. Let $\mathbf{R}$ be a $q \times (k+1)$ matrix, with $q \leq (k+1)$. Assume that the q restrictions on the $(k+1) \times 1$ vector of parameters, $\boldsymbol{\beta}$, can be expressed as H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{r}$ is a $q \times 1$ vector of known constants. Under Assumptions TS.1′ to TS.5′, it can be shown that, under H_0,

$$[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \overset{a}{\sim} \chi^2_q, \qquad \text{[E.27]}$$

where $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$, as in the proofs of Theorems 11.1 and 11.2. The intuition behind equation (E.25) is simple. Because $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is roughly distributed as Normal$(\mathbf{0},\sigma^2\mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is approximately Normal$(0,\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ by Property 3 of the multivariate normal distribution in Advanced Treatment D. Under H_0, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, so $\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \overset{a}{\sim}$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ under H_0. By Property 3 of the chi-square distribution, $z'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}z \sim \chi^2_q$ if $z \sim$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. To obtain the final result formally, we need to use an asymptotic version of this property, which can be found in Wooldridge (2010, Chapter 3).

Given the result in (E.25), we obtain a computable statistic by replacing $\mathbf{A}$ and σ^2 with their consistent estimators; doing so does not change the asymptotic distribution. The result is the so-called Wald statistic, which, after canceling the sample sizes and doing a little algebra, can be written as

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \qquad \text{[E.28]}$$

Under H_0, $W \overset{a}{\sim} \chi^2_q$, where we recall that q is the number of restrictions being tested. If $\hat{\sigma}^2 = \mathrm{SSR}/(n-k-1)$, it can be shown that W/q is exactly the F statistic we obtained in Chapter 4 for testing multiple linear restrictions. [See, for example, Greene (1997, Chapter 7).] Therefore, under the classical linear model assumptions TS.1 to TS.6 in Chapter 10, W/q has an exact $F_{q,n-k-1}$ distribution. Under Assumptions TS.1′ to TS.5′, we only have the asymptotic result in (E.26). Nevertheless, it is appropriate, and common, to treat the usual F statistic as having an approximate $F_{q,n-k-1}$ distribution.

A Wald statistic that is robust to heteroskedasticity of unknown form is obtained by using the matrix in (E.26) in place of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, and similarly for a test statistic robust to both heteroskedasticity and serial correlation. The robust versions of the test statistics cannot be computed via sums of squared residuals or R-squareds from the restricted and unrestricted regressions.

Summary

This Advanced Treatment has provided a brief treatment of the linear regression model using matrix notation. This material is included for more advanced classes that use matrix algebra, but it is not needed to read the text. In effect, this Advanced Treatment proves some of the results that we either stated without proof, proved only in special cases, or proved through a more cumbersome method of proof. Other topics—such as asymptotic properties, instrumental variables estimation, and panel data models—can be given concise treatments using matrices. Advanced texts in econometrics, including Davidson and MacKinnon (1993), Greene (1997), Hayashi (2000), and Wooldridge (2010), can be consulted for details.

Key Terms

Classical Linear Model (CLM) Assumptions
First Order Condition
Frisch-Waugh (FW) theorem
Gauss-Markov Assumptions
Matrix Notation
Minimum Variance Unbiased Estimator
Scalar Variance-Covariance Matrix
Variance-Covariance Matrix of the OLS Estimator
Wald Statistic
Quasi-Maximum Likelihood Estimator (QMLE)

Problems

1 Let $\mathbf{x}_t$ be the $1 \times (k + 1)$ vector of explanatory variables for observation t. Show that the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n} \mathbf{x}_t'y_t\right).$$

Dividing each summation by n shows that $\hat{\boldsymbol{\beta}}$ is a function of sample averages.

2 Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimates.

(i) Show that for any $(k + 1) \times 1$ vector $\mathbf{b}$, we can write the sum of squared residuals as

$$\text{SSR}(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

{*Hint*: Write $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = [\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ and use the fact that $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.}

(ii) Explain how the expression for SSR($\mathbf{b}$) in part (i) proves that $\hat{\boldsymbol{\beta}}$ uniquely minimizes SSR($\mathbf{b}$) over all possible values of $\mathbf{b}$, assuming $\mathbf{X}$ has rank $k + 1$.

3 Let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the regression of $\mathbf{y}$ on $\mathbf{X}$. Let $\mathbf{A}$ be a $(k + 1) \times (k + 1)$ nonsingular matrix and define $\mathbf{z}_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Therefore, $\mathbf{z}_t$ is $1 \times (k + 1)$ and is a nonsingular linear combination of $\mathbf{x}_t$. Let $\mathbf{Z}$ be the $n \times (k + 1)$ matrix with rows $\mathbf{z}_t$. Let $\tilde{\boldsymbol{\beta}}$ denote the OLS estimate from a regression of $\mathbf{y}$ on $\mathbf{Z}$.

(i) Show that $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.

(ii) Let $\hat{y}_t$ be the fitted values from the original regression and let $\tilde{y}_t$ be the fitted values from regressing $\mathbf{y}$ on $\mathbf{Z}$. Show that $\tilde{y}_t = \hat{y}_t$, for all $t = 1, 2, \ldots, n$. How do the residuals from the two regressions compare?

(iii) Show that the estimated variance matrix for $\tilde{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1'}$, where $\hat{\sigma}^2$ is the usual variance estimate from regressing $\mathbf{y}$ on $\mathbf{X}$.

(iv) Let the $\hat{\beta}_j$ be the OLS estimates from regressing y_t on $1, x_{t1}, \ldots, x_{tk}$, and let the $\tilde{\beta}_j$ be the OLS estimates from the regression of y_t on $1, a_1x_{t1}, \ldots, a_kx_{tk}$, where $a_i \neq 0$, $j = 1, \ldots, k$. Use the results from part (i) to find the relationship between the $\tilde{\beta}_j$ and the $\hat{\beta}_j$.

(v) Assuming the setup of part (iv), use part (iii) to show that $\text{se}(\tilde{\beta}_j) = \text{se}(\hat{\beta}_j)/|a_j|$.

(vi) Assuming the setup of part (iv), show that the absolute values of the t statistics for $\tilde{\beta}_j$ and $\hat{\beta}_j$ are identical.

4 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions, let $\mathbf{G}$ be a $(k + 1) \times (k + 1)$ nonsingular, nonrandom matrix, and define $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, so that $\boldsymbol{\delta}$ is also a $(k + 1) \times 1$ vector. Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimators and define $\hat{\boldsymbol{\delta}} = \mathbf{G}\hat{\boldsymbol{\beta}}$ as the OLS estimator of $\boldsymbol{\delta}$.

(i) Show that $\text{E}(\hat{\boldsymbol{\delta}}|\mathbf{X}) = \boldsymbol{\delta}$.

(ii) Find $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ in terms of σ^2, $\mathbf{X}$, and $\mathbf{G}$.

(iii) Use Problem E.3 to verify that $\hat{\boldsymbol{\delta}}$ and the appropriate estimate of $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ are obtained from the regression of $\mathbf{y}$ on $\mathbf{X}\mathbf{G}^{-1}$.

(iv) Now, let $\mathbf{c}$ be a $(k + 1) \times 1$ vector with at least one nonzero entry. For concreteness, assume that $c_k \neq 0$. Define $\theta = \mathbf{c}'\boldsymbol{\beta}$, so that θ is a scalar. Define $\delta_j = \boldsymbol{\beta}_j$, $j = 0, 1, \ldots, k - 1$ and $\delta_k = \theta$. Show how to define a $(k + 1) \times (k + 1)$ nonsingular matrix $\mathbf{G}$ so that $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$. (*Hint*: Each of the first k rows of $\mathbf{G}$ should contain k zeros and a one. What is the last row?)

(v) Show that for the choice of $\mathbf{G}$ in part (iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & 1 & 0 \\ -c_0/c_k & -c_1/c_k & \cdot & \cdot & \cdot & -c_{k-1}/c_k & 1/c_k \end{bmatrix}$$

Use this expression for $\mathbf{G}^{-1}$ and part (iii) to conclude that $\hat{\theta}$ and its standard error are obtained as the coefficient on x_{tk}/c_k in the regression of

$$y_t \text{ on } [1 - (c_0/c_k)x_{tk}], [x_{t1} - (c_1/c_k)x_{tk}], \ldots, [x_{t,k-1} - (c_{k-1}/c_k)x_{tk}], x_{tk}/c_k, t = 1, \ldots, n.$$

This regression is exactly the one obtained by writing β_k in terms of θ and $\beta_0, \beta_1, \ldots, \beta_{k-1}$, plugging the result into the original model, and rearranging. Therefore, we can formally justify the trick we use throughout the text for obtaining the standard error of a linear combination of parameters.

5 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions and let $\hat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$. Let $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ be an $n \times (k+1)$ matrix function of $\mathbf{X}$ and assume that $\mathbf{Z}'\mathbf{X}$[a $(k+1) \times (k+1)$ matrix] is nonsingular. Define a new estimator of $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

(i) Show that $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, so that $\tilde{\boldsymbol{\beta}}$ is also unbiased conditional on $\mathbf{X}$.

(ii) Find $\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Make sure this is a symmetric, $(k+1) \times (k+1)$ matrix that depends on $\mathbf{Z}$, $\mathbf{X}$, and σ^2.

(iii) Which estimator do you prefer, $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$? Explain.

6 Consider the setup of the Frisch-Waugh Theorem.

(i) Using partitioned matrices, show that the first order conditions $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ can be written as

$$\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$$
$$\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}.$$

(ii) Multiply the first set of equations by $\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and subtract the result from the second set of equations to show that

$$(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

where $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Conclude that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\mathbf{y}.$$

(iii) Use part (ii) to show that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\ddot{\mathbf{y}}.$$

(iv) Use the fact that $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ to show that the residuals $\ddot{\mathbf{u}}$ from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ are identical to the residuals $\hat{\mathbf{u}}$ from the regression $\mathbf{y}$ on $\mathbf{X}_1$, $\mathbf{X}_2$. [*Hint*: By definition and the FW theorem,

$$\ddot{\mathbf{u}} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2).$$

Now you do the rest.]

7 Suppose that the linear model, written in matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

satisfies Assumptions E.1, E.2, and E.3. Partition the model as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and $\mathbf{X}_2$ is $n \times k_2$.

(i) Consider the following proposal for estimating $\boldsymbol{\beta}_2$. First, regress $\mathbf{y}$ on $\mathbf{X}_1$ and obtain the residuals, say, $\ddot{\mathbf{y}}$. Then, regress $\ddot{\mathbf{y}}$ on $\mathbf{X}_2$ to get $\breve{\boldsymbol{\beta}}_2$. Show that $\breve{\boldsymbol{\beta}}_2$ is generally biased and show what the bias is. [You should find $\mathrm{E}(\breve{\boldsymbol{\beta}}_2|\mathbf{X})$ in terms of $\boldsymbol{\beta}_2$, $\mathbf{X}_2$, and the residual-making matrix $\mathbf{M}_1$.]

(ii) As a special case, write

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \beta_k\mathbf{X}_k + \mathbf{u},$$

where $\mathbf{X}_k$ is an $n \times 1$ vector on the variable x_{tk}. Show that

$$\mathrm{E}(\tilde{\beta}_k|\mathbf{X}) = \left(\frac{\mathrm{SSR}_k}{\sum_{t=1}^{n} x_{tk}^2}\right)\beta_k,$$

SSR_k is the sum of squared residuals from regressing x_{tk} on $1, x_{t1}, x_{t2}, \ldots, x_{t,k-1}$. Why is the factor multiplying β_k never greater than one?

(iii) Suppose you know $\boldsymbol{\beta}_1$. Show that the regression $\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$ on $\mathbf{X}_2$ produces an unbiased estimator of $\boldsymbol{\beta}_2$ (conditional on $\mathbf{X}$).

8 In the context of multiple regression, define the $n \times n$ matrix

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

(i) Show that $\mathbf{M}$ is symmetric and idempotent.

(ii) Prove that m_{tt}, the diagonals of the matrix $\mathbf{M}$, satisfy $0 \leq m_{tt} \leq 1$ for $t = 1, 2, \ldots, n$.

(iii) Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov Assumptions. Let $\hat{\mathbf{u}}$ be the vector of OLS residuals. Show that

$$\mathrm{E}(\hat{\mathbf{u}}\hat{\mathbf{u}}'|\mathbf{X}) = \sigma^2\mathbf{M}$$

(iv) Conclude that while the errors $\{u_t: t = 1, 2, \ldots, \mathrm{n}\}$ are homoskedastic and uncorrelated under the Gauss-Markov Assumptions, the OLS residuals are heteroskedastic and correlated.

9 Consider the population model

$$\mathrm{y} = \mathbf{x}\boldsymbol{\beta} + u$$
$$\mathrm{E}(u|\mathbf{x}) = \mathbf{0},$$

where the $1 \times (k + 1)$ vector $\mathbf{x}$ is

$$\mathbf{x} = (1, x_1, x_2, \ldots, x_k).$$

Let $\{(\mathbf{x}_i,\mathrm{y}_i): i = 1, 2, \ldots, \mathrm{n}\}$ be a random sample. Show that Assumptions E.3 and E.5 hold.