Chapter 1

Regression Models

1.1 Introduction

Regression models form the core of the discipline of econometrics. Although econometricians routinely estimate a wide variety of statistical models, using many different types of data, the vast majority of these are either regression models or close relatives of them. In this chapter, we introduce the concept of a regression model, discuss several varieties of them, and introduce the estimation method that is most commonly used with regression models, namely, **least squares**. This estimation method is derived by using the **method of moments**, which is a very general principle of estimation that has many applications in econometrics.

The most elementary type of regression model is the **simple linear regression model**, which can be expressed by the following equation:

$$y_t = \beta_1 + \beta_2 X_t + u_t. \tag{1.01}$$

The subscript t is used to index the **observations** of a **sample**. The total number of observations, also called the **sample size**, will be denoted by n. Thus, for a sample of size n, the subscript t runs from 1 to n. Each observation comprises an observation on a **dependent variable**, written as y_t for observation t, and an observation on a single **explanatory variable**, or **independent variable**, written as X_t.

The relation (1.01) links the observations on the dependent and the explanatory variables for each observation in terms of two unknown **parameters**, β_1 and β_2, and an unobserved **error term**, u_t. Thus, of the five quantities that appear in (1.01), two, y_t and X_t, are observed, and three, β_1, β_2, and u_t, are not. Three of them, y_t, X_t, and u_t, are specific to observation t, while the other two, the parameters, are common to all n observations.

Here is a simple example of how a regression model like (1.01) could arise in economics. Suppose that the index t is a time index, as the notation suggests. Each value of t could represent a year, for instance. Then y_t could be household consumption as measured in year t, and X_t could be measured disposable income of households in the same year. In that case, (1.01) would represent what in elementary macroeconomics is called a **consumption function**.

If for the moment we ignore the presence of the error terms, β_2 is the **marginal propensity to consume** out of disposable income, and β_1 is what is sometimes called **autonomous consumption**. As is true of a great many econometric models, the parameters in this example can be seen to have a direct interpretation in terms of economic theory. The variables, income and consumption, do indeed vary in value from year to year, as the term "variables" suggests. In contrast, the parameters reflect aspects of the economy that do not vary, but take on the same values each year.

The purpose of formulating the model (1.01) is to try to explain the observed values of the dependent variable in terms of those of the explanatory variable. According to (1.01), for each t, the value of y_t is given by a linear function of X_t, plus what we have called the error term, u_t. The linear (strictly speaking, affine[1]) function, which in this case is $\beta_1 + \beta_2 X_t$, is called the **regression function**. At this stage we should note that, as long as we say nothing about the unobserved quantity u_t, (1.01) does not tell us anything. In fact, we can allow the parameters β_1 and β_2 to be quite arbitrary, since, for any given β_1 and β_2, (1.01) can always be made to be true by defining u_t suitably.

If we wish to make sense of the regression model (1.01), then, we must make some assumptions about the properties of the error term u_t. Precisely what those assumptions are will vary from case to case. In all cases, though, it is assumed that u_t is a **random variable**. Most commonly, it is assumed that, whatever the value of X_t, the expectation of the random variable u_t is zero. This assumption usually serves to **identify** the unknown parameters β_1 and β_2, in the sense that, under the assumption, (1.01) can be true only for specific values of those parameters.

The presence of error terms in regression models means that the explanations these models provide are at best partial. This would not be so if the error terms could be directly observed as economic variables, for then u_t could be treated as a further explanatory variable. In that case, (1.01) would be a relation linking y_t to X_t and u_t in a completely unambiguous fashion. Given X_t and u_t, y_t would be completely explained without error.

Of course, error terms are not observed in the real world. They are included in regression models because we are not able to specify all of the real-world factors that determine y_t. When we set up our models with u_t as a random variable, what we are really doing is using the mathematical concept of randomness to model our *ignorance* of the details of economic mechanisms. What we are doing when we suppose that the mean of an error term is zero is supposing that the factors determining y_t that we ignore are just as likely to make y_t bigger than it would have been if those factors were absent as they are to make y_t smaller. Thus we are assuming that, on average, the effects of the neglected determinants tend to cancel out. This does not mean that

[1] A function $g(x)$ is said to be **affine** if it takes the form $g(x) = a + bx$ for two real numbers a and b.

those effects are necessarily small. The proportion of the variation in y_t that is accounted for by the error term will depend on the nature of the data and the extent of our ignorance. Even if this proportion is large, as it will be in some cases, regression models like (1.01) can be useful if they allow us to see how y_t is related to the variables, like X_t, that we can actually observe.

Much of the literature in econometrics, and therefore much of this book, is concerned with how to estimate, and test hypotheses about, the parameters of regression models. In the case of (1.01), these parameters are the **constant term**, or **intercept**, β_1, and the **slope coefficient**, β_2. Although we will begin our discussion of estimation in this chapter, most of it will be postponed until later chapters. In this chapter, we are primarily concerned with understanding regression models as statistical models, rather than with estimating them or testing hypotheses about them.

In the next section, we review some elementary concepts from probability theory, including random variables and their expectations. Many readers will already be familiar with these concepts. They will be useful in Section 1.3, where we discuss the meaning of regression models and some of the forms that such models can take. In Section 1.4, we review some topics from matrix algebra and show how multiple regression models can be written using matrix notation. Finally, in Section 1.5, we introduce the method of moments and show how it leads to ordinary least squares as a way of estimating regression models.

1.2 Distributions, Densities, and Moments

The variables that appear in an econometric model are treated as what statisticians call **random variables**. In order to characterize a random variable, we must first specify the set of all the possible values that the random variable can take on. The simplest case is a **scalar random variable**, or **scalar r.v.** The set of possible values for a scalar r.v. may be the real line or a subset of the real line, such as the set of nonnegative real numbers. It may also be the set of integers or a subset of the set of integers, such as the numbers 1, 2, and 3.

Since a random variable is a collection of possibilities, random variables cannot be observed as such. What we do observe are **realizations** of random variables, a realization being one value out of the set of possible values. For a scalar random variable, each realization is therefore a single real value.

If X is any random variable, **probabilities** can be assigned to subsets of the full set of possibilities of values for X, in some cases to each point in that set. Such subsets are called **events**, and their probabilities are assigned by a **probability distribution**, according to a few general rules.

Discrete and Continuous Random Variables

The easiest sort of probability distribution to consider arises when X is a **discrete random variable**, which can take on a finite, or perhaps a countably infinite number of values, which we may denote as $x_1, x_2, \ldots$. The probability distribution simply assigns probabilities, that is, numbers between 0 and 1, to each of these values, in such a way that the probabilities sum to 1:

$$\sum_{i=1}^{\infty} p(x_i) = 1,$$

where $p(x_i)$ is the probability assigned to x_i. Any assignment of nonnegative probabilities that sum to one automatically respects all the general rules alluded to above.

In the context of econometrics, the most commonly encountered discrete random variables occur in the context of **binary data**, which can take on the values 0 and 1, and in the context of **count data**, which can take on the values 0, 1, 2,...; see Chapter 11.

Another possibility is that X may be a **continuous random variable**, which, for the case of a scalar r.v., can take on any value in some continuous subset of the real line, or possibly the whole real line. The dependent variable in a regression model is normally a continuous r.v. For a continuous r.v., the probability distribution can be represented by a **cumulative distribution function**, or **CDF**. This function, which is often denoted $F(x)$, is defined on the real line. Its value is $\Pr(X \le x)$, the probability of the event that X is equal to or less than some value x. In general, the notation $\Pr(A)$ signifies the probability assigned to the event A, a subset of the full set of possibilities. Since X is continuous, it does not really matter whether we define the CDF as $\Pr(X \le x)$ or as $\Pr(X < x)$ here, but it is conventional to use the former definition.

Notice that, in the preceding paragraph, we used X to denote a random variable and x to denote a realization of X, that is, a particular value that the random variable X may take on. This distinction is important when discussing the meaning of a probability distribution, but it will rarely be necessary in most of this book.

Probability Distributions

We may now make explicit the general rules that must be obeyed by probability distributions in assigning probabilities to events. There are just three of these rules:

(i) All probabilities lie between 0 and 1;

(ii) The null set is assigned probability 0, and the full set of possibilities is assigned probability 1;

(iii) The probability assigned to an event that is the union of two disjoint events is the sum of the probabilities assigned to those disjoint events.

We will not often need to make explicit use of these rules, but we can use them now in order to derive some properties of any well-defined CDF for a scalar r.v. First, a CDF $F(x)$ tends to 0 as $x \to -\infty$. This follows because the event $(X \leq x)$ tends to the null set as $x \to -\infty$, and the null set has probability 0. By similar reasoning, $F(x)$ tends to 1 when $x \to +\infty$, because then the event $(X \leq x)$ tends to the entire real line. Further, $F(x)$ must be a weakly increasing function of x. This is true because, if $x_1 < x_2$, we have

$$(X \leq x_2) = (X \leq x_1) \cup (x_1 < X \leq x_2), \tag{1.02}$$

where $\cup$ is the symbol for set union. The two subsets on the right-hand side of (1.02) are clearly disjoint, and so

$$\Pr(X \leq x_2) = \Pr(X \leq x_1) + \Pr(x_1 < X \leq x_2).$$

Since all probabilities are nonnegative, it follows that the probability that $(X \leq x_2)$ must be no smaller than the probability that $(X \leq x_1)$.

For a continuous r.v., the CDF assigns probabilities to every interval on the real line. However, if we try to assign a probability to a single point, the result is always just zero. Suppose that X is a scalar r.v. with CDF $F(x)$. For any interval $[a, b]$ of the real line, the fact that $F(x)$ is weakly increasing allows us to compute the probability that $X \in [a, b]$. If $a < b$,

$$\Pr(X \leq b) = \Pr(X \leq a) + \Pr(a < X \leq b),$$

whence it follows directly from the definition of a CDF that

$$\Pr(a \leq X \leq b) = F(b) - F(a), \tag{1.03}$$

since, for a continuous r.v., we make no distinction between $\Pr(a < X \leq b)$ and $\Pr(a \leq X \leq b)$. If we set $b = a$, in the hope of obtaining the probability that $X = a$, then we get $F(a) - F(a) = 0$.

Probability Density Functions

For continuous random variables, the concept of a **probability density function**, or **PDF**, is very closely related to that of a CDF. Whereas a distribution function exists for any well-defined random variable, a PDF exists only when the random variable is continuous, and when its CDF is differentiable. For a scalar r.v., the density function, often denoted by f, is just the derivative of the CDF:

$$f(x) \equiv F'(x).$$

Because $F(-\infty) = 0$ and $F(\infty) = 1$, every PDF must be **normalized** to integrate to unity. By the Fundamental Theorem of Calculus,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} F'(x)\,dx = F(\infty) - F(-\infty) = 1. \tag{1.04}$$

It is obvious that a PDF is nonnegative, since it is the derivative of a weakly increasing function.

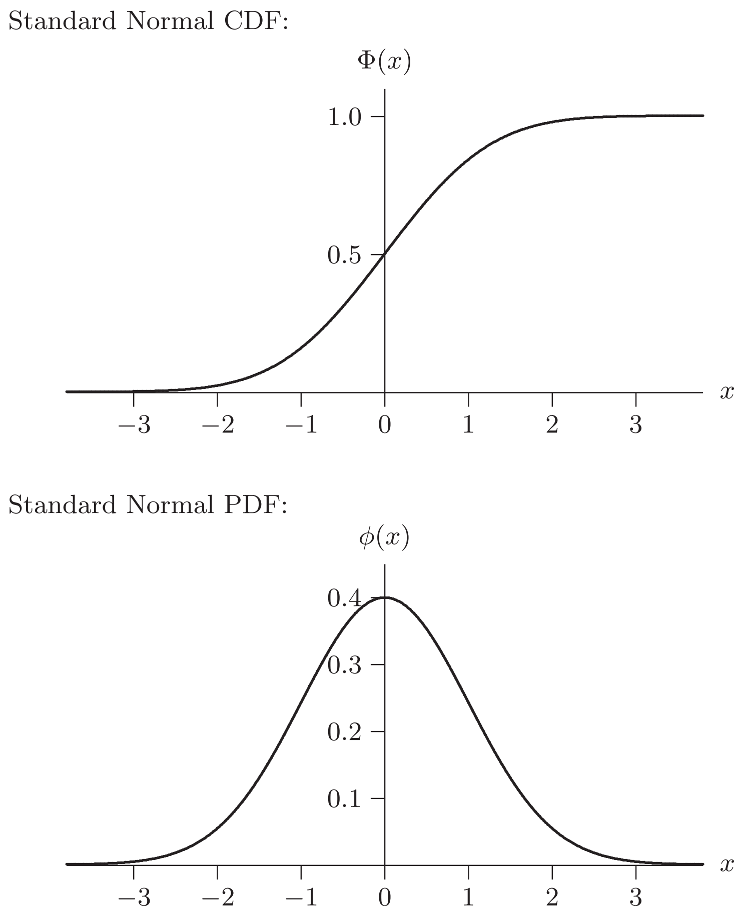


Figure 1.1 The CDF and PDF of the standard normal distribution

Probabilities can be computed in terms of the PDF as well as the CDF. Note that, by (1.03) and the Fundamental Theorem of Calculus once more,

$$\Pr(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\,dx. \tag{1.05}$$

Since (1.05) must hold for arbitrary a and b, it is clear why $f(x)$ must always be nonnegative. However, it is important to remember that $f(x)$ is not bounded above by unity, because the value of a PDF at a point x is not a probability. Only when a PDF is integrated over some interval, as in (1.05), does it yield a probability.

The most common example of a continuous distribution is provided by the **normal distribution**. This is the distribution that generates the famous or infamous "bell curve" sometimes thought to influence students' grade distributions. The fundamental member of the normal family of distributions is the **standard normal distribution**. It is a continuous scalar distribution, defined

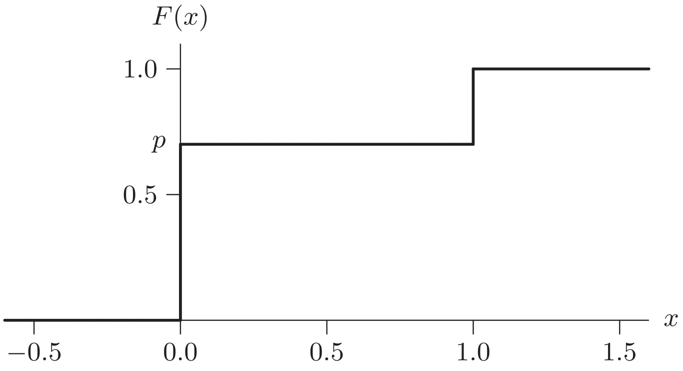


Figure 1.2 The CDF of a binary random variable

on the entire real line. The PDF of the standard normal distribution is often denoted $\phi(\cdot)$. Its explicit expression, which we will need later in the book, is

$$\phi(x) = (2\pi)^{-1/2} \exp\bigl(-\tfrac{1}{2}x^2\bigr). \tag{1.06}$$

Unlike $\phi(\cdot)$, the CDF, usually denoted $\Phi(\cdot)$, has no elementary closed-form expression. However, by (1.05) with $a = -\infty$ and $b = x$, we have

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\,dy.$$

The functions $\Phi(\cdot)$ and $\phi(\cdot)$ are graphed in Figure 1.1. Since the PDF is the derivative of the CDF, it achieves a maximum at $x = 0$, where the CDF is rising most steeply. As the CDF approaches both 0 and 1, and consequently, becomes very flat, the PDF approaches 0.

Although it may not be obvious at once, discrete random variables can be characterized by a CDF just as well as continuous ones can be. Consider a binary r.v. X that can take on only two values, 0 and 1, and let the probability that $X = 0$ be p. It follows that the probability that $X = 1$ is $1-p$. Then the CDF of X, according to the definition of $F(x)$ as $\Pr(X \le x)$, is the following discontinuous, "staircase" function:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ p & \text{for } 0 \le x < 1 \\ 1 & \text{for } x \ge 1. \end{cases}$$

This CDF is graphed in Figure 1.2. Obviously, we cannot graph a corresponding PDF, for it does not exist. For general discrete random variables, the discontinuities of the CDF occur at the discrete permitted values of X, and the jump at each discontinuity is equal to the probability of the corresponding value. Since the sum of the jumps is therefore equal to 1, the limiting value of F, to the right of all permitted values, is also 1.

Using a CDF is a reasonable way to deal with random variables that are neither completely discrete nor completely continuous. Such hybrid variables can be produced by the phenomenon of **censoring**. A random variable is said to be censored if not all of its potential values can actually be observed. For instance, in some data sets, a household's measured income is set equal to 0 if it is actually negative. It might be negative if, for instance, the household lost more on the stock market than it earned from other sources in a given year. Even if the true income variable is continuously distributed over the positive and negative real line, the observed, censored, variable will have an **atom**, or bump, at 0, since the single value of 0 now has a nonzero probability attached to it, namely, the probability that an individual's income is nonpositive. As with a purely discrete random variable, the CDF will have a discontinuity at 0, with a jump equal to the probability of a negative or zero income.

Moments of Random Variables

A fundamental property of a random variable is its **expectation**. For a discrete r.v. that can take on m possible finite values $x_1, x_2, \ldots, x_m$, the expectation is simply

$$\mathrm{E}(X) \equiv \sum_{i=1}^{m} p(x_i)\,x_i. \tag{1.07}$$

Thus each possible value x_i is multiplied by the probability associated with it. If m is infinite, the sum above has an infinite number of terms.

For a continuous r.v., the expectation is defined analogously using the PDF:

$$\mathrm{E}(X) \equiv \int_{-\infty}^{\infty} x\,f(x)\,dx. \tag{1.08}$$

Not every r.v. has an expectation, however. The integral of a density function always exists and equals 1. But since X can range from $-\infty$ to ∞, the integral (1.08) may well diverge at either limit of integration, or both, if the density f does not tend to zero fast enough. Similarly, if m in (1.07) is infinite, the sum may diverge. The expectation of a random variable is sometimes called the **mean** or, to prevent confusion with the usual meaning of the word as the mean of a sample, the **population mean**. A common notation for it is μ.

The expectation of a random variable is often referred to as its **first moment**. The so-called **higher moments**, if they exist, are the expectations of the r.v. raised to a power. Thus the **second moment** of a random variable X is the expectation of X^2, the **third moment** is the expectation of X^3, and so on. In general, the k^{th} moment of a continuous random variable X is

$$m_k(X) \equiv \int_{-\infty}^{\infty} x^k f(x)\,dx.$$

Observe that the value of any moment depends only on the probability distribution of the r.v. in question. For this reason, we often speak of the moments of the distribution rather than the moments of a specific random variable. If a distribution possesses a $k^{\rm th}$ moment, it also possesses all moments of order less than k.

The higher moments just defined are called the **uncentered moments** of a distribution, because, in general, X does not have mean zero. It is often more useful to work with the **central moments**, which are defined as the ordinary moments of the difference between the random variable and its expectation. Thus the $k^{\rm th}$ central moment of the distribution of a continuous r.v. X is

$$\mu_k \equiv {\rm E}\big(X - {\rm E}(X)\big)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx,$$

where $\mu \equiv {\rm E}(X)$. For a discrete X, the $k^{\rm th}$ central moment is

$$\mu_k \equiv {\rm E}\big(X - {\rm E}(X)\big)^k = \sum_{i=1}^{m} p(x_i)(x_i-\mu)^k.$$

By far the most important central moment is the second. It is called the **variance** of the random variable and is frequently written as ${\rm Var}(X)$. Another common notation for a variance is σ^2. This notation underlines the important fact that a variance cannot be negative. The square root of the variance, σ, is called the **standard deviation** of the distribution. Estimates of standard deviations are often referred to as **standard errors**, especially when the random variable in question is an estimated parameter.

Multivariate Distributions

A **vector-valued random variable** takes on values that are vectors. It can be thought of as several scalar random variables that have a single, joint distribution. For simplicity, we will focus on the case of **bivariate random variables**, where the vector is of length 2. A continuous, bivariate r.v. (X_1, X_2) has a distribution function

$$F(x_1, x_2) = {\rm Pr}\big((X_1 \le x_1) \cap (X_2 \le x_2)\big),$$

where $\cap$ is the symbol for set intersection. Thus $F(x_1, x_2)$ is the joint probability that both $X_1 \le x_1$ and $X_2 \le x_2$. For continuous variables, the PDF, if it exists, is the **joint density function**[2]

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}. \qquad (1.09)$$

[2] Here we are using what computer scientists would call “overloaded function” notation. This means that $F(\cdot)$ and $f(\cdot)$ denote respectively the CDF and the PDF of whatever their argument(s) happen to be. This practice is harmless provided there is no ambiguity.

This function has exactly the same properties as an ordinary PDF. In particular, as in (1.04),

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 dx_2 = 1.$$

More generally, the probability that X_1 and X_2 jointly lie in any region is the integral of $f(x_1, x_2)$ over that region. A case of particular interest is

$$\begin{aligned} F(x_1, x_2) &= \Pr\bigl((X_1 \leq x_1) \cap (X_2 \leq x_2)\bigr) \\ &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(y_1, y_2)\, dy_1 dy_2, \end{aligned} \tag{1.10}$$

which shows how to compute the CDF given the PDF.

The concept of joint probability distributions leads naturally to the important notion of **statistical independence**. Let (X_1, X_2) be a bivariate random variable. Then X_1 and X_2 are said to be **statistically independent**, or often just **independent**, if the joint CDF of (X_1, X_2) is the product of the CDFs of X_1 and X_2. In straightforward notation, this means that

$$F(x_1, x_2) = F(x_1, \infty)F(\infty, x_2). \tag{1.11}$$

The first factor here is the joint probability that $X_1 \leq x_1$ and $X_2 \leq \infty$. Since the second inequality imposes no constraint, this factor is just the probability that $X_1 \leq x_1$. The function $F(x_1, \infty)$, which is called the **marginal CDF** of X_1, is thus just the CDF of X_1 considered by itself. Similarly, the second factor on the right-hand side of (1.11) is the marginal CDF of X_2.

It is also possible to express statistical independence in terms of the **marginal density** of X_1 and the marginal density of X_2. The marginal density of X_1 is, as one would expect, the derivative of the marginal CDF of X_1,

$$f(x_1) \equiv F_1(x_1, \infty),$$

where $F_1(\cdot)$ denotes the partial derivative of $F(\cdot)$ with respect to its first argument. It can be shown from (1.10) that the marginal density can also be expressed in terms of the joint density, as follows:

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2. \tag{1.12}$$

Thus $f(x_1)$ is obtained by integrating X_2 out of the joint density. Similarly, the marginal density of X_2 is obtained by integrating X_1 out of the joint density. From (1.09), it can be shown that, if X_1 and X_2 are independent, so that (1.11) holds, then

$$f(x_1, x_2) = f(x_1)f(x_2). \tag{1.13}$$

Thus, when densities exist, statistical independence means that the joint density factorizes as the product of the marginal densities, just as the joint CDF factorizes as the product of the marginal CDFs.

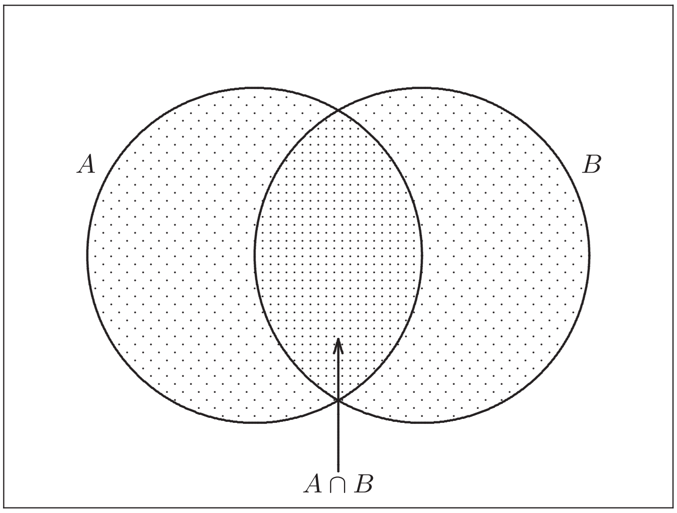


Figure 1.3 Conditional probability

Conditional Probabilities

Suppose that A and B are any two events. Then the probability of event A **conditional** on B, or given B, is denoted as $\Pr(A \mid B)$ and is defined implicitly by the equation

$$\Pr(A \cap B) = \Pr(B) \Pr(A \mid B). \tag{1.14}$$

For this equation to make sense as a definition of $\Pr(A \mid B)$, it is necessary that $\Pr(B) \neq 0$. The idea underlying the definition is that, if we know somehow that the event B has been realized, this knowledge can provide information about whether event A has also been realized. For instance, if A and B are disjoint, and B is realized, then it is certain that A has not been. As we would wish, this does indeed follow from the definition (1.14), since $A \cap B$ is the null set, of zero probability, if A and B are disjoint. Similarly, if B is a subset of A, knowing that B has been realized means that A must have been realized as well. Since in this case $\Pr(A \cap B) = \Pr(B)$, (1.14) tells us that $\Pr(A \mid B) = 1$, as required.

To gain a better understanding of (1.14), consider Figure 1.3. The bounding rectangle represents the full set of possibilities, and events A and B are subsets of the rectangle that overlap as shown. Suppose that the figure has been drawn in such a way that probabilities of subsets are proportional to their areas. Thus the probabilities of A and B are the ratios of the areas of the corresponding circles to the area of the bounding rectangle, and the probability of the intersection $A \cap B$ is the ratio of its area to that of the rectangle.

Suppose now that it is known that B has been realized. This fact leads us to redefine the probabilities so that everything outside B now has zero probability, while, inside B, probabilities remain proportional to areas. Event B

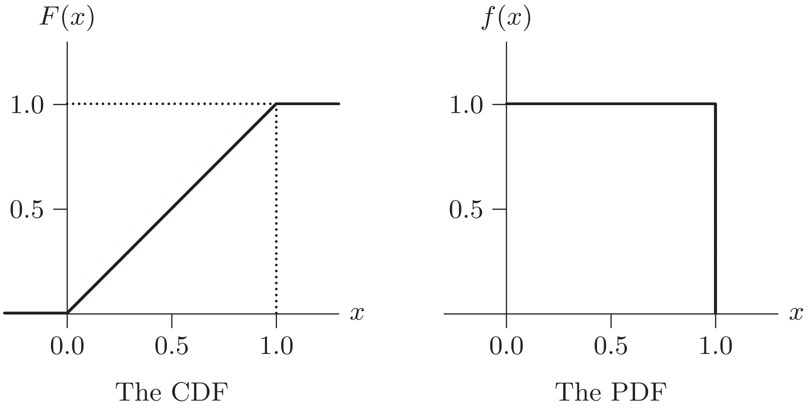


Figure 1.4 The CDF and PDF of the uniform distribution on $[0, 1]$

will now have probability 1, in order to keep the total probability equal to 1. Event A can be realized only if the realized point is in the intersection $A \cap B$, since the set of all points of A outside this intersection have zero probability. The probability of A, conditional on knowing that B has been realized, is thus the ratio of the area of $A \cap B$ to that of B. This construction leads directly to (1.14).

There are many ways to associate a random variable X with the rectangle shown in Figure 1.3. Such a random variable could be any function of the two coordinates that define a point in the rectangle. For example, it could be the horizontal coordinate of the point measured from the origin at the lower left-hand corner of the rectangle, or its vertical coordinate, or the Euclidean distance of the point from the origin. The realization of X is the value of the function it corresponds to at the realized point in the rectangle.

For concreteness, let us assume that the function is simply the horizontal coordinate, and let the width of the rectangle be equal to 1. Then, since all values of the horizontal coordinate between 0 and 1 are equally probable, the random variable X has what is called the **uniform distribution** on the interval $[0, 1]$. The CDF of this distribution is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1. \end{cases}$$

Because $F(x)$ is not differentiable at $x = 0$ and $x = 1$, the PDF of the uniform distribution does not exist at those points. Elsewhere, the derivative of $F(x)$ is 0 outside $[0, 1]$ and 1 inside. The CDF and PDF are illustrated in Figure 1.4. This special case of the uniform distribution is often denoted the $U(0, 1)$ distribution.

If the information were available that B had been realized, then the distribution of X conditional on this information would be very different from the

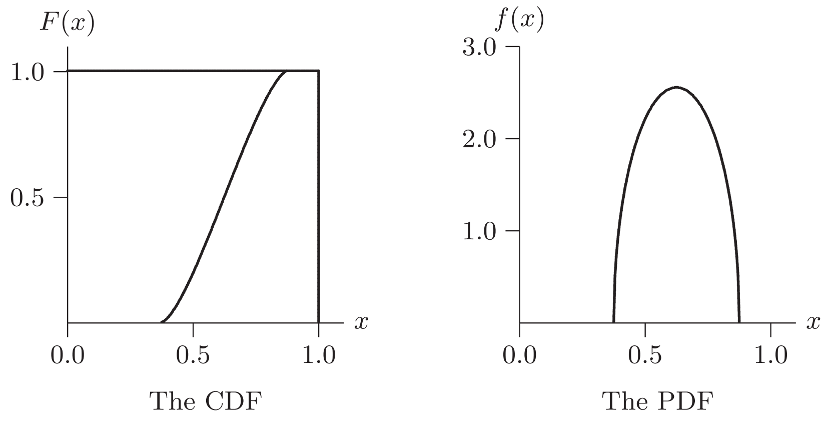


Figure 1.5 The CDF and PDF conditional on event B

$U(0,1)$ distribution. Now only values between the extreme horizontal limits of the circle of B are allowed. If one computes the area of the part of the circle to the left of a given vertical line, then for each event $a \equiv (X \leq x)$ the probability of this event conditional on B can be worked out. The result is just the CDF of X conditional on the event B. Its derivative is the PDF of X conditional on B. These are shown in Figure 1.5.

The concept of conditional probability can be extended beyond probability conditional on an event to probability conditional on a random variable. Suppose that X_1 is a r.v. and X_2 is a discrete r.v. with permitted values $z_1, \ldots, z_m$. For each $i = 1, \ldots, m$, the CDF of X_1, and, if X_1 is continuous, its PDF, can be computed conditional on the event $(X_2 = z_i)$. If X_2 is also a continuous r.v., then things are a little more complicated, because events like $(X_2 = x_2)$ for some real x_2 have zero probability, and so cannot be conditioned on in the manner of (1.14).

On the other hand, it makes perfect intuitive sense to think of the distribution of X_1 conditional on some specific realized value of X_2. This conditional distribution gives us the probabilities of events concerning X_1 when we know that the realization of X_2 was actually x_2. We therefore make use of the **conditional density** of X_1 for a given value x_2 of X_2. This conditional density, or **conditional PDF**, is defined as

$$f(x_1 \,|\, x_2) = \frac{f(x_1, x_2)}{f(x_2)}. \tag{1.15}$$

Thus, for a given value x_2 of X_2, the conditional density is proportional to the joint density of X_1 and X_2. Of course, (1.15) is well defined only if $f(x_2) > 0$. In some cases, more sophisticated definitions can be found that would allow $f(x_1 \,|\, x_2)$ to be defined for all x_2 even if $f(x_2) = 0$, but we will not need these in this book. See, among others, Billingsley (1979).

Conditional Expectations

Whenever we can describe the distribution of a random variable, X_1, conditional on another, X_2, either by a conditional CDF or a conditional PDF, we can consider the **conditional expectation** or **conditional mean** of X_1. If it exists, this conditional expectation is just the ordinary expectation computed using the conditional distribution. If x_2 is a possible value for X_2, then this conditional expectation is written as $\mathrm{E}(X_1 \,|\, x_2)$.

For a given value x_2, the conditional expectation $\mathrm{E}(X_1 \,|\, x_2)$ is, like any other ordinary expectation, a deterministic, that is, nonrandom, quantity. But we can consider the expectation of X_1 conditional on *every* possible realization of X_2. In this way, we can construct a new random variable, which we denote by $\mathrm{E}(X_1 \,|\, X_2)$, the realization of which is $\mathrm{E}(X_1 \,|\, x_2)$ when the realization of X_2 is x_2. We can call $\mathrm{E}(X_1 \,|\, X_2)$ a deterministic function of the random variable X_2, because the realization of $\mathrm{E}(X_1 \,|\, X_2)$ is unambiguously determined by the realization of X_2.

Conditional expectations defined as random variables in this way have a number of interesting and useful properties. The first, called the **Law of Iterated Expectations**, can be expressed as follows:

$$\mathrm{E}\big(\mathrm{E}(X_1 \,|\, X_2)\big) = \mathrm{E}(X_1). \tag{1.16}$$

If a conditional expectation of X_1 can be treated as a random variable, then the conditional expectation itself may have an expectation. According to (1.16), this expectation is just the ordinary expectation of X_1.

Another property of conditional expectations is that any deterministic function of a conditioning variable X_2 is its own conditional expectation. Thus, for example, $\mathrm{E}(X_2 \,|\, X_2) = X_2$, and $\mathrm{E}(X_2^2 \,|\, X_2) = X_2^2$. Similarly, conditional on X_2, the expectation of a product of another random variable X_1 and a deterministic function of X_2 is the product of that deterministic function and the expectation of X_1 conditional on X_2:

$$\mathrm{E}\big(X_1 h(X_2) \,|\, X_2\big) = h(X_2)\,\mathrm{E}(X_1 \,|\, X_2), \tag{1.17}$$

for any deterministic function $h(\cdot)$. An important special case of this, which we will make use of in Section 1.5, arises when $\mathrm{E}(X_1 \,|\, X_2) = 0$. In that case, for any function $h(\cdot)$, $\mathrm{E}(X_1 h(X_2)) = 0$, because

$$\begin{aligned} E\big(X_1 h(X_2)\big) &= \mathrm{E}\big(\mathrm{E}(X_1 h(X_2) \,|\, X_2)\big) \\ &= \mathrm{E}\big(h(X_2)\mathrm{E}(X_1 \,|\, X_2)\big) \\ &= \mathrm{E}(0) = 0. \end{aligned}$$

The first equality here follows from the Law of Iterated Expectations, (1.16). The second follows from (1.17). Since $\mathrm{E}(X_1 \,|\, X_2) = 0$, the third line then follows immediately. We will present other properties of conditional expectations as the need arises.

1.3 The Specification of Regression Models

We now return our attention to the regression model (1.01) and revert to the notation of Section 1.1 in which y_t and X_t respectively denote the dependent and independent variables. The model (1.01) can be interpreted as a model for the mean of y_t conditional on X_t. Let us assume that the error term u_t has mean 0 conditional on X_t. Then, taking conditional expectations of both sides of (1.01), we see that

$$\mathrm{E}(y_t \,|\, X_t) = \beta_1 + \beta_2 X_t + \mathrm{E}(u_t \,|\, X_t) = \beta_1 + \beta_2 X_t.$$

Without the key assumption that $\mathrm{E}(u_t \,|\, X_t) = 0$, the second equality here would not hold. As we pointed out in Section 1.1, it is impossible to make any sense of a regression model unless we make strong assumptions about the error terms. Of course, we could define u_t as the difference between y_t and $\mathrm{E}(y_t \,|\, X_t)$, which would give $\mathrm{E}(u_t \,|\, X_t) = 0$ by definition. But if we require that $\mathrm{E}(u_t \,|\, X_t) = 0$ and also specify (1.01), we must necessarily have $\mathrm{E}(y_t \,|\, X_t) = \beta_1 + \beta_2 X_t$.

As an example, suppose that we estimate the model (1.01) when in fact

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t^2 + v_t \tag{1.18}$$

with $\beta_3 \neq 0$ and an error term v_t such that $\mathrm{E}(v_t \,|\, X_t) = 0$. If the data were generated by (1.18), the error term u_t in (1.01) would be equal to $\beta_3 X_t^2 + v_t$. By the results on conditional expectations in the last section, we see that

$$\mathrm{E}(u_t \,|\, X_t) = \mathrm{E}\big(\beta_3 X_t^2 + v_t \,|\, X_t\big) = \beta_3 X_t^2,$$

which we have assumed to be nonzero. This example shows the force of the assumption that the error term has mean zero conditional on X_t. Unless the mean of y_t conditional on X_t really is a linear function of X_t, the regression function in (1.01) is not **correctly specified**, in the precise sense that (1.01) cannot hold with an error term that has mean zero conditional on X_t. It will become clear in later chapters that estimating incorrectly specified models usually leads to results that are meaningless or, at best, seriously misleading.

Information Sets

In a more general setting, what we are interested in is usually not the mean of y_t conditional on a single explanatory variable X_t but the mean of y_t conditional on a set of potential explanatory variables. This set is often called an **information set**, and it is denoted Ω_t. Typically, the information set will contain more variables than would actually be used in a regression model. For example, it might consist of all the variables observed by the economic agents whose actions determine y_t at the time they make the decisions that cause them to perform those actions. Such an information set could be very large.

As a consequence, much of the art of constructing, or specifying, a regression model is deciding which of the variables that belong to Ω_t should be included in the model and which of the variables should be excluded.

In some cases, economic theory makes it fairly clear what the information set Ω_t should consist of, and sometimes also which variables in Ω_t should make their way into a regression model. In many others, however, it may not be at all clear how to specify Ω_t. In general, we want to condition on **exogenous** variables but not on **endogenous** ones. These terms refer to the *origin* or *genesis* of the variables: An exogenous variable has its origins *outside* the model under consideration, while the mechanism generating an endogenous variable is *inside* the model. When we write a single equation like (1.01), the only endogenous variable allowed is the dependent variable, y_t.

Recall the example of the consumption function that we looked at in Section 1.1. That model seeks to explain household consumption in terms of disposable income, but it makes no claim to explain disposable income, which is simply taken as given. The consumption function model can be correctly specified only if two conditions hold:

(i) The mean of consumption conditional on disposable income is a linear function of the latter.

(ii) Consumption is *not* a variable that contributes to the determination of disposable income.

The second condition means that the origin of disposable income, that is, the mechanism by which disposable income is generated, lies outside the model for consumption. In other words, disposable income is exogenous in that model. If the simple consumption model we have presented is correctly specified, the two conditions above must be satisfied. Needless to say, we do not claim that this model is in fact correctly specified.

It is not always easy to decide just what information set to condition on. As the above example shows, it is often not clear whether or not a variable is exogenous. This sort of question will be discussed in Chapter 8. Moreover, even if a variable clearly is exogenous, we may not want to include it in Ω_t. For example, if the ultimate purpose of estimating a regression model is to use it for forecasting, there may be no point in conditioning on information that will not be available at the time the forecast is to be made.

Error Terms

Whenever we specify a regression model, it is essential to make assumptions about the properties of the error terms. The simplest assumption is that all of the error terms have mean 0, come from the same distribution, and are independent of each other. Although this is a rather strong assumption, it is very commonly made in practice.

Mutual independence of the error terms, when coupled with the assumption that $\mathrm{E}(u_t) = 0$, implies that the mean of u_t is 0 conditional on all of the other error terms u_s, $s \neq t$. However, the implication does not work in the other direction, because the assumption of mutual independence is stronger than the assumption about the conditional means. A very strong assumption which is often made is that the error terms are **independently and identically distributed**, or **IID**. According to this assumption, the error terms are mutually independent, and they are in addition realizations from the same, identical, probability distribution.

When the successive observations are ordered by time, it often seems plausible that an error term will be correlated with neighboring error terms. Thus u_t might well be correlated with u_s when the value of $|t - s|$ is small. This could occur, for example, if there is correlation across time periods of random factors that influence the dependent variable but are not explicitly accounted for in the regression function. This phenomenon is called **serial correlation**, and it often appears to be observed in practice. When there is serial correlation, the error terms cannot be IID because they are not independent.

Another possibility is that the variance of the error terms may be systematically larger for some observations than for others. This will happen if the conditional variance of y_t depends on some of the same variables as the conditional mean. This phenomenon is called **heteroskedasticity**, and it too is often observed in practice. For example, in the case of the consumption function, the variance of consumption may well be higher for households with high incomes than for households with low incomes. When there is heteroskedasticity, the error terms cannot be IID, because they are not identically distributed. It is perfectly possible to take explicit account of both serial correlation and heteroskedasticity, but doing so would take us outside the context of regression models like (1.01).

It may sometimes be desirable to write a regression model like the one we have been studying as

$$\mathrm{E}(y_t \,|\, \Omega_t) = \beta_1 + \beta_2 X_t, \tag{1.19}$$

in order to stress the fact that this is a model for the mean of y_t conditional on a certain information set. However, by itself, (1.19) is just as incomplete a specification as (1.01). In order to see this point, we must now state what we mean by a **complete specification** of a regression model. Probably the best way to do this is to say that a complete specification of any econometric model is one that provides an unambiguous recipe for simulating the model on a computer. After all, if we can use the model to generate simulated data, it must be completely specified.

Simulating Econometric Models

Consider equation (1.01). When we say that we **simulate** this model, we mean that we generate numbers for the dependent variable, y_t, according to equation (1.01). Obviously, one of the first things we must fix for the simulation is the sample size, n. That done, we can generate each of the y_t,

$t = 1, \ldots, n$, by evaluating the right-hand side of the equation n times. For this to be possible, we need to know the value of each variable or parameter that appears on the right-hand side.

If we suppose that the explanatory variable X_t is exogenous, then we simply take it as given. So if, in the context of the consumption function example, we had data on the disposable income of households in some country every year for a period of n years, we could just use those data. Our simulation would then be specific to the country in question and to the time period of the data. Alternatively, it could be that we or some other econometricians had previously specified another model, for the explanatory variable this time, and we could then use simulated data provided by that model.

Besides the explanatory variable, the other elements of the right-hand side of (1.01) are the parameters, β_1 and β_2, and the error term u_t. The key feature of the parameters is that we do not know their true values. We will have more to say about this point in Chapter 3, when we define the twin concepts of models and data-generating processes. However, for purposes of simulation, we could use either values suggested by economic theory or values obtained by estimating the model. Evidently, the simulation results will depend on precisely what values we use.

Unlike the parameters, the error terms cannot be taken as given; instead, we wish to treat them as random. Luckily, it is easy to use a computer to generate "random" numbers by using a program called a **random number generator**; we will discuss these programs in Chapter 4. The "random" numbers generated by computers are not random according to some meanings of the word. For instance, a computer can be made to spit out exactly the same sequence of supposedly random numbers more than once. In addition, a digital computer is a perfectly deterministic device. Therefore, if random means the opposite of deterministic, only computers that are not functioning properly would be capable of generating truly random numbers. Because of this, some people prefer to speak of computer-generated random numbers as **pseudo-random**. However, for the purposes of simulations, the numbers computers provide have all the properties of random numbers that we need, and so we will call them simply random rather than pseudo-random.

Computer-generated random numbers are mutually independent **drawings**, or realizations, from specific probability distributions, usually the uniform $U(0, 1)$ distribution or the standard normal distribution, both of which were defined in Section 1.2. Of course, techniques exist for generating drawings from many other distributions as well, as do techniques for generating drawings that are not independent. For the moment, the essential point is that we must always specify the probability distribution of the random numbers we use in a simulation. It is important to note that specifying the expectation of a distribution, or even the expectation conditional on some other variables, is not enough to specify the distribution in full.

Let us now summarize the various steps in performing a simulation by giving a sort of generic recipe for simulations of regression models. In the model specification, it is convenient to distinguish between the **deterministic specification** and the **stochastic specification**. In model (1.01), the deterministic specification consists of the regression function, of which the ingredients are the explanatory variable and the parameters. The stochastic specification ("stochastic" is another word for "random") consists of the probability distribution of the error terms, and the requirement that the error terms should be IID drawings from this distribution. Then, in order to simulate the dependent variable y_t in (1.01), we do as follows:

- Fix the sample size, n;
- Choose the parameters (here β_1 and β_2) of the deterministic specification;
- Obtain the n successive values X_t, $t = 1, \ldots, n$, of the explanatory variable. As explained above, these values may be real-world data or the output of another simulation;
- Evaluate the n successive values of the regression function $\beta_1 + \beta_2 X_t$, for $t = 1, \ldots, n$;
- Choose the probability distribution of the error terms, if necessary specifying parameters such as its mean and variance;
- Use a random-number generator to generate the n successive and mutually independent values u_t of the error terms;
- Form the n successive values y_t of the dependent variable by adding the error terms to the values of the regression function.

The n values y_t, $t = 1, \ldots, n$, thus generated are the output of the simulation; they are the **simulated values** of the dependent variable.

The chief interest of such a simulation is that, if the model we simulate is correctly specified and thus reflects the real-world generating process for the dependent variable, our simulation mimics the real world accurately, because it makes use of the same data-generating mechanism as that in operation in the real world.

A complete specification, then, is anything that leads unambiguously to a recipe like the one given above. We will define a **fully specified parametric model** as a model for which it is possible to simulate the dependent variable once the values of the parameters are known. A **partially specified parametric model** is one for which more information, over and above the parameter values, must be supplied before simulation is possible. Both sorts of models are frequently encountered in econometrics.

To conclude this discussion of simulations, let us return to the specifications (1.01) and (1.19). Both are obviously incomplete as they stand. In order to complete either one, it is necessary to specify the information set Ω_t and the distribution of u_t conditional on Ω_t. In particular, it is necessary to know whether the error terms u_s with $s \neq t$ belong to Ω_t. In (1.19), one aspect of the conditional distribution is given, namely, the conditional mean. Unfortunately, because (1.19) contains no explicit error term, it is easy to forget that it is there. Perhaps as a result, it is more common to write regression models in the form of (1.01) than in the form of (1.19). However, writing a model in the form of (1.01) does have the disadvantage that it obscures both the dependence of the model on the choice of an information set and the fact that the distribution of the error term must be specified conditional on that information set.

Linear and Nonlinear Regression Models

The simple linear regression model (1.01) is by no means the only reasonable model for the mean of y_t conditional on X_t. Consider, for example, the models

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t^2 + u_t \tag{1.20}$$

$$y_t = \gamma_1 + \gamma_2 \log X_t + u_t, \quad \text{and} \tag{1.21}$$

$$y_t = \delta_1 + \delta_2 \frac{1}{X_t} + u_t. \tag{1.22}$$

These are all models that might be plausible in some circumstances.[3] In equation (1.20), there is an extra parameter, β_3, which allows $\mathrm{E}(y_t \,|\, X_t)$ to vary quadratically with X_t whenever β_3 is nonzero. In effect, X_t and X_t^2 are being treated as separate explanatory variables. Thus (1.20) is the first example we have seen of a **multiple linear regression model**. It reduces to the simple linear regression model (1.01) when $\beta_3 = 0$.

In the models (1.21) and (1.22), on the other hand, there are no extra parameters. Instead, a nonlinear transformation of X_t is used in place of X_t itself. As a consequence, the relationship between X_t and $\mathrm{E}(y_t \,|\, X_t)$ in these two models is necessarily nonlinear. Nevertheless, (1.20), (1.21), and (1.22) are all said to be linear regression models, because, even though the mean of y_t may depend nonlinearly on X_t, it always depends linearly on the unknown parameters of the regression function. As we will see in Section 1.5, it is quite easy to estimate a linear regression model. In contrast, genuinely nonlinear models, in which the regression function depends nonlinearly on the parameters, are somewhat harder to estimate; see Chapter 6.

Because it is very easy to estimate linear regression models, a great deal of applied work in econometrics makes use of them. It may seem that the linearity assumption is very restrictive. However, as the examples (1.20), (1.21), and (1.22) illustrate, this assumption need not be unduly restrictive in practice, at least not if the econometrician is at all creative. If we are willing to transform the dependent variable as well as the independent ones,

[3] In this book, all logarithms are natural logarithms. Thus $a = \log x$ implies that $x = e^a$. Some authors use "ln" to denote natural logarithms and "log" to denote base 10 logarithms. Since econometricians should never have any use for base 10 logarithms, we avoid this aesthetically displeasing notation.

the linearity assumption can be made even less restrictive. As an example, consider the nonlinear regression model

$$y_t = e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3} + u_t, \tag{1.23}$$

in which there are two explanatory variables, X_{t2} and X_{t3}, and the regression function is multiplicative. If the notation seems odd, suppose that there is implicitly a third explanatory variable, X_{t1}, which is constant and always equal to e. Notice that the regression function in (1.23) can be evaluated only when X_{t2} and X_{t3} are positive for all t. It is a genuinely nonlinear regression function, since it is clearly linear neither in parameters nor in variables. For reasons that will shortly become apparent, a nonlinear model like (1.23) is very rarely estimated in practice.

A model like (1.23) is not as outlandish as may appear at first glance. It could arise, for instance, if we wanted to estimate a Cobb-Douglas production function. In that case, y_t would be output for observation t, and X_{t2} and X_{t3} would be inputs, say labor and capital. Since e^{β_1} is just a positive constant, it plays the role of the scale factor that is present in every Cobb-Douglas production function.

As (1.23) is written, everything enters multiplicatively except the error term. But it is easy to modify (1.23) so that the error term also enters multiplicatively. One way to do this is to write

$$y_t = e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3} + u_t \equiv \big(e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3}\big)(1 + v_t), \tag{1.24}$$

where the error factor $1 + v_t$ multiplies the regression function. If we now assume that the underlying errors v_t are IID, it follows that the additive errors u_t are proportional to the regression function. This may well be a more plausible specification than that in which the u_t are supposed to be IID, as was implicitly assumed in (1.23). To see this, notice first that the additive error u_t has the same units of measurement as y_t. If (1.23) is interpreted as a production function, then u_t is measured in units of output. However, the multiplicative error v_t is dimensionless. In other words, it is a pure number, like 0.02, which could be expressed as 2 percent. If the u_t are assumed to be IID, then we are assuming that the error in output is of the same order of magnitude regardless of the scale of production. If, on the other hand, the v_t are assumed to be IID, then the error is proportional to total output. This second assumption is almost always more reasonable than the first.

If the model (1.24) is a good one, the v_t should be quite small, usually less than about 0.05. For small values of the argument w, a standard approximation to the exponential function gives us that $e^w \cong 1 + w$. As a consequence, (1.24) will be very similar to the model

$$y_t = e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3} e^{v_t}, \tag{1.25}$$

whenever the error terms are reasonably small.

Now suppose we take logarithms of both sides of (1.25). The result is

$$\log y_t = \beta_1 + \beta_2 \log X_{t2} + \beta_3 \log X_{t3} + v_t, \qquad (1.26)$$

which is a **loglinear regression model**. This model is linear in the parameters and in the logarithms of all the variables, and so it is very much easier to estimate than the nonlinear model (1.23). Since (1.25) is at least as plausible as (1.23), it is not surprising that loglinear regression models, like (1.26), are estimated very frequently in practice, while multiplicative models with additive error terms, like (1.23), are very rarely estimated. Of course, it is important to remember that (1.26) is not a model for the mean of y_t conditional on X_{t2} and X_{t3}. Instead, it is a model for the mean of $\log y_t$ conditional on those variables. If it is really the conditional mean of y_t that we are interested in, we will not want to estimate a loglinear model like (1.26).

1.4 Matrix Algebra

It is impossible to study econometrics beyond the most elementary level without using matrix algebra. Most readers are probably already quite familiar with matrix algebra. This section reviews some basic results that will be used throughout the book. It also shows how regression models can be written very compactly using matrix notation. More advanced material will be discussed in later chapters, as it is needed.

An $n \times m$ **matrix** $\boldsymbol{A}$ is a rectangular array that consists of nm elements arranged in n rows and m columns. The name of the matrix is conventionally shown in boldface. A typical element of $\boldsymbol{A}$ might be denoted by either A_{ij} or a_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The first subscript always indicates the row, and the second always indicates the column. It is sometimes necessary to show the elements of a matrix explicitly, in which case they are arrayed in rows and columns and surrounded by large brackets, as in

$$\boldsymbol{B} = \begin{bmatrix} 2 & 3 & 6 \\ 4 & 5 & 8 \end{bmatrix}.$$

Here $\boldsymbol{B}$ is a 2×3 matrix.

If a matrix has only one column or only one row, it is called a **vector**. There are two types of vectors, **column vectors** and **row vectors**. Since column vectors are more common than row vectors, a vector that is not specified to be a row vector is normally treated as a column vector. If a column vector has n elements, it may be referred to as an n–vector. Boldface is used to denote vectors as well as matrices. It is conventional to use uppercase letters for matrices and lowercase letters for column vectors. However, it is sometimes necessary to ignore this convention.

If a matrix has the same number of columns and rows, it is said to be **square**. A square matrix $\boldsymbol{A}$ is **symmetric** if $A_{ij} = A_{ji}$ for all i and j. Symmetric matrices occur very frequently in econometrics. A square matrix is said to be **diagonal** if $A_{ij} = 0$ for all $i \neq j$; in this case, the only nonzero entries are those on what is called the **principal diagonal**. Sometimes a square matrix has all zeros above or below the principal diagonal. Such a matrix is said to be **triangular**. If the nonzero elements are all above the diagonal, it is said to be **upper-triangular**; if the nonzero elements are all below the diagonal, it is said to be **lower-triangular**. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 6 \\ 4 & 6 & 5 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 5 & 2 & 6 \end{bmatrix}.$$

In this case, $\boldsymbol{A}$ is symmetric, $\boldsymbol{B}$ is diagonal, and $\boldsymbol{C}$ is lower-triangular.

The **transpose** of a matrix is obtained by interchanging its row and column subscripts. Thus the ij^{th} element of $\boldsymbol{A}$ becomes the ji^{th} element of its transpose, which is denoted $\boldsymbol{A}^\top$. Note that many authors use $\boldsymbol{A}'$ rather than $\boldsymbol{A}^\top$ to denote the transpose of $\boldsymbol{A}$. The transpose of a symmetric matrix is equal to the matrix itself. The transpose of a column vector is a row vector, and vice versa. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 5 & 7 \\ 3 & 8 & 4 \end{bmatrix} \quad \boldsymbol{A}^\top = \begin{bmatrix} 2 & 3 \\ 5 & 8 \\ 7 & 4 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \quad \boldsymbol{b}^\top = [\,2 \quad 4 \quad 6\,].$$

Note that a matrix $\boldsymbol{A}$ is symmetric if and only if $\boldsymbol{A} = \boldsymbol{A}^\top$.

Arithmetic Operations on Matrices

Addition and **subtraction** of matrices works exactly the way it does for scalars, with the proviso that matrices can be added or subtracted only if they are **conformable**. In the case of addition and subtraction, this just means that they must have the same dimensions, that is, the same number of rows and the same number of columns. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable, then a typical element of $\boldsymbol{A} + \boldsymbol{B}$ is simply $A_{ij} + B_{ij}$, and a typical element of $\boldsymbol{A} - \boldsymbol{B}$ is $A_{ij} - B_{ij}$.

Matrix multiplication actually involves both additions and multiplications. It is based on what is called the **inner product**, or **scalar product**, of two vectors. Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are n–vectors. Then their inner product is

$$\boldsymbol{a}^\top\boldsymbol{b} = \boldsymbol{b}^\top\boldsymbol{a} = \sum_{i=1}^{n} a_i b_i.$$

As the name suggests, this is just a scalar.

When two matrices are multiplied together, the $ij^{\rm th}$ element of the result is equal to the inner product of the $i^{\rm th}$ row of the first matrix with the $j^{\rm th}$ column of the second matrix. Thus, if $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$,

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}. \tag{1.27}$$

For (1.27) to make sense, we must assume that $\boldsymbol{A}$ has m columns and that $\boldsymbol{B}$ has m rows. In general, if two matrices are to be conformable for multiplication, the first matrix must have as many columns as the second has rows. Further, as is clear from (1.27), the result has as many rows as the first matrix and as many columns as the second. One way to make this explicit is to write something like

$$\underset{n\times m}{\boldsymbol{A}}\ \underset{m\times l}{\boldsymbol{B}} = \underset{n\times l}{\boldsymbol{C}}.$$

One rarely sees this type of notation in a book or journal article. However, it is often useful to employ it when doing calculations, in order to verify that the matrices being multiplied are indeed conformable and to derive the dimensions of their product.

The rules for multiplying matrices and vectors together are the same as the rules for multiplying matrices with each other; vectors are simply treated as matrices that have only one column or only one row. For instance, if we multiply an n−vector $\boldsymbol{a}$ by the transpose of an n−vector $\boldsymbol{b}$, we obtain what is called the **outer product** of the two vectors. The result, written as $\boldsymbol{a}\boldsymbol{b}^\top$, is an $n \times n$ matrix with typical element $a_i b_j$.

Matrix multiplication is, in general, not commutative. The fact that it is possible to **premultiply** $\boldsymbol{B}$ by $\boldsymbol{A}$ does not imply that it is possible to **postmultiply** $\boldsymbol{B}$ by $\boldsymbol{A}$. In fact, it is easy to see that both operations are possible if and only if one of the matrix products is square, in which case the other matrix product will be square also, although generally with different dimensions. Even when both operations are possible, $\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A}$ except in special cases.

A special matrix that econometricians frequently make use of is $\mathbf{I}$, which denotes the **identity matrix**. It is a diagonal matrix with every diagonal element equal to 1. A subscript is sometimes used to indicate the number of rows and columns. Thus

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The identity matrix is so called because when it is either premultiplied or postmultiplied by any matrix, it leaves the latter unchanged. Thus, for any matrix $\boldsymbol{A}$, $\boldsymbol{A}\mathbf{I} = \mathbf{I}\boldsymbol{A} = \boldsymbol{A}$, provided, of course, that the matrices are conformable for multiplication. It is easy to see why the identity matrix has this property. Recall that the only nonzero elements of $\mathbf{I}$ are equal to 1 and are on the principal diagonal. This fact can be expressed simply with the help of the symbol known as the **Kronecker delta**, written as δ_{ij}. The definition is

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{1.28}$$

The ij^{th} element of $\mathbf{I}$ is just δ_{ij}. By (1.27), the ij^{th} element of $\boldsymbol{A}\mathbf{I}$ is

$$\sum_{k=1}^{m} A_{ik}\mathrm{I}_{kj} = \sum_{k=1}^{m} A_{ik}\delta_{kj} = A_{ij},$$

since all the terms in the sum over k vanish except that for which $k = j$.

A special vector that we frequently use in this book is $\boldsymbol{\iota}$. It denotes a column vector every element of which is 1. This special vector comes in handy whenever one wishes to sum the elements of another vector, because, for any n–vector $\boldsymbol{b}$,

$$\boldsymbol{\iota}^{\top}\boldsymbol{b} = \sum_{i=1}^{n} b_i. \tag{1.29}$$

Matrix multiplication and matrix addition interact in an intuitive way. It is easy to check from the definitions of the respective operations that the **distributive** properties hold. That is, assuming that the dimensions of the matrices are conformable for the various operations,

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{AB}+\boldsymbol{AC}, \quad \text{and} \\ (\boldsymbol{B}+\boldsymbol{C})\boldsymbol{A} &= \boldsymbol{BA}+\boldsymbol{CA}. \end{aligned}$$

In addition, both operations are **associative**, which means that

$$\begin{aligned} (\boldsymbol{A}+\boldsymbol{B})+\boldsymbol{C} &= \boldsymbol{A}+(\boldsymbol{B}+\boldsymbol{C}), \quad \text{and} \\ (\boldsymbol{AB})\boldsymbol{C} &= \boldsymbol{A}(\boldsymbol{BC}). \end{aligned}$$

The transpose of the product of two matrices is the product of the transposes of the matrices with the order reversed. Thus

$$(\boldsymbol{AB})^{\top} = \boldsymbol{B}^{\top}\boldsymbol{A}^{\top}. \tag{1.30}$$

The reversal of the order is necessary for the transposed matrices to be conformable for multiplication. The result (1.30) can be proved immediately by writing out the typical entries of both sides and checking that

$$(\boldsymbol{AB})^{\top}_{ij} = (\boldsymbol{AB})_{ji} = \sum_{k=1}^{m} A_{jk}B_{ki} = \sum_{k=1}^{m} (\boldsymbol{B}^{\top})_{ik}(\boldsymbol{A}^{\top})_{kj} = (\boldsymbol{B}^{\top}\boldsymbol{A}^{\top})_{ij},$$

where m is the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$. It is always possible to multiply a matrix by its own transpose: If $\boldsymbol{A}$ is $n \times m$, then

$\boldsymbol{A}^{\top}$ is $m \times n$, $\boldsymbol{A}^{\top}\boldsymbol{A}$ is $m \times m$, and $\boldsymbol{A}\boldsymbol{A}^{\top}$ is $n \times n$. It follows directly from (1.30) that both of these matrix products are symmetric:

$$\boldsymbol{A}^{\top}\boldsymbol{A} = (\boldsymbol{A}^{\top}\boldsymbol{A})^{\top} \quad \text{and} \quad \boldsymbol{A}\boldsymbol{A}^{\top} = (\boldsymbol{A}\boldsymbol{A}^{\top})^{\top}.$$

It is frequently necessary to multiply a matrix, say $\boldsymbol{B}$, by a scalar, say α. **Multiplication by a scalar** works exactly the way one would expect: Every element of $\boldsymbol{B}$ is multiplied by α. Since multiplication by a scalar is commutative, we can write this either as $\alpha\boldsymbol{B}$ or as $\boldsymbol{B}\alpha$, but $\alpha\boldsymbol{B}$ is the more common notation.

Occasionally, it is necessary to multiply two matrices together element by element. The result is called the **direct product** of the two matrices. The direct product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted $\boldsymbol{A} {*} \boldsymbol{B}$, and a typical element of it is equal to $A_{ij}B_{ij}$.

A square matrix may or may not be **invertible**. If $\boldsymbol{A}$ is invertible, then it has an **inverse matrix** $\boldsymbol{A}^{-1}$ with the property that

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbf{I}.$$

If $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$. If $\boldsymbol{A}$ is triangular, then so is $\boldsymbol{A}^{-1}$. Except in certain special cases, it is not easy to calculate the inverse of a matrix by hand. One such special case is that of a diagonal matrix, say $\boldsymbol{D}$, with typical diagonal element D_{ii}. It is easy to verify that $\boldsymbol{D}^{-1}$ is also a diagonal matrix, with typical diagonal element D_{ii}^{-1}.

If an $n \times n$ square matrix $\boldsymbol{A}$ is invertible, then its **rank** is n. Such a matrix is said to have **full rank**. If a square matrix does not have full rank, and therefore is not invertible, it is said to be **singular**. If a square matrix is singular, its rank must be less than its dimension. If, by omitting j rows and j columns of $\boldsymbol{A}$, we can obtain a matrix $\boldsymbol{A}'$ that is invertible, and if j is the smallest number for which this is true, the rank of $\boldsymbol{A}$ is $n - j$. More generally, for matrices that are not necessarily square, the rank is the largest number m for which an $m \times m$ nonsingular matrix can be constructed by omitting some rows and some columns from the original matrix. The rank of a matrix is closely related to the geometry of vector spaces, which will be discussed in the next chapter.

Regression Models and Matrix Notation

The simple linear regression model (1.01) can easily be written in matrix notation. If we stack the model for all the observations, we obtain

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 X_1 + u_1 \\ y_2 &= \beta_1 + \beta_2 X_2 + u_2 \\ \vdots \quad & \quad \vdots \qquad \vdots \qquad \vdots \\ y_n &= \beta_1 + \beta_2 X_n + u_n\,. \end{aligned} \tag{1.31}$$

Let $\boldsymbol{y}$ denote an n–vector with typical element y_t, $\boldsymbol{u}$ an n–vector with typical element u_t, $\boldsymbol{X}$ an $n \times 2$ matrix that consists of a column of 1s and a column with typical element X_t, and $\boldsymbol{\beta}$ a 2–vector with typical element β_i, $i = 1, 2$. Thus we have

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Equations (1.31) can now be rewritten as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}. \tag{1.32}$$

It is easy to verify from the rules of matrix multiplication that a typical row of (1.32) is a typical row of (1.31). When we postmultiply the matrix $\boldsymbol{X}$ by the vector $\boldsymbol{\beta}$, we obtain a vector $\boldsymbol{X\beta}$ with typical element $\beta_1 + \beta_2 X_t$.

When a regression model is written in the form (1.32), the separate columns of the matrix $\boldsymbol{X}$ are called **regressors**, and the column vector $\boldsymbol{y}$ is called the **regressand**. In (1.31), there are just two regressors, corresponding to the constant and one explanatory variable. One advantage of writing the regression model in the form (1.32) is that we are not restricted to just one or two regressors. Suppose that we have k regressors, one of which may or may not correspond to a constant, and the others to a number of explanatory variables. Then the matrix $\boldsymbol{X}$ becomes

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}, \tag{1.33}$$

where X_{ti} denotes the t^{th} observation on the i^{th} regressor, and the vector $\boldsymbol{\beta}$ now has k elements, β_1 through β_k. Equation (1.32) remains perfectly valid when $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are redefined in this way. A typical row of this equation is

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t = \sum_{i=1}^{k} \beta_i X_{ti} + u_t, \tag{1.34}$$

where we have used $\boldsymbol{X}_t$ to denote the t^{th} row of $\boldsymbol{X}$.

In (1.32), we used the rules of matrix multiplication to write the regression function, for the entire sample, in a very simple form. These rules make it possible to find equally convenient expressions for other aspects of regression models. The key fact is that every element of the product of two matrices is a summation. Thus it is often very convenient to use matrix algebra when dealing with summations. Consider, for example, the matrix of sums of squares and cross-products of the $\boldsymbol{X}$ matrix. This is a $k \times k$ symmetric matrix, of which a typical element is either

$$\sum_{t=1}^{n} X_{ti}^2 \quad \text{or} \quad \sum_{t=1}^{n} X_{ti} X_{tj},$$

the former being a typical diagonal element and the latter a typical off-diagonal one. This entire matrix can be written very compactly as $\boldsymbol{X}^\top\boldsymbol{X}$. Similarly, the vector with typical element

$$\sum_{t=1}^{n} X_{ti}\, y_t$$

can be written as $\boldsymbol{X}^\top\boldsymbol{y}$. As we will see in the next section, the least squares estimates of $\boldsymbol{\beta}$ depend only on the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ and the vector $\boldsymbol{X}^\top\boldsymbol{y}$.

Partitioned Matrices

There are many ways of writing an $n \times k$ matrix $\boldsymbol{X}$ that are intermediate between the straightforward notation $\boldsymbol{X}$ and the full element-by-element decomposition of $\boldsymbol{X}$ given in (1.33). We might wish to separate the columns while grouping the rows, as

$$\underset{n\times k}{\boldsymbol{X}} = \big[\, \underset{n\times 1}{\boldsymbol{x}_1} \quad \underset{n\times 1}{\boldsymbol{x}_2} \quad \underset{\dots}{\cdots} \quad \underset{n\times 1}{\boldsymbol{x}_k} \,\big],$$

or we might wish to separate the rows but not the columns, as

$$\boldsymbol{X} = \underset{n\times k}{\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}} \begin{array}{l} 1\times k \\ 1\times k \\ \\ 1\times k \end{array} .$$

To save space, we can also write this as $\boldsymbol{X} = \big[\boldsymbol{X}_1 \,\vdots\, \boldsymbol{X}_2 \,\vdots\, \dots \,\vdots\, \boldsymbol{X}_n\big]$. There is no restriction on how a matrix can be partitioned, so long as all the **submatrices** or **blocks** fit together correctly. Thus we might have

$$\boldsymbol{X} = \begin{array}{c} \begin{array}{cc} k_1 & k_2 \end{array} \\ \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix} \end{array} \begin{array}{l} \\ n_1 \\ n_2 \end{array}$$

with the submatrix $\boldsymbol{X}_{11}$ of dimensions $n_1 \times k_1$, $\boldsymbol{X}_{12}$ of dimensions $n_1 \times k_2$, $\boldsymbol{X}_{21}$ of dimensions $n_2 \times k_1$, and $\boldsymbol{X}_{22}$ of dimensions $n_2 \times k_2$, with $n_1 + n_2 = n$ and $k_1 + k_2 = k$. Thus $\boldsymbol{X}_{11}$ and $\boldsymbol{X}_{12}$ have the same number of rows, and also $\boldsymbol{X}_{21}$ and $\boldsymbol{X}_{22}$, as required for the submatrices to fit together horizontally. Similarly, $\boldsymbol{X}_{11}$ and $\boldsymbol{X}_{21}$ have the same number of columns, and also $\boldsymbol{X}_{12}$ and $\boldsymbol{X}_{22}$, as required for the submatrices to fit together vertically as well.

If two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimensions are partitioned in exactly the same way, they can be added or subtracted block by block. A simple example is

$$\boldsymbol{A}+\boldsymbol{B} = [\,\boldsymbol{A}_1 \quad \boldsymbol{A}_2\,] + [\,\boldsymbol{B}_1 \quad \boldsymbol{B}_2\,] = [\,\boldsymbol{A}_1+\boldsymbol{B}_1 \quad \boldsymbol{A}_2+\boldsymbol{B}_2\,],$$

where $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ have the same dimensions, as do $\boldsymbol{A}_2$ and $\boldsymbol{B}_2$.

More interestingly, as we now explain, matrix multiplication can sometimes be performed block by block on partitioned matrices. If the product $\boldsymbol{AB}$ exists, then $\boldsymbol{A}$ has as many columns as $\boldsymbol{B}$ has rows. Now suppose that the columns of $\boldsymbol{A}$ are partitioned in the same way as the rows of $\boldsymbol{B}$. Then

$$\boldsymbol{AB} = [\,\boldsymbol{A}_1 \quad \boldsymbol{A}_2 \quad \cdots \quad \boldsymbol{A}_p\,] \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_p \end{bmatrix}.$$

Here each $\boldsymbol{A}_i$, $i = 1, \ldots, p$, has as many columns as the corresponding $\boldsymbol{B}_i$ has rows. The product can be computed following the usual rules for matrix multiplication just as though the blocks were scalars, yielding the result

$$\boldsymbol{AB} = \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{B}_i. \tag{1.35}$$

To see this, it is enough to compute the typical element of each side of equation (1.35) directly and observe that they are the same. Matrix multiplication can also be performed block by block on matrices that are partitioned both horizontally and vertically, provided all the submatrices are conformable; see Exercise 1.17.

These results on multiplying partitioned matrices lead to a useful corollary. Suppose that we are interested only in the first m rows of a product $\boldsymbol{AB}$, where $\boldsymbol{A}$ has more than m rows. Then we can partition the rows of $\boldsymbol{A}$ into two blocks, the first with m rows, the second with all the rest. We need not partition $\boldsymbol{B}$ at all. Then

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_1\boldsymbol{B} \\ \boldsymbol{A}_2\boldsymbol{B} \end{bmatrix}. \tag{1.36}$$

This works because $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ both have the full number of columns of $\boldsymbol{A}$, which must be the same as the number of rows of $\boldsymbol{B}$, since $\boldsymbol{AB}$ exists. It is clear from the rightmost expression in (1.36) that the first m rows of $\boldsymbol{AB}$ are given by $\boldsymbol{A}_1\boldsymbol{B}$. In order to obtain any subset of the rows of a matrix product of arbitrarily many factors, the rule is that we take the submatrix of the leftmost factor that contains just the rows we want, and then multiply it by all the other factors unchanged. Similarly, if we want to select a subset of columns of a matrix product, we can just select them from the rightmost factor, leaving all the factors to the left unchanged.

1.5 Method of Moments Estimation

Almost all econometric models contain unknown parameters. For most of the uses to which such models can be put, it is necessary to have **estimates** of these parameters. To compute parameter estimates, we need both a model containing the parameters and a sample made up of observed data. If the model is correctly specified, it describes the real-world mechanism which generated the data in our sample.

It is common in statistics to speak of the "population" from which a sample is drawn. Recall the use of the term "population mean" as a synonym for the mathematical term "expectation"; see Section 1.2. The expression is a holdover from the time when statistics was biostatistics, and the object of study was the human population, usually that of a specific town or country, from which random samples were drawn by statisticians for study. The average weight of all members of the population, for instance, would then be estimated by the mean of the weights of the individuals in the sample, that is, by the **sample mean** of individuals' weights. The sample mean was thus an estimate of the **population mean**. The underlying idea is just that the sample *represents* the population from which it has been drawn.

In econometrics, the use of the term population is simply a metaphor. A better concept is that of a **data-generating process**, or **DGP**. By this term, we mean whatever mechanism is at work in the real world of economic activity giving rise to the numbers in our samples, that is, precisely the mechanism that our econometric model is supposed to describe. A data-generating process is thus the analog in econometrics of a population in biostatistics. Samples may be drawn from a DGP just as they may be drawn from a population. In both cases, the samples are assumed to be representative of the DGP or population from which they are drawn.

A very natural way to estimate parameters is to replace population means by sample means. This technique is called the **method of moments**, and it is one of the most widely-used estimation methods in statistics. As the name implies, it can be used with moments other than the mean. In general, the method of moments, sometimes called **MM** for short, estimates population moments by the corresponding sample moments. In order to apply this method to regression models, we must use the facts that population moments are expectations, and that regression models are specified in terms of the conditional expectations of the error terms.

Estimating the Simple Linear Regression Model

Let us now see how the principle of replacing population means by sample means works for the simple linear regression model (1.01). The error term for observation t is

$$u_t = y_t - \beta_1 - \beta_2 X_t,$$

and, according to our model, the expectation of this error term is zero. Since we have n error terms for a sample of size n, we can consider the sample mean of the error terms:

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta_2 X_t). \tag{1.37}$$

We would like to set this sample mean equal to zero.

Suppose to begin with that $\beta_2 = 0$. This reduces the number of parameters in the model to just one. In that case, there is just one value of β_1 which will allow (1.37) to be zero. The equation defining this value is

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1) = 0. \tag{1.38}$$

Since β_1 is common to all the observations and thus does not depend on the index t, (1.38) can be written as

$$\frac{1}{n}\sum_{t=1}^{n} y_t - \beta_1 = 0.$$

We can easily solve this equation to obtain an estimate $\hat\beta_1$. This estimate is just the mean of the observed values of the dependent variable,

$$\hat\beta_1 = \frac{1}{n}\sum_{t=1}^{n} y_t. \tag{1.39}$$

Thus, if we wish to estimate the population mean of the y_t, which is what β_1 is in our model when $\beta_2 = 0$, the method of moments tells us to use the sample mean as our estimate.

It is not obvious at first glance how to use the method of moments if we put the second parameter β_2 back into the model. Equation (1.38) would become

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta_2 X_t) = 0, \tag{1.40}$$

but this is just one equation, and there are two unknowns. In order to obtain another equation, we can use the fact that our model specifies that the mean of u_t is 0 *conditional* on the explanatory variable X_t. Actually, it may well specify that the mean of u_t is 0 conditional on many other things as well, depending on our choice of the information set Ω_t, but we will ignore this for now. The conditional mean assumption implies that not only is $\mathrm{E}(u_t) = 0$, but that $\mathrm{E}(X_t u_t) = 0$ as well, since, by (1.16) and (1.17),

$$\mathrm{E}(X_t u_t) = \mathrm{E}\big(\mathrm{E}(X_t u_t \,|\, X_t)\big) = \mathrm{E}\big(X_t \mathrm{E}(u_t \,|\, X_t)\big) = 0. \tag{1.41}$$

Thus we can supplement (1.40) by the following equation, which replaces the population mean in (1.41) by the corresponding sample mean,

$$\frac{1}{n}\sum_{t=1}^{n} X_t(y_t - \beta_1 - \beta_2 X_t) = 0. \tag{1.42}$$

The equations (1.40) and (1.42) are two linear equations in two unknowns, β_1 and β_2. Except in rare conditions, which can easily be ruled out, they will have a unique solution that is not difficult to calculate. Solving these equations yields the MM estimates.

We could just solve (1.40) and (1.42) directly, but it is far more illuminating to rewrite them in matrix form. Since β_1 and β_2 do not depend on t, these two equations can be written as

$$\begin{aligned} \beta_1 + \Big(\frac{1}{n}\sum_{t=1}^{n} X_t\Big)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} y_t \\ \Big(\frac{1}{n}\sum_{t=1}^{n} X_t\Big)\beta_1 + \Big(\frac{1}{n}\sum_{t=1}^{n} X_t^2\Big)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} X_t y_t. \end{aligned}$$

Multiplying both equations by n and using the rules of matrix multiplication that were discussed in the last section, we can also write them as

$$\begin{bmatrix} n & \sum_{t=1}^{n} X_t \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} X_t y_t \end{bmatrix}. \tag{1.43}$$

Equations (1.43) can be rewritten much more compactly. As we saw in the last section, the model (1.01) is simply a special case of the **multiple linear regression model**

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1.44}$$

where the n–vector $\boldsymbol{y}$ has typical element y_t, the k–vector $\boldsymbol{\beta}$ has typical element β_i, and, in general, the matrix $\boldsymbol{X}$ is $n \times k$. In this case, $\boldsymbol{X}$ is $n \times 2$; it can be written as $\boldsymbol{X} = [\boldsymbol{\iota} \;\; \boldsymbol{x}]$, where $\boldsymbol{\iota}$ denotes a column of 1s, and $\boldsymbol{x}$ denotes a column with typical element X_t. Thus, recalling (1.29), we see that

$$\boldsymbol{X}^\top\boldsymbol{y} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} X_t y_t \end{bmatrix}$$

and

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} n & \sum_{t=1}^{n} X_t \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \end{bmatrix}.$$

These are the principal quantities that appear in the equations (1.43). Thus it is clear that we can rewrite those equations as

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{y}. \tag{1.45}$$

To find the estimator $\hat{\boldsymbol{\beta}}$ that solves (1.45), we simply multiply it by the inverse of the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, assuming that this inverse exists. This yields the famous formula

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{1.46}$$

The estimator $\hat{\boldsymbol{\beta}}$ given by this formula is generally called the **ordinary least squares**, or **OLS**, estimator for the linear regression model.[4] Why it is called this, rather than the MM estimator, will be explained shortly.

Estimating the Multiple Linear Regression Model

The formula (1.46) gives us the OLS, and MM, estimator for the simple linear regression model (1.01), but in fact it does far more than that. As we now show, it also gives us the MM estimator for the multiple linear regression model (1.44). Since each of the explanatory variables is required to be in the information set Ω_t, we have, for $i = 1, \ldots, k$,

$$\mathrm{E}(X_{ti}\, u_t) = 0;$$

which, in the corresponding sample mean form, yields

$$\frac{1}{n}\sum_{t=1}^{n} X_{ti}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = 0. \tag{1.47}$$

(Recall from (1.34) that $\boldsymbol{X}_t$ denotes the t^{th} row of $\boldsymbol{X}$.) As i varies from 1 to k, equation (1.47) yields k equations for the k unknown components of $\boldsymbol{\beta}$. In most cases, there will be a constant, which we may take to be the first regressor. If so, $X_{t1} = 1$, and the first of these equations simply says that the sample mean of the error terms is 0.

In matrix form, after multiplying them by n, the k equations of (1.47) can be written as

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{1.48}$$

The notation $\mathbf{0}$ is used to signify a **zero vector**, here a k--vector, each element of which is zero. Equations (1.48) are clearly equivalent to equations (1.45). Thus solving them yields the estimator (1.46), which applies no matter what the number of regressors.

It is easy to see that the OLS estimator (1.46) depends on $\boldsymbol{y}$ and $\boldsymbol{X}$ exclusively through a number of scalar products. Each column $\boldsymbol{x}_i$ of the matrix $\boldsymbol{X}$ corresponds to one of the regressors, as does each row $\boldsymbol{x}_i^\top$ of the transposed

[4] Econometricians generally make a distinction between an **estimate**, which is simply a number used to estimate some parameter, normally based on a particular data set, and an **estimator**, which is a rule, such as (1.46), for obtaining estimates from any set of data.

matrix $\boldsymbol{X}^\top$. Thus we can write $\boldsymbol{X}^\top\boldsymbol{y}$ as

$$\boldsymbol{X}^\top\boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_1{}^\top \\ \boldsymbol{x}_2{}^\top \\ \vdots \\ \boldsymbol{x}_k{}^\top \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_1{}^\top\boldsymbol{y} \\ \boldsymbol{x}_2{}^\top\boldsymbol{y} \\ \vdots \\ \boldsymbol{x}_k{}^\top\boldsymbol{y} \end{bmatrix}.$$

The elements of the rightmost expression here are just the scalar products of the regressors $\boldsymbol{x}_i$ with the regressand $\boldsymbol{y}$. Similarly, we can write $\boldsymbol{X}^\top\boldsymbol{X}$ as

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1{}^\top \\ \boldsymbol{x}_2{}^\top \\ \vdots \\ \boldsymbol{x}_k{}^\top \end{bmatrix} [\,\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k\,] = \begin{bmatrix} \boldsymbol{x}_1{}^\top\boldsymbol{x}_1 & \boldsymbol{x}_1{}^\top\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1{}^\top\boldsymbol{x}_k \\ \boldsymbol{x}_2{}^\top\boldsymbol{x}_1 & \boldsymbol{x}_2{}^\top\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2{}^\top\boldsymbol{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_k{}^\top\boldsymbol{x}_1 & \boldsymbol{x}_k{}^\top\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k{}^\top\boldsymbol{x}_k \end{bmatrix}.$$

Once more, all the elements of the rightmost expression are scalar products of pairs of regressors. Since $\boldsymbol{X}^\top\boldsymbol{X}$ can be expressed exclusively in terms of scalar products of the variables of the regression, the same is true of its inverse, the elements of which will be in general complicated functions of those scalar products. Thus $\hat{\boldsymbol{\beta}}$ is a function solely of scalar products of pairs of variables.

Least Squares Estimation

We have derived the estimator (1.46) by using the method of moments. Deriving it in this way has at least two major advantages. Firstly, the method of moments is a very general and very powerful principle of estimation, one that we will encounter again and again throughout this book. Secondly, by using the method of moments, we were able to obtain (1.46) without making any use of calculus. However, as we have already remarked, (1.46) is generally referred to as the OLS estimator, not the MM estimator. It is interesting to see why this is so.

For the multiple linear regression model (1.44), the expression $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ is equal to the error term for the t^{th} observation, but only if the correct value of the parameter vector $\boldsymbol{\beta}$ is used. If the same expression is thought of as a function of $\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ allowed to vary arbitrarily, then it is called a **residual**, more specifically, the residual associated with the t^{th} observation. Similarly, the n–vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ is called the vector of residuals. The sum of the squares of the components of the vector of residuals is called the **sum of squared residuals**, or **SSR**. Since this sum is a scalar, the sum of squared residuals is a scalar-valued function of the k–vector $\boldsymbol{\beta}$:

$$\text{SSR}(\boldsymbol{\beta}) = \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2. \tag{1.49}$$

The notation here emphasizes the fact that this function can be computed for arbitrary values of the argument $\boldsymbol{\beta}$ purely in terms of the observed data $\boldsymbol{y}$ and $\boldsymbol{X}$.

The idea of **least squares** estimation is to minimize the sum of squared residuals associated with a regression model. At this point, it may not be at all clear why we would wish to do such a thing. However, it can be shown that the parameter vector $\hat{\boldsymbol{\beta}}$ which minimizes (1.49) is the same as the MM estimator (1.46). This being so, we will regularly use the traditional terminology associated with linear regressions, based on least squares. Thus, the parameter estimates which are the components of the vector $\hat{\boldsymbol{\beta}}$ that minimizes the SSR (1.49) are called the **least squares estimates**, and the corresponding vector of residuals is called the vector of **least squares residuals**. When least squares is used to estimate a linear regression model like (1.01), it is called **ordinary least squares**, or **OLS**, to distinguish it from other varieties of least squares that we will encounter later, such as nonlinear least squares (Chapter 6) and generalized least squares (Chapter 7).

Consider briefly the simplest case of (1.01), in which $\beta_2 = 0$ and the model contains only a constant term. Expression (1.49) becomes

$$\mathrm{SSR}(\beta_1) = \sum_{t=1}^{n}(y_t - \beta_1)^2 = \sum_{t=1}^{n} y_t^2 + n\beta_1^2 - 2\beta_1\sum_{t=1}^{n} y_t. \tag{1.50}$$

Differentiating the rightmost expression in (1.50) with respect to β_1 and setting the derivative equal to zero gives the following first-order condition for a minimum:

$$\frac{\partial \mathrm{SSR}}{\partial \beta_1} = 2\beta_1 n - 2\sum_{t=1}^{n} y_t = 0. \tag{1.51}$$

For this simple model, the matrix $\boldsymbol{X}$ consists solely of the constant vector, $\boldsymbol{\iota}$. Therefore, by (1.29), $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{\iota}^\top\boldsymbol{\iota} = n$, and $\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{\iota}^\top\boldsymbol{y} = \sum_{t=1}^{n} y_t$. Thus, if the first-order condition (1.51) is multiplied by one-half, it can be rewritten as $\boldsymbol{\iota}^\top\boldsymbol{\iota}\beta_1 = \boldsymbol{\iota}^\top\boldsymbol{y}$, which is clearly just a special case of (1.45). Solving (1.51) for β_1 yields the sample mean of the y_t,

$$\hat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t = (\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top\boldsymbol{y}. \tag{1.52}$$

We already saw, in (1.39), that this is the MM estimator for the model with $\beta_2 = 0$. The rightmost expression in (1.52) makes it clear that the sample mean is just a special case of the famous formula (1.46).

Not surprisingly, the OLS and MM estimators are also equivalent in the multiple linear regression model. For this model,

$$\mathrm{SSR}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{1.53}$$

If this inner product is written out in terms of the scalar components of $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$, it is easy enough to show that the first-order conditions for minimizing the SSR (1.53) can be written as (1.45); see Exercise 1.20. Thus we conclude that (1.46) provides a general formula for the OLS estimator $\hat{\boldsymbol{\beta}}$ in the multiple linear regression model.

Final Remarks

We have seen that it is perfectly easy to obtain an algebraic expression, (1.46), for the OLS estimator $\hat{\boldsymbol{\beta}}$. With modern computers and appropriate software, it is also easy to obtain OLS estimates numerically, even for regressions with millions of observations and dozens of explanatory variables; the time-honored term for doing so is "running a regression". What is not so easy, and will occupy us for most of the next four chapters, is to understand the properties of these estimates.

We will be concerned with two types of properties. The first type, **numerical properties**, arise as a consequence of the way that OLS estimates are obtained. These properties hold for every set of OLS estimates, no matter how the data were generated. That they hold for any data set can easily be verified by direct calculation. The numerical properties of OLS will be discussed in Chapter 2. The second type, **statistical properties**, depend on the way in which the data were generated. They can be verified theoretically, under certain assumptions, and they can be illustrated by simulation, but we can never prove that they are true for any given data set. The statistical properties of OLS will be discussed in detail in Chapters 3, 4, and 5.

Readers who seek a deeper treatment of the topics dealt with in the first two sections may wish to consult Gallant (1997) or Mittelhammer (1996).

1.6 Notes on the Exercises

Each chapter of this book is followed by a set of exercises. These exercises are of various sorts, and they have various intended functions. Some are, quite simply, just for practice. Some serve chiefly to extend the material presented in the chapter. In many cases, the new material in such exercises recurs later in the book, and it is hoped that readers who have worked through them will follow later discussions more easily. A case in point concerns the **bootstrap**. Some of the exercises in this chapter and the next two are designed to familiarize readers with the tools that are used to implement the bootstrap, so that, when it is introduced formally in Chapter 4, the bootstrap will appear as a natural development. Other exercises have a tidying-up function. Details left out of the discussions in the main text are taken up, and conscientious readers can check that unproved claims made in the text are in fact justified.

Many of the exercises require the reader to make use of a computer, sometimes to compute estimates and test statistics using real or simulated data, and sometimes for the purpose of doing simulations. There are a great many computer packages that are capable of doing the things we ask for in the exercises, and it seems unnecessary to make any specific recommendations as to what software would be best. Besides, we expect that many readers will already have developed their own personal preferences for software packages, and we know better than to try to upset such preferences.

Some exercises require, not only a computer, but also actual economic data. It cannot be stressed enough that econometrics is an empirical discipline, and that the analysis of economic data is its *raison d'être*. All of the data needed for the exercises are available from the World Wide Web site for this book. The address is

http://www.econ.queensu.ca/ETM/

This web site will ultimately contain corrections and updates to the book as well as the data needed for the exercises.

1.7 Exercises

1.1 Consider a sample of n observations, $y_1, y_2, \ldots, y_n$, on some random variable Y. The **empirical distribution function**, or **EDF**, of this sample is a discrete distribution with n possible points. These points are just the n observed points, $y_1, y_2, \ldots, y_n$. Each point is assigned the same probability, which is just ${}^1\!/\!_n$, in order to ensure that all the probabilities sum to 1.

Compute the expectation of the discrete distribution characterized by the EDF, and show that it is equal to the **sample mean**, that is, the unweighted average of the n sample points, $y_1, y_2, \ldots, y_n$.

1.2 A random variable computed as the ratio of two independent standard normal variables follows what is called the **Cauchy distribution**. It can be shown that the density of this distribution is

$$f(x) = \frac{1}{\pi(1+x^2)}\,.$$

Show that the Cauchy distribution has no first moment, which means that its expectation does not exist.

Use your favorite random number generator to generate samples of 10, 100, 1,000, and 10,000 drawings from the Cauchy distribution, and as many intermediate values of n as you have patience or computer time for. For each sample, compute the sample mean. Do these sample means seem to converge to zero as the sample size increases? Repeat the exercise with drawings from the standard normal density. Do these sample means tend to converge to zero as the sample size increases?

1.3 Consider two events A and B such that $A \subset B$. Compute $\Pr(A \,|\, B)$ in terms of $\Pr(A)$ and $\Pr(B)$. Interpret the result.

1.4 Prove **Bayes' Theorem**. This famous theorem states that, for any two events A and B with nonzero probabilities,

$$\Pr(A \,|\, B) = \frac{\Pr(B \,|\, A)\Pr(A)}{\Pr(B)}\,.$$

Another form of the theorem deals with two continuous random variables X_1 and X_2, which have a joint density $f(x_1, x_2)$. Show that, for any values x_1 and x_2 that are permissible for X_1 and X_2, respectively,

$$f(x_1 \,|\, x_2) = \frac{f(x_2 \,|\, x_1)\, f(x_1)}{f(x_2)}\,.$$

1.5 Suppose that X and Y are two binary random variables. Their joint distribution is given in the following table.

	$Y = 0$	$Y = 1$
$X = 0$	.16	.37
$X = 1$	.29	.18

What is the marginal distribution of Y? What is the distribution of Y conditional on $X = 0$? What is the distribution of Y conditional on $X = 1$?

Demonstrate the Law of Iterated Expectations explicitly by showing that $\mathrm{E}(\mathrm{E}(X \,|\, Y)) = \mathrm{E}(X)$. Let $h(Y) = Y^3$. Show explicitly that $\mathrm{E}(Xh(Y) \,|\, Y) = h(Y)\mathrm{E}(X \,|\, Y)$ in this case.

1.6 Using expression (1.06) for the density $\phi(x)$ of the standard normal distribution, show that the derivative of $\phi(x)$ is the function $-x\phi(x)$, and that the second derivative is $(x^2 - 1)\phi(x)$. Use these facts to show that the expectation of a standard normal random variable is 0, and that its variance is 1. These two properties account for the use of the term "standard."

1.7 A normally distributed random variable can have any mean μ and any positive variance σ^2. Such a random variable is said to follow the $N(\mu, \sigma^2)$ distribution. A standard normal variable therefore has the $N(0, 1)$ distribution. Suppose that X has the standard normal distribution. Show that the random variable $Z \equiv \mu + \sigma X$ has mean μ and variance σ^2.

1.8 Compute the CDF of the $N(\mu, \sigma^2)$ distribution in terms of $\Phi(\cdot)$, the CDF of the standard normal distribution. Differentiate your answer so as to obtain the PDF of $N(\mu, \sigma^2)$.

1.9 If two random variables X_1 and X_2 are statistically independent, show that $\mathrm{E}(X_1 \,|\, X_2) = \mathrm{E}(X_1)$.

1.10 The **covariance** of two random variables X_1 and X_2, which is often written as $\mathrm{Cov}(X_1, X_2)$, is defined as the expectation of the product of $X_1 - \mathrm{E}(X_1)$ and $X_2 - \mathrm{E}(X_2)$. Consider a random variable X_1 with mean zero. Show that the covariance of X_1 and any other random variable X_2, whether it has mean zero or not, is just the expectation of the product of X_1 and X_2.

Show that the covariance of the random variables $\mathrm{E}(X_1 \,|\, X_2)$ and $X_1 - \mathrm{E}(X_1 \,|\, X_2)$ is zero. It is easiest to show this result by first showing that it is true when the covariance is computed conditional on X_2.

Show also that the variance of the random variable $X_1 - \mathrm{E}(X_1 \,|\, X_2)$ cannot be greater than the variance of X_1, and that the two variances will be equal if X_1 and X_2 are independent. This result shows how one random variable can be informative about another: Conditioning on it reduces variance unless the two variables are independent.

1.11 Prove that, if X_1 and X_2 are statistically independent, $\mathrm{Cov}(X_1, X_2) = 0$.

1.12 Let a random variable X_1 be distributed as $N(0, 1)$. Now suppose that a second random variable, X_2, is constructed as the product of X_1 and an independent random variable Z, which equals 1 with probability $1/2$ and -1 with probability $1/2$.

What is the (marginal) distribution of X_2? What is the covariance between X_1 and X_2? What is the distribution of X_1 conditional on X_2?

1.13 Consider the linear regression models

$$\begin{aligned} H_1\colon \quad & y_t = \beta_1 + \beta_2 X_t + u_t \quad \text{and} \\ H_2\colon \quad & \log y_t = \gamma_1 + \gamma_2 \log X_t + u_t. \end{aligned}$$

Suppose that the data are actually generated by H_2, with $\gamma_1 = 1.5$ and $\gamma_2 = 0.5$, and that the value of X_t varies from 10 to 110 with an average value of 60. Ignore the error terms and consider the deterministic relations between y_t and X_t implied by the two models. Find the values of β_1 and β_2 that make the relation given by H_1 have the same level and the same value of dy_t/dX_t as the level and value of dy_t/dX_t implied by the relation given by H_2 when it is evaluated at the average value of the regressor.

Using the deterministic relations, plot y_t as a function of X_t for both models for $10 \leq X_t \leq 110$. Also plot $\log y_t$ as a function of $\log X_t$ for both models for the same range of X_t. How well do the two models approximate each other in each of the plots?

1.14 Consider two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of dimensions such that the product $\boldsymbol{AB}$ exists. Show that the i^{th} row of $\boldsymbol{AB}$ is the matrix product of the i^{th} row of $\boldsymbol{A}$ with the entire matrix $\boldsymbol{B}$. Show that this result implies that the i^{th} row of a product $\boldsymbol{ABC}\ldots$, with arbitrarily many factors, is the product of the i^{th} row of $\boldsymbol{A}$ with $\boldsymbol{BC}\ldots$.

What is the corresponding result for the columns of $\boldsymbol{AB}$? What is the corresponding result for the columns of $\boldsymbol{ABC}\ldots$?

1.15 Consider two invertible square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, of the same dimensions. Show that the inverse of the product $\boldsymbol{AB}$ exists and is given by the formula

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}.$$

This shows that there is a **reversal rule** for inverses as well as for transposes; see (1.30).

1.16 Show that the transpose of the product of an arbitrary number of factors is the product of the transposes of the individual factors in completely reversed order:

$$(\boldsymbol{ABC}\cdots)^{\top} = \cdots \boldsymbol{C}^{\top}\boldsymbol{B}^{\top}\boldsymbol{A}^{\top}.$$

Show also that an analogous result holds for the inverse of the product of an arbitrary number of factors.

1.17 Consider the following example of multiplying partitioned matrices:

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix}.$$

Check all the expressions on the right-hand side, verifying that all products are well defined and that all sums are of matrices of the same dimensions.

1.18 Suppose that $\boldsymbol{X} = [\boldsymbol{\iota}\ \boldsymbol{X}_1\ \boldsymbol{X}_2]$, where $\boldsymbol{X}$ is $n \times k$, $\boldsymbol{\iota}$ is an n–vector of 1s, $\boldsymbol{X}_1$ is $n \times k_1$, and $\boldsymbol{X}_2$ is $n \times k_2$. What is the matrix $\boldsymbol{X}^{\top}\boldsymbol{X}$ in terms of the components of $\boldsymbol{X}$? What are the dimensions of its component matrices? What is the element in the upper left-hand corner of $\boldsymbol{X}^\top\boldsymbol{X}$ equal to?

1.19 Fix a sample size of $n = 100$, and simulate the very simplest regression model, namely, $y_t = \beta + u_t$. Set $\beta = 1$, and let the error terms u_t be drawings from the standard normal distribution. Compute the sample mean of the y_t,

$$\bar{y} \equiv \frac{1}{n}\sum_{t=1}^{n} y_t.$$

Use your favorite econometrics software package to run a regression with $\boldsymbol{y}$, the 100×1 vector with typical element y_t, as the dependent variable, and a constant as the sole explanatory variable. Show that the OLS estimate of the constant is equal to the sample mean. Why is this a necessary consequence of the formula (1.46)?

1.20 For the multiple linear regression model (1.44), the sum of squared residuals can be written as

$$\mathrm{SSR}(\boldsymbol{\beta}) = \sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Show that, if we minimize $\mathrm{SSR}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, the minimizing value of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, the OLS estimator given by (1.46). The easiest way is to show that the first-order conditions for a minimum are exactly the equations (1.47), or (1.48), that arise from MM estimation. This can be done without using matrix calculus.

1.21 The file **consumption.data** contains data on real personal disposable income and consumption expenditures in Canada, seasonally adjusted in 1986 dollars, from the first quarter of 1947 until the last quarter of 1996. The simplest imaginable model of the Canadian consumption function would have consumption expenditures as the dependent variable, and a constant and personal disposable income as explanatory variables. Run this regression for the period 1953:1 to 1996:4. What is your estimate of the marginal propensity to consume out of disposable income?

Plot a graph of the OLS residuals for the consumption function regression against time. All modern regression packages will generate these residuals for you on request. Does the appearance of the residuals suggest that this model of the consumption function is well specified?

1.22 Simulate the consumption function model you have just estimated in exercise 1.21 for the same sample period, using the actual data on disposable income. For the parameters, use the OLS estimates obtained in exercise 1.21. For the error terms, use drawings from the $N(0, s^2)$ distribution, where s^2 is the estimate of the error variance produced by the regression package.

Next, run a regression using the simulated consumption data as the dependent variable and the constant and disposable income as explanatory variables. Are the parameter estimates the same as those obtained using the real data? Why or why not?

Plot the residuals from the regression with simulated data. Does the plot look substantially different from the one obtained using the real data? It should!

Chapter 2

The Geometry of Linear Regression

2.1 Introduction

In Chapter 1, we introduced regression models, both linear and nonlinear, and discussed how to estimate linear regression models by using the method of moments. We saw that all n observations of a linear regression model with k regressors can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \tag{2.01}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are n-vectors, $\boldsymbol{X}$ is an $n \times k$ matrix, one column of which may be a constant term, and $\boldsymbol{\beta}$ is a k-vector. We also saw that the MM estimates, usually called the ordinary least squares or OLS estimates, of the vector $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{2.02}$$

In this chapter, we will be concerned with the **numerical properties** of these OLS estimates. We refer to certain properties of estimates as “numerical” if they have nothing to do with how the data were actually generated. Such properties hold for every set of data by virtue of the way in which $\hat{\boldsymbol{\beta}}$ is computed, and the fact that they hold can always be verified by direct calculation. In contrast, the **statistical properties** of OLS estimates, which will be discussed in Chapter 3, necessarily depend on unverifiable assumptions about how the data were generated, and they can never be verified for any actual data set.

In order to understand the numerical properties of OLS estimates, it is useful to look at them from the perspective of Euclidean geometry. This geometrical interpretation is remarkably simple. Essentially, it involves using Pythagoras’ Theorem and a little bit of high-school trigonometry in the context of finite-dimensional vector spaces. Although this approach is simple, it is very powerful. Once one has a thorough grasp of the geometry involved in ordinary least squares, one can often save oneself many tedious lines of algebra by a simple geometrical argument. We will encounter many examples of this throughout the book.

In the next section, we review some relatively elementary material on the geometry of vector spaces and Pythagoras’ Theorem. In Section 2.3, we then discuss the most important numerical properties of OLS estimation from a geometrical perspective. In Section 2.4, we introduce an extremely useful result called the FWL Theorem, and in Section 2.5 we present a number of applications of this theorem. Finally, in Section 2.6, we discuss how and to what extent individual observations influence parameter estimates.

2.2 The Geometry of Vector Spaces

In Section 1.4, an n–vector was defined as a column vector with n elements, that is, an $n \times 1$ matrix. The elements of such a vector are real numbers. The usual notation for the **real line** is $\mathbb{R}$, and it is therefore natural to denote the set of n–vectors as $\mathbb{R}^n$. However, in order to use the insights of Euclidean geometry to enhance our understanding of the algebra of vectors and matrices, it is desirable to introduce the notion of a **Euclidean space** in n dimensions, which we will denote as E^n. The difference between $\mathbb{R}^n$ and E^n is not that they consist of different sorts of vectors, but rather that a wider set of operations is defined on E^n. A shorthand way of saying that a vector $\boldsymbol{x}$ belongs to an n–dimensional Euclidean space is to write $\boldsymbol{x} \in E^n$.

Addition and subtraction of vectors in E^n is no different from the addition and subtraction of $n \times 1$ matrices discussed in Section 1.4. The same thing is true of multiplication by a scalar in E^n. The final operation essential to E^n is that of the scalar or inner product. For any two vectors $\boldsymbol{x}, \boldsymbol{y} \in E^n$, their scalar product is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv \boldsymbol{x}^\top \boldsymbol{y}.$$

The notation on the left is generally used in the context of the geometry of vectors, while the notation on the right is generally used in the context of matrix algebra. Note that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$, since $\boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x}$. Thus the scalar product is **commutative**.

The scalar product is what allows us to make a close connection between n–vectors considered as matrices and considered as geometrical objects. It allows us to define the **length** of any vector in E^n. The length, or **norm**, of a vector $\boldsymbol{x}$ is simply

$$\|\boldsymbol{x}\| \equiv (\boldsymbol{x}^\top \boldsymbol{x})^{1/2}.$$

This is just the square root of the inner product of $\boldsymbol{x}$ with itself. In scalar terms, it is

$$\|\boldsymbol{x}\| \equiv \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}. \tag{2.03}$$

Pythagoras' Theorem

The definition (2.03) is inspired by the celebrated theorem of Pythagoras, which says that the square on the longest side of a right-angled triangle is equal to the sum of the squares on the other two sides. This longest side

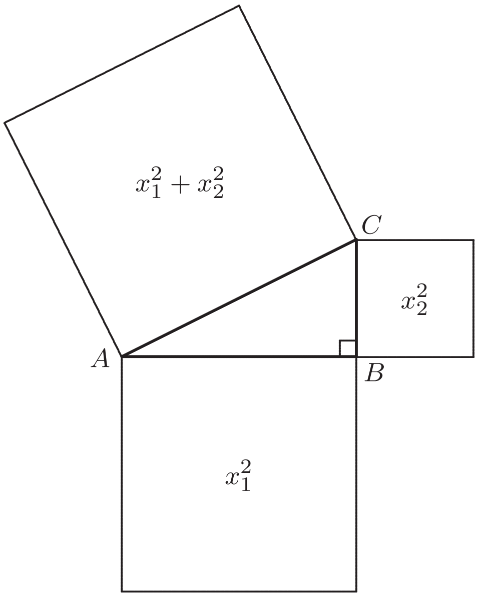


Figure 2.1 Pythagoras' Theorem

is called the **hypotenuse**. Pythagoras' Theorem is illustrated in Figure 2.1. The figure shows a right-angled triangle, ABC, with hypotenuse AC, and two other sides, AB and BC, of lengths x_1 and x_2 respectively. The squares on each of the three sides of the triangle are drawn, and the area of the square on the hypotenuse is shown as $x_1^2 + x_2^2$, in accordance with the theorem.

A beautiful proof of Pythagoras' Theorem, not often found in geometry texts, is shown in Figure 2.2. Two squares of equal area are drawn. Each square contains four copies of the same right-angled triangle. The square on the left also contains the squares on the two shorter sides of the triangle, while the

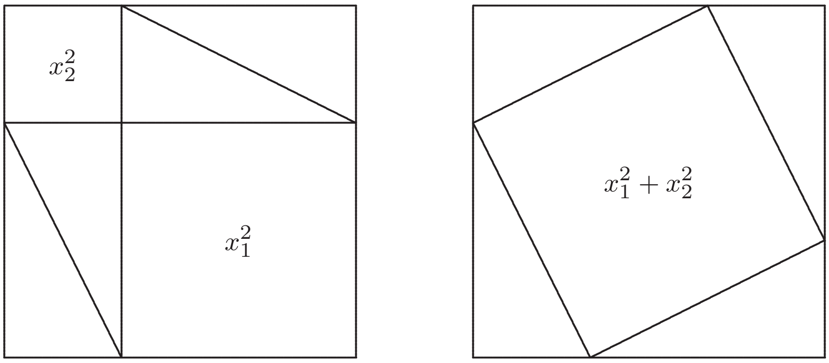


Figure 2.2 Proof of Pythagoras' Theorem

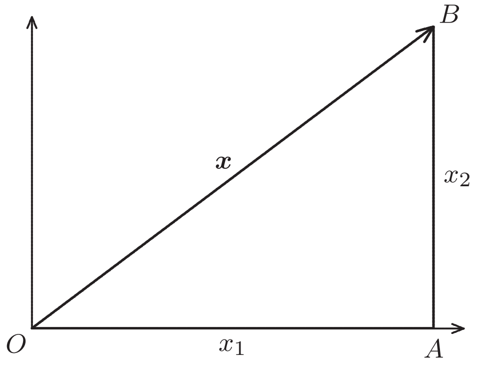


Figure 2.3 A vector $\boldsymbol{x}$ in E^2

square on the right contains the square on the hypotenuse. The theorem follows at once.

Any vector $\boldsymbol{x} \in E^2$ has two components, usually denoted as x_1 and x_2. These two components can be interpreted as the **Cartesian coordinates** of the vector in the plane. The situation is illustrated in Figure 2.3. With O as the origin of the coordinates, a right-angled triangle is formed by the lines OA, AB, and OB. The length of the horizontal side of the triangle, OA, is the horizontal coordinate x_1. The length of the vertical side, AB, is the vertical coordinate x_2. Thus the point B has Cartesian coordinates (x_1, x_2). The vector $\boldsymbol{x}$ itself is usually represented as the hypotenuse of the triangle, OB, that is, the directed line (depicted as an arrow) joining the origin to the point B, with coordinates (x_1, x_2). By Pythagoras' Theorem, the length of the vector $\boldsymbol{x}$, the hypotenuse of the triangle, is $(x_1^2 + x_2^2)^{1/2}$. This is what (2.03) becomes for the special case $n = 2$.

Vector Geometry in Two Dimensions

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two vectors in E^2, with components (x_1, x_2) and (y_1, y_2), respectively. Then, by the rules of matrix addition, the components of $\boldsymbol{x} + \boldsymbol{y}$ are $(x_1 + y_1, x_2 + y_2)$. Figure 2.4 shows how the addition of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be performed geometrically in two different ways. The vector $\boldsymbol{x}$ is drawn as the directed line segment, or arrow, from the origin O to the point A with coordinates (x_1, x_2). The vector $\boldsymbol{y}$ can be drawn similarly and represented by the arrow OB. However, we could also draw $\boldsymbol{y}$ starting, not at O, but at the point reached after drawing $\boldsymbol{x}$, namely A. The arrow AC has the same length and direction as OB, and we will see in general that arrows with the same length and direction can be taken to represent the same vector. It is clear by construction that the coordinates of C are $(x_1 + y_1, x_2 + y_2)$, that is, the coordinates of $\boldsymbol{x} + \boldsymbol{y}$. Thus the sum $\boldsymbol{x} + \boldsymbol{y}$ is represented geometrically by the arrow OC.

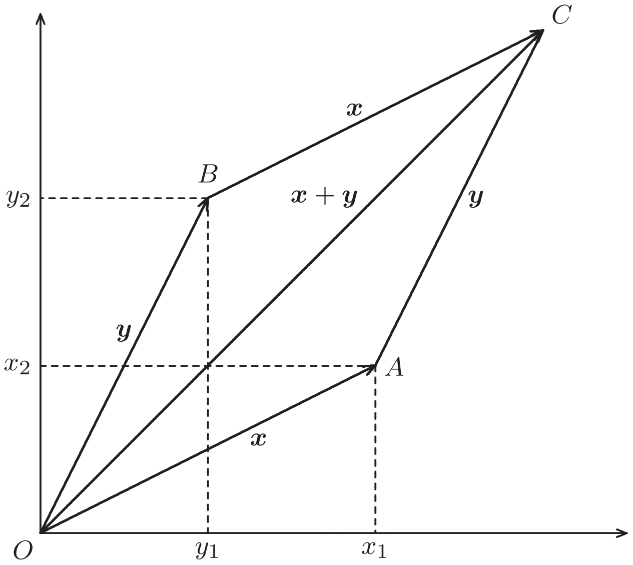


Figure 2.4 Addition of vectors

The classical way of adding vectors geometrically is to form a parallelogram using the line segments OA and OB that represent the two vectors as adjacent sides of the parallelogram. The sum of the two vectors is then the diagonal through O of the resulting parallelogram. It is easy to see that this classical method also gives the result that the sum of the two vectors is represented by the arrow OC, since the figure $OACB$ is just the parallelogram required by the construction, and OC is its diagonal through O. The parallelogram construction also shows clearly that vector addition is commutative, since $\boldsymbol{y}+\boldsymbol{x}$ is represented by OB, for $\boldsymbol{y}$, followed by BC, for $\boldsymbol{x}$. The end result is once more OC.

Multiplying a vector by a scalar is also very easy to represent geometrically. If a vector $\boldsymbol{x}$ with components (x_1, x_2) is multiplied by a scalar α, then $\alpha\boldsymbol{x}$ has components $(\alpha x_1, \alpha x_2)$. This is depicted in Figure 2.5, where $\alpha = 2$. The line segments OA and OB represent $\boldsymbol{x}$ and $\alpha\boldsymbol{x}$, respectively. It is clear that even if we move $\alpha\boldsymbol{x}$ so that it starts somewhere other than O, as with CD in the figure, the vectors $\boldsymbol{x}$ and $\alpha\boldsymbol{x}$ are always **parallel**. If α were negative, then $\alpha\boldsymbol{x}$ would simply point in the opposite direction. Thus, for $\alpha = -2$, $\alpha\boldsymbol{x}$ would be represented by DC, rather than CD.

Another property of multiplication by a scalar is clear from Figure 2.5. By direct calculation,

$$\|\alpha\boldsymbol{x}\| = \langle\alpha\boldsymbol{x}, \alpha\boldsymbol{x}\rangle^{1/2} = |\alpha|\,(\boldsymbol{x}^{\top}\boldsymbol{x})^{1/2} = |\alpha|\,\|\boldsymbol{x}\|. \tag{2.04}$$

Since $\alpha = 2$, OB and CD in the figure are twice as long as OA.

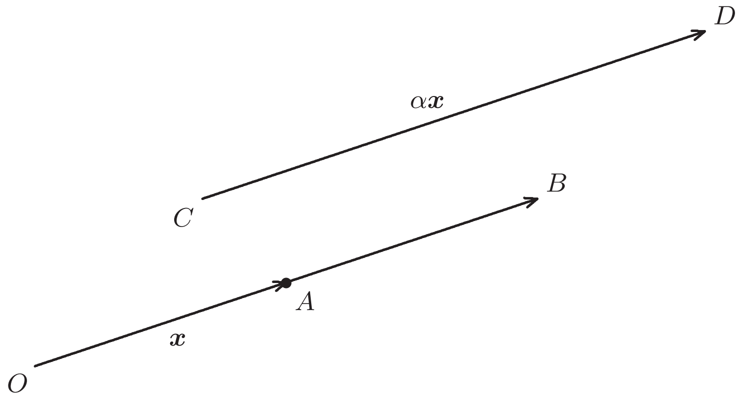


Figure 2.5 Multiplication by a scalar

The Geometry of Scalar Products

The scalar product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, whether in E^2 or E^n, can be expressed geometrically in terms of the lengths of the two vectors and the **angle** between them, and this result will turn out to be very useful. In the case of E^2, it is natural to think of the angle between two vectors as the angle between the two line segments that represent them. As we will now show, it is also quite easy to define the angle between two vectors in E^n.

If the angle between two vectors is 0, they must be **parallel**. The vector $\boldsymbol{y}$ is parallel to the vector $\boldsymbol{x}$ if $\boldsymbol{y} = \alpha\boldsymbol{x}$ for some suitable α. In that event,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \alpha\boldsymbol{x} \rangle = \alpha\boldsymbol{x}^\top\boldsymbol{x} = \alpha\|\boldsymbol{x}\|^2.$$

From (2.04), we know that $\|\boldsymbol{y}\| = |\alpha|\,\|\boldsymbol{x}\|$, and so, if $\alpha > 0$, it follows that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|. \tag{2.05}$$

Of course, this result is true only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel and point in the same direction (rather than in opposite directions).

For simplicity, consider initially two vectors, $\boldsymbol{w}$ and $\boldsymbol{z}$, both of length 1, and let θ denote the angle between them. This is illustrated in Figure 2.6. Suppose that the first vector, $\boldsymbol{w}$, has coordinates $(1, 0)$. It is therefore represented by a horizontal line of length 1 in the figure. Suppose that the second vector, $\boldsymbol{z}$, is also of length 1, that is, $\|\boldsymbol{z}\| = 1$. Then, by elementary trigonometry, the coordinates of $\boldsymbol{z}$ must be $(\cos\theta, \sin\theta)$. To show this, note first that, if so,

$$\|\boldsymbol{z}\|^2 = \cos^2\theta + \sin^2\theta = 1, \tag{2.06}$$

as required. Next, consider the right-angled triangle OAB, in which the hypotenuse OB represents $\boldsymbol{z}$ and is of length 1, by (2.06). The length of the side AB opposite O is $\sin\theta$, the vertical coordinate of $\boldsymbol{z}$. Then the sine of

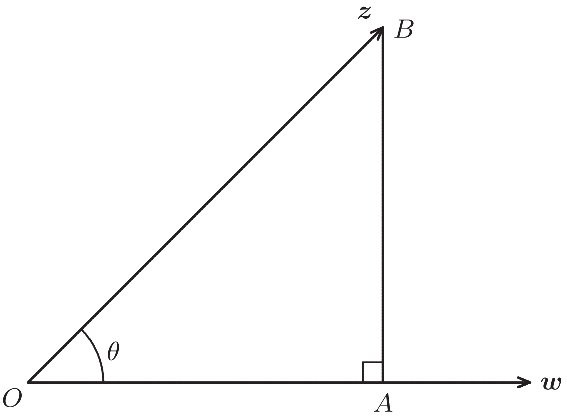


Figure 2.6 The angle between two vectors

the angle BOA is given, by the usual trigonometric rule, by the ratio of the length of the opposite side AB to that of the hypotenuse OB. This ratio is $\sin\theta/1 = \sin\theta$, and so the angle BOA is indeed equal to θ.

Now let us compute the scalar product of $\boldsymbol{w}$ and $\boldsymbol{z}$. It is

$$\langle \boldsymbol{w}, \boldsymbol{z} \rangle = \boldsymbol{w}^\top \boldsymbol{z} = w_1 z_1 + w_2 z_2 = z_1 = \cos\theta,$$

because $w_1 = 1$ and $w_2 = 0$. This result holds for vectors $\boldsymbol{w}$ and $\boldsymbol{z}$ of length 1. More generally, let $\boldsymbol{x} = \alpha\boldsymbol{w}$ and $\boldsymbol{y} = \gamma\boldsymbol{z}$, for positive scalars α and γ. Then $\|\boldsymbol{x}\| = \alpha$ and $\|\boldsymbol{y}\| = \gamma$. Thus we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \alpha\gamma\boldsymbol{w}^\top \boldsymbol{z} = \alpha\gamma\langle \boldsymbol{w}, \boldsymbol{z} \rangle.$$

Because $\boldsymbol{x}$ is parallel to $\boldsymbol{w}$, and $\boldsymbol{y}$ is parallel to $\boldsymbol{z}$, the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the same as that between $\boldsymbol{w}$ and $\boldsymbol{z}$, namely θ. Therefore,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|\cos\theta. \tag{2.07}$$

This is the general expression, in geometrical terms, for the scalar product of two vectors. It is true in E^n just as it is in E^2, although we have not proved this. In fact, we have not quite proved (2.07) even for the two-dimensional case, because we made the simplifying assumption that the direction of $\boldsymbol{x}$ and $\boldsymbol{w}$ is horizontal. In Exercise 2.1, we ask the reader to provide a more complete proof.

The cosine of the angle between two vectors provides a natural way to measure how close two vectors are in terms of their directions. Recall that $\cos\theta$ varies between -1 and 1; if we measure angles in radians, $\cos 0 = 1$, $\cos\pi/2 = 0$, and $\cos\pi = -1$. Thus $\cos\theta$ will be 1 for vectors that are parallel, 0 for vectors that are at right angles to each other, and -1 for vectors that point in directly opposite directions. If the angle θ between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is a right angle, its cosine is 0, and so, from (2.07), the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is 0. Conversely, if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, then $\cos\theta = 0$ unless $\boldsymbol{x}$ or $\boldsymbol{y}$ is a zero vector. If $\cos\theta = 0$, it follows that $\theta = \pi/2$. Thus, if two nonzero vectors have a zero scalar product, they are at right angles. Such vectors are often said to be **orthogonal**, or, less commonly, **perpendicular**. This definition implies that the zero vector is orthogonal to everything.

Since the cosine function can take on values only between -1 and 1, a consequence of (2.07) is that

$$|\boldsymbol{x}^\top \boldsymbol{y}| \le \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|. \tag{2.08}$$

This result, which is called the **Cauchy-Schwartz inequality**, says that the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$ can never be greater than the length of the vector $\boldsymbol{x}$ times the length of the vector $\boldsymbol{y}$. Only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel does the inequality in (2.08) become the equality (2.05). Readers are asked to prove this result in Exercise 2.2.

Subspaces of Euclidean Space

For arbitrary positive integers n, the elements of an n–vector can be thought of as the coordinates of a point in E^n. In particular, in the regression model (2.01), the regressand $\boldsymbol{y}$ and each column of the matrix of regressors $\boldsymbol{X}$ can be thought of as vectors in E^n. This makes it possible to represent a relationship like (2.01) geometrically.

It is obviously impossible to represent all n dimensions of E^n physically when $n > 3$. For the pages of a book, even three dimensions can be too many, although a proper use of perspective drawings can allow three dimensions to be shown. Fortunately, we can represent (2.01) without needing to draw in n dimensions. The key to this is that there are only three vectors in (2.01): $\boldsymbol{y}$, $\boldsymbol{X\beta}$, and $\boldsymbol{u}$. Since only two vectors, $\boldsymbol{X\beta}$ and $\boldsymbol{u}$, appear on the right-hand side of (2.01), only two dimensions are needed to represent it. Because $\boldsymbol{y}$ is equal to $\boldsymbol{X\beta} + \boldsymbol{u}$, these two dimensions suffice for $\boldsymbol{y}$ as well.

To see how this works, we need the concept of a **subspace** of a Euclidean space E^n. Normally, such a subspace will have a dimension lower than n. The easiest way to define a subspace of E^n is in terms of a set of **basis vectors**. A subspace that is of particular interest to us is the one for which the columns of $\boldsymbol{X}$ provide the basis vectors. We may denote the k columns of $\boldsymbol{X}$ as $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\ldots \boldsymbol{x}_k$. Then the subspace associated with these k basis vectors will be denoted by $\mathcal{S}(\boldsymbol{X})$ or $\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. The basis vectors are said to **span** this subspace, which will in general be a k–dimensional subspace.

The subspace $\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ consists of every vector that can be formed as a **linear combination** of the $\boldsymbol{x}_i$, $i = 1, \ldots, k$. Formally, it is defined as

$$\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \equiv \Big\{ \boldsymbol{z} \in E^n \;\Big|\; \boldsymbol{z} = \sum_{i=1}^{k} b_i \boldsymbol{x}_i, \quad b_i \in \mathbb{R} \Big\}. \tag{2.09}$$

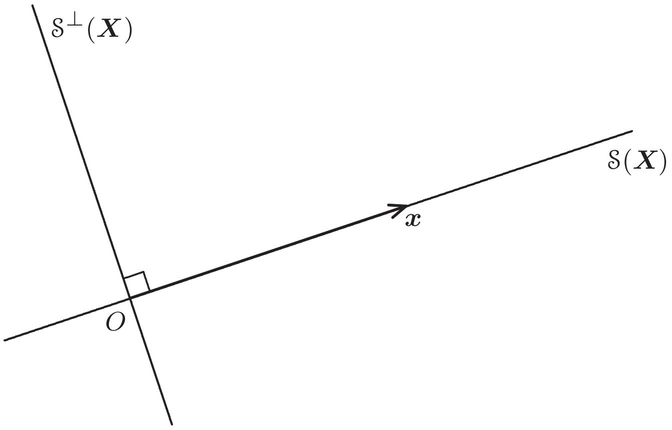

$\mathcal{S}^\perp(\boldsymbol{X})$
$\mathcal{S}(\boldsymbol{X})$
$\boldsymbol{x}$
O

Figure 2.7 The spaces $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$

The subspace defined in (2.09) is called the subspace spanned by the $\boldsymbol{x}_i$, $i = 1, \ldots, k$, or the **column space** of $\boldsymbol{X}$; less formally, it may simply be referred to as the **span** of $\boldsymbol{X}$, or the span of the $\boldsymbol{x}_i$.

The **orthogonal complement** of $\mathcal{S}(\boldsymbol{X})$ in E^n, which is denoted $\mathcal{S}^\perp(\boldsymbol{X})$, is the set of all vectors $\boldsymbol{w}$ in E^n that are orthogonal to everything in $\mathcal{S}(\boldsymbol{X})$. This means that, for every $\boldsymbol{z}$ in $\mathcal{S}(\boldsymbol{X})$, $\langle \boldsymbol{w}, \boldsymbol{z} \rangle = \boldsymbol{w}^\top \boldsymbol{z} = 0$. Formally,

$$\mathcal{S}^\perp(\boldsymbol{X}) \equiv \bigl\{ \boldsymbol{w} \in E^n \mid \boldsymbol{w}^\top \boldsymbol{z} = 0 \text{ for all } \boldsymbol{z} \in \mathcal{S}(\boldsymbol{X}) \bigr\}.$$

If the dimension of $\mathcal{S}(\boldsymbol{X})$ is k, then the dimension of $\mathcal{S}^\perp(\boldsymbol{X})$ is $n - k$.

Figure 2.7 illustrates the concepts of a subspace and its orthogonal complement for the simplest case, in which $n = 2$ and $k = 1$. The matrix $\boldsymbol{X}$ has only one column in this case, and it is therefore represented in the figure by a single vector, denoted $\boldsymbol{x}$. As a consequence, $\mathcal{S}(\boldsymbol{X})$ is 1–dimensional, and, since $n = 2$, $\mathcal{S}^\perp(\boldsymbol{X})$ is also 1–dimensional. Notice that $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$ would be the same if $\boldsymbol{x}$ were *any* vector, except for the origin, parallel to the straight line that represents $\mathcal{S}(\boldsymbol{X})$.

Now let us return to E^n. Suppose, to begin with, that $k = 2$. We have two vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which span a subspace of, at most, two dimensions. It is always possible to represent vectors in a 2–dimensional space on a piece of paper, whether that space is E^2 itself or, as in this case, the 2–dimensional subspace of E^n spanned by the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. To represent the first vector, $\boldsymbol{x}_1$, we choose an origin and a direction, both of which are entirely arbitrary, and draw an arrow of length $\|\boldsymbol{x}_1\|$ in that direction. Suppose that the origin is the point O in Figure 2.8, and that the direction is the horizontal direction in the plane of the page. Then an arrow to represent $\boldsymbol{x}_1$ can be drawn as shown in the figure. For $\boldsymbol{x}_2$, we compute its length, $\|\boldsymbol{x}_2\|$, and the angle, θ, that it makes with $\boldsymbol{x}_1$. Suppose for now that $\theta \neq 0$. Then we choose

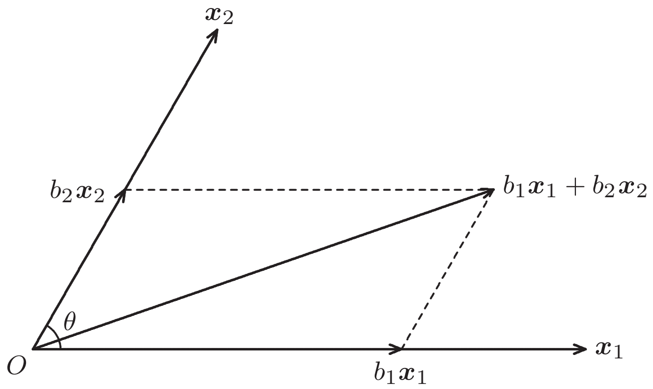


Figure 2.8 A 2-dimensional subspace

as our second dimension the vertical direction in the plane of the page, with the result that we can draw an arrow for $\boldsymbol{x}_2$, as shown.

Any vector in $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be drawn in the plane of Figure 2.8. Consider, for instance, the linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ given by the expression $\boldsymbol{z} \equiv b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$. We could draw the vector $\boldsymbol{z}$ by computing its length and the angle that it makes with $\boldsymbol{x}_1$. Alternatively, we could apply the rules for adding vectors geometrically that were illustrated in Figure 2.4 to the vectors $b_1\boldsymbol{x}_1$ and $b_2\boldsymbol{x}_2$. This is illustrated in the figure for the case in which $b_1 = {}^2\!/_3$ and $b_2 = {}^1\!/_2$.

In precisely the same way, we can represent any three vectors by arrows in 3–dimensional space, but we leave this task to the reader. It will be easier to appreciate the renderings of vectors in three dimensions in perspective that appear later on if one has already tried to draw 3–dimensional pictures, or even to model relationships in three dimensions with the help of a computer.

We can finally represent the regression model (2.01) geometrically. This is done in Figure 2.9. The horizontal direction is chosen for the vector $\boldsymbol{X\beta}$, and then the other two vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ are shown in the plane of the page. It is clear that, by construction, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. Notice that $\boldsymbol{u}$, the error vector, is not orthogonal to $\boldsymbol{X\beta}$. The figure contains no reference to any system of axes, because there would be n of them, and we would not be able to avoid needing n dimensions to treat them all.

Linear Independence

In order to define the OLS estimator by the formula (1.46), it is necessary to assume that the $k \times k$ square matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, or nonsingular. Equivalently, as we saw in Section 1.4, we may say that $\boldsymbol{X}^\top\boldsymbol{X}$ has full rank. This condition is equivalent to the condition that the columns of $\boldsymbol{X}$ should be **linearly independent**. This is a very important concept for econometrics. Note that the meaning of linear independence is quite different from the meaning

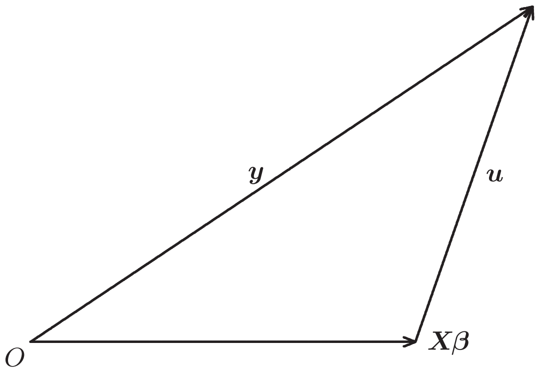


Figure 2.9 The geometry of the linear regression model

of statistical independence, which we discussed in Section 1.2. It is important not to confuse these two concepts.

The vectors $\boldsymbol{x}_1$ through $\boldsymbol{x}_k$ are said to be **linearly dependent** if we can write one of them as a linear combination of the others. In other words, there is a vector $\boldsymbol{x}_j$, $1 \leq j \leq k$, and coefficients c_i such that

$$\boldsymbol{x}_j = \sum_{i \neq j} c_i \boldsymbol{x}_i. \tag{2.10}$$

Another, equivalent, definition is that there exist coefficients b_i, at least one of which is nonzero, such that

$$\sum_{i=1}^{k} b_i \boldsymbol{x}_i = \mathbf{0}. \tag{2.11}$$

Recall that $\mathbf{0}$ denotes the **zero vector**, every component of which is 0. It is clear from the definition (2.11) that, if any of the $\boldsymbol{x}_i$ is itself equal to the zero vector, then the $\boldsymbol{x}_i$ are linearly dependent. If $\boldsymbol{x}_j = \mathbf{0}$, for example, then (2.11) will be satisfied if we make b_j nonzero and set $b_i = 0$ for all $i \neq j$.

If the vectors $\boldsymbol{x}_i$, $i = 1, \ldots, k$, are the columns of an $n \times k$ matrix $\boldsymbol{X}$, then another way of writing (2.11) is

$$\boldsymbol{Xb} = \mathbf{0}, \tag{2.12}$$

where $\boldsymbol{b}$ is a k–vector with typical element b_i. In order to see that (2.11) and (2.12) are equivalent, it is enough to check that the typical elements of the two left-hand sides are the same; see Exercise 2.5. The set of vectors $\boldsymbol{x}_i$, $i = 1, \ldots, k$, is linearly independent if it is not linearly dependent, that is, if there are no coefficients c_i such that (2.10) is true, or (equivalently) no coefficients b_i such that (2.11) is true, or (equivalently, once more) no vector $\boldsymbol{b}$ such that (2.12) is true.

It is easy to show that if the columns of $\boldsymbol{X}$ are linearly dependent, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is not invertible. Premultiplying (2.12) by $\boldsymbol{X}^\top$ yields

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}. \tag{2.13}$$

Thus, if the columns of $\boldsymbol{X}$ are linearly dependent, there is a nonzero k–vector $\boldsymbol{b}$ which is annihilated by $\boldsymbol{X}^\top\boldsymbol{X}$. The existence of such a vector $\boldsymbol{b}$ means that $\boldsymbol{X}^\top\boldsymbol{X}$ cannot be inverted. To see this, consider any vector $\boldsymbol{a}$, and suppose that

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{c}.$$

If $\boldsymbol{X}^\top\boldsymbol{X}$ could be inverted, then we could premultiply the above equation by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ to obtain

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c} = \boldsymbol{a}. \tag{2.14}$$

However, (2.13) also allows us to write

$$\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{a}+\boldsymbol{b}) = \boldsymbol{c},$$

which would give

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c} = \boldsymbol{a}+\boldsymbol{b}. \tag{2.15}$$

But (2.14) and (2.15) cannot both be true, and so $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ cannot exist. Thus a necessary condition for the existence of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is that the columns of $\boldsymbol{X}$ should be linearly independent. With a little more work, it can be shown that this condition is also sufficient, and so, if the regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are linearly independent, $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible.

If the k columns of $\boldsymbol{X}$ are not linearly independent, then they will span a subspace of dimension less than k, say k', where k' is the largest number of columns of $\boldsymbol{X}$ that are linearly independent of each other. The number k' is called the **rank** of $\boldsymbol{X}$. Look again at Figure 2.8, and imagine that the angle θ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ tends to zero. If $\theta = 0$, then $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are parallel, and we can write $\boldsymbol{x}_1 = \alpha\boldsymbol{x}_2$, for some scalar α. But this means that $\boldsymbol{x}_1 - \alpha\boldsymbol{x}_2 = \boldsymbol{0}$, and so a relation of the form (2.11) holds between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which are therefore linearly dependent. In the figure, if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are parallel, then only one dimension is used, and there is no need for the second dimension in the plane of the page. Thus, in this case, $k = 2$ and $k' = 1$.

When the dimension of $\mathcal{S}(\boldsymbol{X})$ is $k' < k$, $\mathcal{S}(\boldsymbol{X})$ will be identical to $\mathcal{S}(\boldsymbol{X}')$, where $\boldsymbol{X}'$ is an $n \times k'$ matrix consisting of any k' linearly independent columns of $\boldsymbol{X}$. For example, consider the following $\boldsymbol{X}$ matrix, which is 5×3:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{2.16}$$

The columns of this matrix are not linearly independent, since

$$\boldsymbol{x}_1 = .25\boldsymbol{x}_2 + \boldsymbol{x}_3.$$

However, any two of the columns are linearly independent, and so

$$\mathcal{S}(\boldsymbol{X}) = \mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_3) = \mathcal{S}(\boldsymbol{x}_2, \boldsymbol{x}_3);$$

see Exercise 2.8. For the remainder of this chapter, unless the contrary is explicitly assumed, we will assume that the columns of any regressor matrix $\boldsymbol{X}$ are linearly independent.

2.3 The Geometry of OLS Estimation

We studied the geometry of vector spaces in the last section because the numerical properties of OLS estimates are easily understood in terms of that geometry. The geometrical interpretation of OLS estimation, that is, MM estimation of linear regression models, is simple and intuitive. In many cases, it entirely does away with the need for algebraic proofs.

As we saw in the last section, any point in a subspace $\mathcal{S}(\boldsymbol{X})$, where $\boldsymbol{X}$ is an $n \times k$ matrix, can be represented as a linear combination of the columns of $\boldsymbol{X}$. We can partition $\boldsymbol{X}$ in terms of its columns explicitly, as follows:

$$\boldsymbol{X} = [\, \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k \,].$$

In order to compute the matrix product $\boldsymbol{X\beta}$ in terms of this partitioning, we need to partition the vector $\boldsymbol{\beta}$ by its rows. Since $\boldsymbol{\beta}$ has only one column, the elements of the partitioned vector are just the individual elements of $\boldsymbol{\beta}$. Thus we find that

$$\boldsymbol{X\beta} = [\, \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k \,] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \ldots + \boldsymbol{x}_k\beta_k = \sum_{i=1}^{k} \beta_i \boldsymbol{x}_i,$$

which is just a linear combination of the columns of $\boldsymbol{X}$. In fact, it is clear from the definition (2.09) that any linear combination of the columns of $\boldsymbol{X}$, and thus any element of the subspace $\mathcal{S}(\boldsymbol{X}) = \mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, can be written as $\boldsymbol{X\beta}$ for some $\boldsymbol{\beta}$. The specific linear combination (2.09) is constructed by using $\boldsymbol{\beta} = [b_1 \vdots \ldots \vdots b_k]$. Thus every n–vector $\boldsymbol{X\beta}$ belongs to $\mathcal{S}(\boldsymbol{X})$, which is, in general, a k–dimensional subspace of E^n. In particular, the vector $\boldsymbol{X\hat{\beta}}$ constructed using the OLS estimator $\boldsymbol{\hat{\beta}}$ belongs to this subspace.

The estimator $\boldsymbol{\hat{\beta}}$ was obtained by solving the equations (1.48), which we rewrite here for easy reference:

$$\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X\hat{\beta}}) = \boldsymbol{0}. \tag{1.48}$$

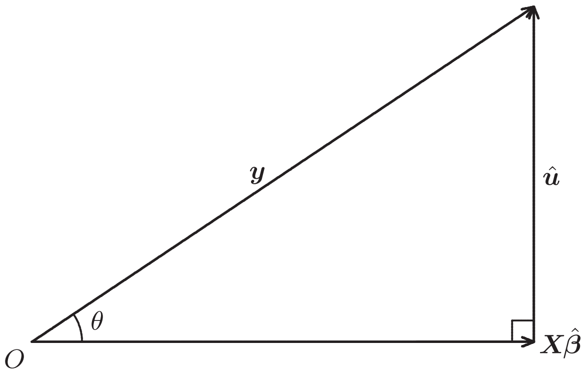


Figure 2.10 Residuals and fitted values

These equations have a simple geometrical interpretation. Note first that each element of the left-hand side of (1.48) is a scalar product. By the rule for selecting a single row of a matrix product (see Section 1.4), the i^{th} element is

$$\boldsymbol{x}_i^{\top}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \langle \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \rangle, \tag{2.17}$$

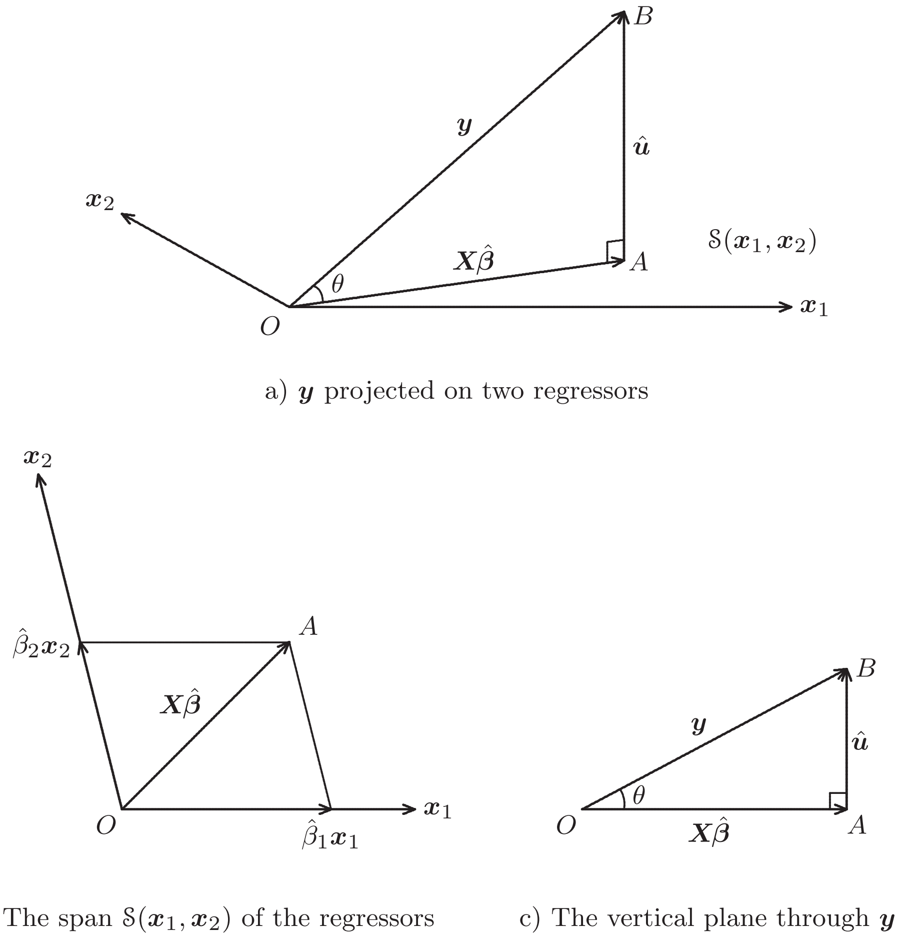
since $\boldsymbol{x}_i$, the i^{th} column of $\boldsymbol{X}$, is the transpose of the i^{th} row of $\boldsymbol{X}^{\top}$. By (1.48), the scalar product in (2.17) is zero, and so the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to all of the regressors, that is, all of the vectors $\boldsymbol{x}_i$ that represent the explanatory variables in the regression. For this reason, equations like (1.48) are often referred to as **orthogonality conditions**.

Recall from Section 1.5 that the vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, treated as a function of $\boldsymbol{\beta}$, is called the vector of residuals. This vector may be written as $\boldsymbol{u}(\boldsymbol{\beta})$. We are interested in $\boldsymbol{u}(\hat{\boldsymbol{\beta}})$, the vector of residuals evaluated at $\hat{\boldsymbol{\beta}}$, which is often called the vector of **least squares residuals** and is usually written simply as $\hat{\boldsymbol{u}}$. We have just seen, in (2.17), that $\hat{\boldsymbol{u}}$ is orthogonal to all the regressors. This implies that $\hat{\boldsymbol{u}}$ is in fact orthogonal to *every* vector in $\mathcal{S}(\boldsymbol{X})$, the span of the regressors. To see this, remember that any element of $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, with the result that, by (1.48),

$$\langle \boldsymbol{X}\boldsymbol{\beta}, \hat{\boldsymbol{u}} \rangle = (\boldsymbol{X}\boldsymbol{\beta})^{\top}\hat{\boldsymbol{u}} = \boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}\hat{\boldsymbol{u}} = \boldsymbol{0}.$$

The vector $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is referred to as the vector of **fitted values**. Clearly, it lies in $\mathcal{S}(\boldsymbol{X})$, and, consequently, it must be orthogonal to $\hat{\boldsymbol{u}}$. Figure 2.10 is similar to Figure 2.9, but it shows the vector of least squares residuals $\hat{\boldsymbol{u}}$ and the vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{u}$ and $\boldsymbol{X}\boldsymbol{\beta}$. The key feature of this figure, which is a consequence of the orthogonality conditions (1.48), is that the vector $\hat{\boldsymbol{u}}$ makes a right angle with the vector $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Some things about the orthogonality conditions (1.48) are clearer if we add a third dimension to the picture. Accordingly, in panel a) of Figure 2.11,


B
$\boldsymbol{y}$
$\hat{\boldsymbol{u}}$
$\boldsymbol{x}_2$
$\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$\boldsymbol{X}\hat{\boldsymbol{\beta}}$
A
θ
O
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
A
$\hat{\beta}_2\boldsymbol{x}_2$
$\boldsymbol{X}\hat{\boldsymbol{\beta}}$
B
$\boldsymbol{y}$
$\hat{\boldsymbol{u}}$
θ
O
$\boldsymbol{x}_1$
$\hat{\beta}_1\boldsymbol{x}_1$
O
$\boldsymbol{X}\hat{\boldsymbol{\beta}}$
A

a) $\boldsymbol{y}$ projected on two regressors

b) The span $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of the regressors

c) The vertical plane through $\boldsymbol{y}$

Figure 2.11 Linear regression in three dimensions

we consider the case of two regressors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which together span the horizontal plane labelled $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, seen in perspective from slightly above the plane. Although the perspective rendering of the figure does not make it clear, both the lengths of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and the angle between them are totally arbitrary, since they do not affect $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ at all. The vector $\boldsymbol{y}$ is intended to be viewed as rising up out of the plane spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

In the 3–dimensional setup, it is clear that, if $\hat{\boldsymbol{u}}$ is to be orthogonal to the horizontal plane, it must itself be vertical. Thus it is obtained by "dropping a perpendicular" from $\boldsymbol{y}$ to the horizontal plane. The least-squares interpretation of the MM estimator $\hat{\boldsymbol{\beta}}$ can now be seen to be a consequence of simple geometry. The shortest distance from $\boldsymbol{y}$ to the horizontal plane is obtained by descending vertically on to it, and the point in the horizontal plane vertically below $\boldsymbol{y}$, labeled A in the figure, is the closest point in the plane to $\boldsymbol{y}$. Thus $\|\hat{\boldsymbol{u}}\|$ minimizes $\|\boldsymbol{u}(\boldsymbol{\beta})\|$, the norm of $\boldsymbol{u}(\boldsymbol{\beta})$, with respect to $\boldsymbol{\beta}$.

The squared norm, $\|\boldsymbol{u}(\boldsymbol{\beta})\|^2$, is just the sum of squared residuals, $\text{SSR}(\boldsymbol{\beta})$; see (1.49). Since minimizing the norm of $\boldsymbol{u}(\boldsymbol{\beta})$ is the same thing as minimizing the squared norm, it follows that $\hat{\boldsymbol{\beta}}$ is the OLS estimator.

Panel b) of the figure shows the horizontal plane $\mathbb{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ as a straightforward 2–dimensional picture, seen from directly above. The point A is the point directly underneath $\boldsymbol{y}$, and so, since $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}$ by definition, the vector represented by the line segment OA is the vector of fitted values, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Geometrically, it is much simpler to represent $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ than to represent just the vector $\hat{\boldsymbol{\beta}}$, because the latter lies in $\mathbb{R}^k$, a different space from the space E^n that contains the variables and all linear combinations of them. However, it is easy to see that the information in panel b) does indeed determine $\hat{\boldsymbol{\beta}}$. Plainly, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ can be decomposed in just one way as a linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, as shown. The numerical value of $\hat{\beta}_1$ can be computed as the ratio of the length of the vector $\hat{\beta}_1\boldsymbol{x}_1$ to that of $\boldsymbol{x}_1$, and similarly for $\hat{\beta}_2$.

In panel c) of Figure 2.11, we show the right-angled triangle that corresponds to dropping a perpendicular from $\boldsymbol{y}$, labelled in the same way as in panel a). This triangle lies in the vertical plane that contains the vector $\boldsymbol{y}$. We can see that $\boldsymbol{y}$ is the hypotenuse of the triangle, the other two sides being $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. Thus this panel corresponds to what we saw already in Figure 2.10. Since we have a right-angled triangle, we can apply Pythagoras' Theorem. It gives

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\hat{\boldsymbol{u}}\|^2. \tag{2.18}$$

If we write out the squared norms as scalar products, this becomes

$$\boldsymbol{y}^\top\boldsymbol{y} = \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{2.19}$$

In words, the **total sum of squares**, or **TSS**, is equal to the **explained sum of squares**, or **ESS**, plus the **sum of squared residuals**, or **SSR**. This is a fundamental property of OLS estimates, and it will prove to be very useful in many contexts. Intuitively, it lets us break down the total variation (TSS) of the dependent variable into the explained variation (ESS) and the unexplained variation (SSR), unexplained because the residuals represent the aspects of $\boldsymbol{y}$ about which we remain in ignorance.

Orthogonal Projections

When we estimate a linear regression model, we implicitly map the regressand $\boldsymbol{y}$ into a vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Geometrically, these mappings are examples of orthogonal projections. A **projection** is a mapping that takes each point of E^n into a point in a subspace of E^n, while leaving all points in that subspace unchanged. Because of this, the subspace is called the **invariant subspace** of the projection. An **orthogonal projection** maps any point into the point of the subspace that is closest to it. If a point is already in the invariant subspace, it is mapped into itself.

The concept of an orthogonal projection formalizes the notion of "dropping a perpendicular" that we used in the last subsection when discussing least squares. Algebraically, an orthogonal projection on to a given subspace can be performed by premultiplying the vector to be projected by a suitable **projection matrix**. In the case of OLS, the two projection matrices that yield the vector of fitted values and the vector of residuals, respectively, are

$$\begin{aligned} \boldsymbol{P_X} &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \quad \text{and} \\ \boldsymbol{M_X} &= \mathbf{I} - \boldsymbol{P_X} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \end{aligned} \tag{2.20}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. To see this, recall (2.02), the formula for the OLS estimates of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

From this, we see that

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{y}. \tag{2.21}$$

Therefore, the first projection matrix in (2.20), $\boldsymbol{P_X}$, projects on to $\mathcal{S}(\boldsymbol{X})$. For any n–vector $\boldsymbol{y}$, $\boldsymbol{P_X}\boldsymbol{y}$ always lies in $\mathcal{S}(\boldsymbol{X})$, because

$$\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{X}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\big).$$

Since this takes the form $\boldsymbol{Xb}$ for $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$, it is a linear combination of the columns of $\boldsymbol{X}$, and hence it belongs to $\mathcal{S}(\boldsymbol{X})$.

From (2.20), it is easy to show that $\boldsymbol{P_X X} = \boldsymbol{X}$. Since any vector in $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{Xb}$ for some $\boldsymbol{b} \in \mathbb{R}^k$, we see that

$$\boldsymbol{P_X X b} = \boldsymbol{X b}. \tag{2.22}$$

We saw from (2.21) that the result of acting on any vector $\boldsymbol{y} \in E^n$ with $\boldsymbol{P_X}$ is a vector in $\mathcal{S}(\boldsymbol{X})$. Thus the invariant subspace of the projection $\boldsymbol{P_X}$ must be contained in $\mathcal{S}(\boldsymbol{X})$. But, by (2.22), *every* vector in $\mathcal{S}(\boldsymbol{X})$ is mapped into itself by $\boldsymbol{P_X}$. Therefore, the **image** of $\boldsymbol{P_X}$, which is a shorter name for its invariant subspace, is precisely $\mathcal{S}(\boldsymbol{X})$.

It is clear from (2.21) that, when $\boldsymbol{P_X}$ is applied to $\boldsymbol{y}$, it yields the vector of fitted values. Similarly, when $\boldsymbol{M_X}$, the second of the two projection matrices in (2.20), is applied to $\boldsymbol{y}$, it yields the vector of residuals:

$$\boldsymbol{M_X}\boldsymbol{y} = \big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{u}}.$$

The image of $\boldsymbol{M_X}$ is $\mathcal{S}^\perp(\boldsymbol{X})$, the orthogonal complement of the image of $\boldsymbol{P_X}$. To see this, consider any vector $\boldsymbol{w} \in \mathcal{S}^\perp(\boldsymbol{X})$. It must satisfy the defining condition $\boldsymbol{X}^\top\boldsymbol{w} = \boldsymbol{0}$. From the definition (2.20) of $\boldsymbol{P_X}$, this implies that $\boldsymbol{P_X}\boldsymbol{w} = \boldsymbol{0}$, the zero vector. Since $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{P_X}$, we find that $\boldsymbol{M_X w} = \boldsymbol{w}$. Thus $\mathcal{S}^\perp(\boldsymbol{X})$ must be contained in the image of $\boldsymbol{M_X}$. Next, consider any vector in the image of $\boldsymbol{M_X}$. It must take the form $\boldsymbol{M_X y}$, where $\boldsymbol{y}$ is some vector in E^n. From this, it will follow that $\boldsymbol{M_X y}$ belongs to $\mathcal{S}^\perp(\boldsymbol{X})$. Observe that

$$(\boldsymbol{M_X y})^\top \boldsymbol{X} = \boldsymbol{y}^\top \boldsymbol{M_X X}, \tag{2.23}$$

an equality that relies on the symmetry of $\boldsymbol{M_X}$. Then, from (2.20), we have

$$\boldsymbol{M_X X} = (\mathbf{I} - \boldsymbol{P_X})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{O}, \tag{2.24}$$

where $\mathbf{O}$ denotes a zero matrix, which in this case is $n \times k$. The result (2.23) says that any vector $\boldsymbol{M_X y}$ in the image of $\boldsymbol{M_X}$ is orthogonal to $\boldsymbol{X}$, and thus belongs to $\mathcal{S}^\perp(\boldsymbol{X})$. We saw above that $\mathcal{S}^\perp(\boldsymbol{X})$ was contained in the image of $\boldsymbol{M_X}$, and so this image must coincide with $\mathcal{S}^\perp(\boldsymbol{X})$. For obvious reasons, the projection $\boldsymbol{M_X}$ is sometimes called the projection **off** $\mathcal{S}(\boldsymbol{X})$.

For any matrix to represent a projection, it must be **idempotent**. An idempotent matrix is one that, when multiplied by itself, yields itself again. Thus,

$$\boldsymbol{P_X P_X} = \boldsymbol{P_X} \quad \text{and} \quad \boldsymbol{M_X M_X} = \boldsymbol{M_X}.$$

These results are easily proved by a little algebra directly from (2.20), but the geometry of the situation makes them obvious. If we take any point, project it on to $\mathcal{S}(\boldsymbol{X})$, and then project it on to $\mathcal{S}(\boldsymbol{X})$ *again*, the second projection can have no effect at all, because the point is *already* in $\mathcal{S}(\boldsymbol{X})$, and so it is left unchanged. Since this implies that $\boldsymbol{P_X P_X y} = \boldsymbol{P_X y}$ for any vector $\boldsymbol{y}$, it must be the case that $\boldsymbol{P_X P_X} = \boldsymbol{P_X}$, and similarly for $\boldsymbol{M_X}$.

Since, from (2.20),

$$\boldsymbol{P_X} + \boldsymbol{M_X} = \mathbf{I}, \tag{2.25}$$

any vector $\boldsymbol{y} \in E^n$ is equal to $\boldsymbol{P_X y} + \boldsymbol{M_X y}$. The pair of projections $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are said to be **complementary projections**, since the sum of $\boldsymbol{P_X y}$ and $\boldsymbol{M_X y}$ restores the original vector $\boldsymbol{y}$.

The fact that $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$ are orthogonal subspaces leads us to say that the two projection matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ define what is called an **orthogonal decomposition** of E^n, because the two vectors $\boldsymbol{M_X y}$ and $\boldsymbol{P_X y}$ lie in the two orthogonal subspaces. Algebraically, the orthogonality depends on the fact that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric matrices. To see this, we start from a further important property of $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$, which is that

$$\boldsymbol{P_X M_X} = \mathbf{O}. \tag{2.26}$$

This equation is true for any complementary pair of projections satisfying (2.25), whether or not they are symmetric; see Exercise 2.9. We may say that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ **annihilate** each other. Now consider any vector $\boldsymbol{z} \in \mathcal{S}(\boldsymbol{X})$

and any other vector $\boldsymbol{w} \in \mathcal{S}^\perp(\boldsymbol{X})$. We have $\boldsymbol{z} = \boldsymbol{P_X z}$ and $\boldsymbol{w} = \boldsymbol{M_X w}$. Thus the scalar product of the two vectors is

$$\langle \boldsymbol{P_X z}, \boldsymbol{M_X w} \rangle = \boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{X}}^\top \boldsymbol{M_X w}.$$

Since $\boldsymbol{P_X}$ is symmetric, $\boldsymbol{P}_{\boldsymbol{X}}^\top = \boldsymbol{P_X}$, and so the above scalar product is zero by (2.26). In general, however, if two complementary projection matrices are not symmetric, the spaces they project on to are not orthogonal.

The projection matrix $\boldsymbol{M_X}$ annihilates all points that lie in $\mathcal{S}(\boldsymbol{X})$, and $\boldsymbol{P_X}$ likewise annihilates all points that lie in $\mathcal{S}^\perp(\boldsymbol{X})$. These properties can be proved by straightforward algebra (see Exercise 2.11), but the geometry of the situation is very simple. Consider Figure 2.7. It is evident that, if we project any point in $\mathcal{S}^\perp(\boldsymbol{X})$ orthogonally on to $\mathcal{S}(\boldsymbol{X})$, we end up at the origin, as we do if we project any point in $\mathcal{S}(\boldsymbol{X})$ orthogonally on to $\mathcal{S}^\perp(\boldsymbol{X})$.

Provided that $\boldsymbol{X}$ has full rank, the subspace $\mathcal{S}(\boldsymbol{X})$ is k−dimensional, and so the first term in the decomposition $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}$ belongs to a k−dimensional space. Since $\boldsymbol{y}$ itself belongs to E^n, which has n dimensions, it follows that the complementary space $\mathcal{S}^\perp(\boldsymbol{X})$ must have $n - k$ dimensions. The number $n - k$ is called the **codimension** of $\boldsymbol{X}$ in E^n.

Geometrically, an orthogonal decomposition $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}$ can be represented by a right-angled triangle, with $\boldsymbol{y}$ as the hypotenuse and $\boldsymbol{P_X y}$ and $\boldsymbol{M_X y}$ as the other two sides. In terms of projections, equation (2.18), which is really just Pythagoras' Theorem, can be rewritten as

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2. \tag{2.27}$$

In Exercise 2.10, readers are asked to provide an algebraic proof of this equation. Since every term in (2.27) is nonnegative, we obtain the useful result that, for any orthogonal projection matrix $\boldsymbol{P_X}$ and any vector $\boldsymbol{y} \in E^n$,

$$\|\boldsymbol{P_X y}\| \leq \|\boldsymbol{y}\|. \tag{2.28}$$

In effect, this just says that the hypotenuse is longer than either of the other sides of a right-angled triangle.

In general, we will use $\boldsymbol{P}$ and $\boldsymbol{M}$ subscripted by matrix expressions to denote the matrices that, respectively, project on to and off the subspaces spanned by the columns of those matrix expressions. Thus $\boldsymbol{P_Z}$ would be the matrix that projects on to $\mathcal{S}(\boldsymbol{Z})$, $\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{W}}$ would be the matrix that projects off $\mathcal{S}(\boldsymbol{X}, \boldsymbol{W})$, or, equivalently, on to $\mathcal{S}^\perp(\boldsymbol{X}, \boldsymbol{W})$, and so on. It is frequently very convenient to express the quantities that arise in econometrics using these matrices, partly because the resulting expressions are relatively compact, and partly because the properties of projection matrices often make it easy to understand what those expressions mean. However, projection matrices are of little use for computation because they are of dimension $n \times n$. It is never efficient to calculate residuals or fitted values by explicitly using projection matrices, and it can be extremely inefficient if n is large.

Linear Transformations of Regressors

The span $\mathcal{S}(\boldsymbol{X})$ of the regressors of a linear regression can be defined in many equivalent ways. All that is needed is a set of k vectors that encompass all the k directions of the k–dimensional subspace. Consider what happens when we postmultiply $\boldsymbol{X}$ by any nonsingular $k \times k$ matrix $\boldsymbol{A}$. This is called a **nonsingular linear transformation**. Let $\boldsymbol{A}$ be partitioned by its columns, which may be denoted $\boldsymbol{a}_i$, $i = 1, \ldots, k$:

$$\boldsymbol{X}\boldsymbol{A} = \boldsymbol{X}\,[\,\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_k\,] = [\,\boldsymbol{X}\boldsymbol{a}_1 \quad \boldsymbol{X}\boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{X}\boldsymbol{a}_k\,]\,.$$

Each block in the product takes the form $\boldsymbol{X}\boldsymbol{a}_i$, which is an n–vector that is a linear combination of the columns of $\boldsymbol{X}$. Thus any element of $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$ must also be an element of $\mathcal{S}(\boldsymbol{X})$. But any element of $\mathcal{S}(\boldsymbol{X})$ is also an element of $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$. To see this, note that any element of $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. Since $\boldsymbol{A}$ is nonsingular, and thus invertible,

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{A}^{-1}\boldsymbol{\beta}).$$

Because $\boldsymbol{A}^{-1}\boldsymbol{\beta}$ is just a k–vector, this expression is a linear combination of the columns of $\boldsymbol{X}\boldsymbol{A}$, that is, an element of $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$. Since every element of $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$ belongs to $\mathcal{S}(\boldsymbol{X})$, and every element of $\mathcal{S}(\boldsymbol{X})$ belongs to $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$, these two subspaces must be identical.

Given the identity of $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}(\boldsymbol{X}\boldsymbol{A})$, it seems intuitively compelling to suppose that the orthogonal projections $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}}$ should be the same. This is in fact the case, as can be verified directly:

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}} &= \boldsymbol{X}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top \\ &= \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top \\ &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}}.\end{aligned}$$

When expanding the inverse of the matrix $\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}$, we used the reversal rule for inverses; see Exercise 1.15.

We have already seen that the vectors of fitted values and residuals depend on $\boldsymbol{X}$ only through $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$. Therefore, they too must be invariant to any nonsingular linear transformation of the columns of $\boldsymbol{X}$. Thus if, in the regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, we replace $\boldsymbol{X}$ by $\boldsymbol{X}\boldsymbol{A}$ for some nonsingular matrix $\boldsymbol{A}$, the residuals and fitted values will not change, even though $\hat{\boldsymbol{\beta}}$ will change. We will discuss an example of this important result shortly.

When the set of regressors contains a constant, it is necessary to express it as a vector, just like any other regressor. The coefficient of this vector is then the parameter we usually call the constant term. The appropriate vector is $\boldsymbol{\iota}$, the vector of which each element equals 1. Consider the n–vector $\beta_1\boldsymbol{\iota} + \beta_2\boldsymbol{x}$, where $\boldsymbol{x}$ is any nonconstant regressor, and β_1 and β_2 are scalar parameters. The $t^{\rm th}$ element of this vector is $\beta_1 + \beta_2 x_t$. Thus adding the vector $\beta_1\boldsymbol{\iota}$ to $\beta_2\boldsymbol{x}$ simply adds the scalar β_1 to each component of $\beta_2\boldsymbol{x}$. For any regression which includes a constant term, then, the fact that we can perform arbitrary nonsingular transformations of the regressors without affecting residuals or fitted values implies that these vectors are unchanged if we add any constant amount to any one or more of the regressors.

Another implication of the invariance of residuals and fitted values under nonsingular transformations of the regressors is that these vectors are unchanged if we change the **units of measurement** of the regressors. Suppose, for instance, that the temperature is one of the explanatory variables in a regression with a constant term. A practical example in which the temperature could have good explanatory power is the modeling of electricity demand: More electrical power is consumed if the weather is very cold, or, in societies where air conditioners are common, very hot. In a few countries, notably the United States, temperatures are still measured in Fahrenheit degrees, while in most countries they are measured in Celsius (centigrade) degrees. It would be disturbing if our conclusions about the effect of temperature on electricity demand depended on whether we used the Fahrenheit or the Celsius scale.

Let the temperature variable, expressed as an n–vector, be denoted as $\boldsymbol{T}$ in Celsius and as $\boldsymbol{F}$ in Fahrenheit, the constant as usual being represented by $\boldsymbol{\iota}$. Then $\boldsymbol{F} = 32\boldsymbol{\iota} + {}^9\!/_5\boldsymbol{T}$, and, if the constant is included in the transformation,

$$[\,\boldsymbol{\iota} \quad \boldsymbol{F}\,] = [\,\boldsymbol{\iota} \quad \boldsymbol{T}\,]\begin{bmatrix} 1 & 32 \\ 0 & {}^9\!/_5 \end{bmatrix}. \tag{2.29}$$

The constant and the two different temperature measures are related by a linear transformation that is easily seen to be nonsingular, since Fahrenheit degrees can be converted back into Celsius. This implies that the residuals and fitted values are unaffected by our choice of temperature scale.

Let us denote the constant term and the slope coefficient as β_1 and β_2 if we use the Celsius scale, and as α_1 and α_2 if we use the Fahrenheit scale. Then it is easy to see that these parameters are related by the equations

$$\beta_1 = \alpha_1 + 32\alpha_2 \quad \text{and} \quad \beta_2 = {}^9\!/_5\alpha_2. \tag{2.30}$$

To see that this makes sense, suppose that the temperature is at freezing point, which is 0° Celsius and 32° Fahrenheit. Then the combined effect of the constant and the temperature on electricity demand is $\beta_1 + 0\beta_2 = \beta_1$ using the Celsius scale, and $\alpha_1 + 32\alpha_2$ using the Fahrenheit scale. These should be the same, and, according to (2.30), they are. Similarly, the effect of a 1-degree increase in the Celsius temperature is given by β_2. Now 1 Celsius degree equals ${}^9\!/_5$ Fahrenheit degrees, and the effect of a temperature increase of ${}^9\!/_5$ Fahrenheit degrees is given by ${}^9\!/_5\alpha_2$. We are assured by (2.30) that the two effects are the same.

2.4 The Frisch-Waugh-Lovell Theorem

In this section, we discuss an extremely useful property of least squares estimates, which we will refer to as the **Frisch-Waugh-Lovell Theorem**, or **FWL Theorem** for short. It was introduced to econometricians by Frisch and Waugh (1933), and then reintroduced by Lovell (1963).

Deviations from the Mean

We begin by considering a particular nonsingular transformation of variables in a regression with a constant term. We saw at the end of the last section that residuals and fitted values are invariant under such transformations of the regressors. For simplicity, consider a model with a constant and just one explanatory variable:

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x} + \boldsymbol{u}. \tag{2.31}$$

In general, $\boldsymbol{x}$ is not orthogonal to $\boldsymbol{\iota}$, but there is a very simple transformation which makes it so. This transformation replaces the observations in $\boldsymbol{x}$ by **deviations from the mean**. In order to perform the transformation, one first calculates the mean of the n observations of the vector $\boldsymbol{x}$,

$$\bar{x} \equiv \frac{1}{n} \sum_{t=1}^{n} x_t,$$

and then subtracts the constant $\bar{x}$ from each element of $\boldsymbol{x}$. This yields the vector of deviations from the mean, $\boldsymbol{z} \equiv \boldsymbol{x} - \bar{x}\boldsymbol{\iota}$. The vector $\boldsymbol{z}$ is easily seen to be orthogonal to $\boldsymbol{\iota}$, because

$$\boldsymbol{\iota}^{\top}\boldsymbol{z} = \boldsymbol{\iota}^{\top}(\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) = n\bar{x} - \bar{x}\boldsymbol{\iota}^{\top}\boldsymbol{\iota} = n\bar{x} - n\bar{x} = 0.$$

The operation of expressing a variable in terms of the deviations from its mean is called **centering** the variable. In this case, the vector $\boldsymbol{z}$ is the **centered** version of the vector $\boldsymbol{x}$.

Since centering leads to a variable that is orthogonal to $\boldsymbol{\iota}$, it can be performed algebraically by the orthogonal projection matrix $\boldsymbol{M}_{\iota}$. This can be verified by observing that

$$\boldsymbol{M}_{\iota}\boldsymbol{x} = (\mathbf{I} - \boldsymbol{P}_{\iota})\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{\iota}(\boldsymbol{\iota}^{\top}\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^{\top}\boldsymbol{x} = \boldsymbol{x} - \bar{x}\boldsymbol{\iota} = \boldsymbol{z}, \tag{2.32}$$

as claimed. Here, we once again used the facts that $\boldsymbol{\iota}^{\top}\boldsymbol{\iota} = n$ and $\boldsymbol{\iota}^{\top}\boldsymbol{x} = n\bar{x}$.

The idea behind the use of deviations from the mean is that it makes sense to separate the overall level of a dependent variable from its dependence on explanatory variables. Specifically, if we write (2.31) in terms of $\boldsymbol{z}$, we get

$$\boldsymbol{y} = (\beta_1 + \beta_2\bar{x})\boldsymbol{\iota} + \beta_2\boldsymbol{z} + \boldsymbol{u} = \alpha_1\boldsymbol{\iota} + \alpha_2\boldsymbol{z} + \boldsymbol{u},$$

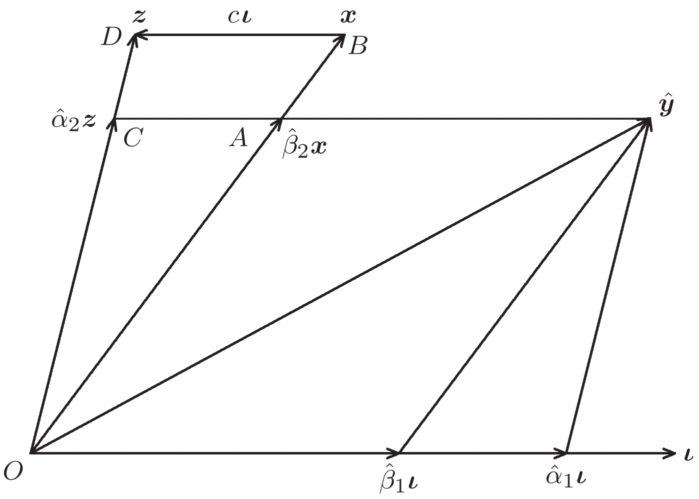

$\boldsymbol{z}$
$c\boldsymbol{\iota}$
$\boldsymbol{x}$
D
B
$\hat{\boldsymbol{y}}$
$\hat{\alpha}_2\boldsymbol{z}$
C
A
$\hat{\beta}_2\boldsymbol{x}$
O
$\hat{\beta}_1\boldsymbol{\iota}$
$\hat{\alpha}_1\boldsymbol{\iota}$
$\boldsymbol{\iota}$

Figure 2.12 Adding a constant does not affect the slope coefficient

where we see that

$$\alpha_1 = \beta_1 + \beta_2\bar{x}, \text{ and } \alpha_2 = \beta_2.$$

If, for some observation t, the value of x_t were exactly equal to the mean value, $\bar{x}$, then $z_t = 0$. Thus we find that $y_t = \alpha_1 + u_t$. We interpret this as saying that the expected value of y_t, when the explanatory variable takes on its average value, is the constant α_1.

The effect on y_t of a change of one unit in x_t is measured by the slope coefficient β_2. If we hold $\bar{x}$ at its value before x_t is changed, then the unit change in x_t induces a unit change in z_t. Thus a unit change in z_t, which is measured by the slope coefficient α_2, should have the same effect as a unit change in x_t. Accordingly, $\alpha_2 = \beta_2$, just as we found above.

The slope coefficients α_2 and β_2 would be the same with any constant in the place of $\bar{x}$. The reason for this can be seen geometrically, as illustrated in Figure 2.12. This figure, which is constructed in the same way as panel b) of Figure 2.11, depicts the span of $\boldsymbol{\iota}$ and $\boldsymbol{x}$, with $\boldsymbol{\iota}$ in the horizontal direction. As before, the vector $\boldsymbol{y}$ is not shown, because a third dimension would be required; the vector would extend from the origin to a point off the plane of the page and directly above (or below) the point labelled $\hat{\boldsymbol{y}}$.

The figure shows the vector of fitted values $\hat{\boldsymbol{y}}$ as the vector sum $\hat{\beta}_1\boldsymbol{\iota} + \hat{\beta}_2\boldsymbol{x}$. The slope coefficient $\hat{\beta}_2$ is the ratio of the length of the vector $\hat{\beta}_2\boldsymbol{x}$ to that of $\boldsymbol{x}$; geometrically, it is given by the ratio OA/OB. Then a new regressor $\boldsymbol{z}$ is defined by adding the constant value c, which is negative in the figure, to each component of $\boldsymbol{x}$, giving $\boldsymbol{z} = \boldsymbol{x} + c\boldsymbol{\iota}$. In terms of this new regressor, the vector $\hat{\boldsymbol{y}}$ is given by $\hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z}$, and $\hat{\alpha}_2$ is given by the ratio OC/OD. Since the ratios OA/OB and OC/OD are clearly the same, we see that $\hat{\alpha}_2 = \hat{\beta}_2$. A formal argument would use the fact that OAC and OBD are similar triangles.

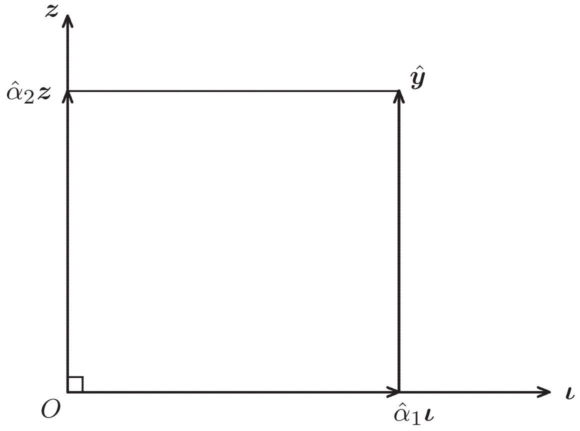

$\boldsymbol{z}$
$\hat{\alpha}_2\boldsymbol{z}$
$\hat{\boldsymbol{y}}$
O
$\hat{\alpha}_1\boldsymbol{\iota}$
$\boldsymbol{\iota}$

Figure 2.13 Orthogonal regressors may be omitted

When the constant c is chosen as $\bar{x}$, the vector $\boldsymbol{z}$ is said to be centered, and, as we saw above, it is orthogonal to $\boldsymbol{\iota}$. In this case, the estimate $\hat{\alpha}_2$ is the same whether it is obtained by regressing $\boldsymbol{y}$ on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$, or just on $\boldsymbol{z}$ alone. This is illustrated in Figure 2.13, which shows what Figure 2.12 would look like when $\boldsymbol{z}$ is orthogonal to $\boldsymbol{\iota}$. Once again, the vector of fitted values $\hat{\boldsymbol{y}}$ is decomposed as $\hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z}$, with $\boldsymbol{z}$ now at right angles to $\boldsymbol{\iota}$.

Now suppose that $\boldsymbol{y}$ is regressed on $\boldsymbol{z}$ alone. This means that $\boldsymbol{y}$ is projected orthogonally on to $\mathcal{S}(\boldsymbol{z})$, which in the figure is the vertical line through $\boldsymbol{z}$. By definition,

$$\boldsymbol{y} = \hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z} + \hat{\boldsymbol{u}}, \tag{2.33}$$

where $\hat{\boldsymbol{u}}$ is orthogonal to both $\boldsymbol{\iota}$ and $\boldsymbol{z}$. But $\boldsymbol{\iota}$ is also orthogonal to $\boldsymbol{z}$, and so the only term on the right-hand side of (2.33) not to be annihilated by the projection on to $\mathcal{S}(\boldsymbol{z})$ is the middle term, which is left unchanged by it. Thus the fitted value vector from regressing $\boldsymbol{y}$ on $\boldsymbol{z}$ alone is just $\hat{\alpha}_2\boldsymbol{z}$, and so the OLS estimate is the same $\hat{\alpha}_2$ as given by the regression on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$. Geometrically, we obtain this result because the projection of $\boldsymbol{y}$ on to $\mathcal{S}(\boldsymbol{z})$ is the same as the projection of $\hat{\boldsymbol{y}}$ on to $\mathcal{S}(\boldsymbol{z})$.

Incidentally, the fact that OLS residuals are orthogonal to all the regressors, including $\boldsymbol{\iota}$, leads to the important result that the residuals in any regression with a constant term sum to zero. In fact,

$$\boldsymbol{\iota}^\top\hat{\boldsymbol{u}} = \sum_{t=1}^{n} \hat{u}_t = 0;$$

recall (1.29). The residuals will also sum to zero in any regression for which $\boldsymbol{\iota} \in \mathcal{S}(\boldsymbol{X})$, even if $\boldsymbol{\iota}$ does not explicitly appear in the list of regressors. This can happen if the regressors include certain sets of **dummy variables**, as we will see in Section 2.5.

Two Groups of Regressors

The results proved in the previous subsection are actually special cases of more general results that apply to any regression in which the regressors can logically be broken up into two groups. Such a regression can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{2.34}$$

where $\boldsymbol{X}_1$ is $n \times k_1$, $\boldsymbol{X}_2$ is $n \times k_2$, and $\boldsymbol{X}$ may be written as the partitioned matrix $[\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$, with $k = k_1 + k_2$. In the case dealt with in the previous subsection, $\boldsymbol{X}_1$ is the constant vector $\boldsymbol{\iota}$ and $\boldsymbol{X}_2$ is either $\boldsymbol{x}$ or $\boldsymbol{z}$. Several other examples of partitioning $\boldsymbol{X}$ in this way will be considered in Section 2.5.

We begin by assuming that all the regressors in $\boldsymbol{X}_1$ are orthogonal to all the regressors in $\boldsymbol{X}_2$, so that $\boldsymbol{X}_2^\top\boldsymbol{X}_1 = \mathbf{O}$. Under this assumption, the vector of least squares estimates $\hat{\boldsymbol{\beta}}_1$ from (2.34) is the same as the one obtained from the regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1, \tag{2.35}$$

and $\hat{\boldsymbol{\beta}}_2$ from (2.34) is likewise the same as the vector of estimates obtained from the regression $\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}_2$. In other words, when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, we can drop either set of regressors from (2.34) without affecting the coefficients of the other set.

The vector of fitted values from (2.34) is $\boldsymbol{P_X y}$, while that from (2.35) is $\boldsymbol{P}_1\boldsymbol{y}$, where we have used the abbreviated notation

$$\boldsymbol{P}_1 \equiv \boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top.$$

As we will show directly,

$$\boldsymbol{P}_1\boldsymbol{P_X} = \boldsymbol{P_X}\boldsymbol{P}_1 = \boldsymbol{P}_1; \tag{2.36}$$

this is true whether or not $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal. Thus

$$\boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{P_X y} = \boldsymbol{P}_1(\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2) = \boldsymbol{P}_1\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1. \tag{2.37}$$

The first equality above, which follows from (2.36), says that the projection of $\boldsymbol{y}$ on to $\mathcal{S}(\boldsymbol{X}_1)$ is the same as the projection of $\hat{\boldsymbol{y}} \equiv \boldsymbol{P_X y}$ on to $\mathcal{S}(\boldsymbol{X}_1)$. The second equality follows from the definition of the fitted value vector from (2.34) as $\boldsymbol{P_X y}$; the third from the orthogonality of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, which implies that $\boldsymbol{P}_1\boldsymbol{X}_2 = \mathbf{O}$; and the last from the fact that $\boldsymbol{X}_1$ is invariant under the action of $\boldsymbol{P}_1$. Since $\boldsymbol{P}_1\boldsymbol{y}$ is equal to $\boldsymbol{X}_1$ postmultiplied by the OLS estimates from (2.35), the equality of the leftmost and rightmost expressions in (2.37) gives us the result that the same $\hat{\boldsymbol{\beta}}_1$ can be obtained either from (2.34) or from (2.35). The analogous result for $\hat{\boldsymbol{\beta}}_2$ is proved in just the same way.

We now drop the assumption that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal and prove (2.36), a very useful result that is true in general. In order to show that $\boldsymbol{P_X P}_1 = \boldsymbol{P}_1$, we proceed as follows:

$$\boldsymbol{P_X P}_1 = \boldsymbol{P_X X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top} = \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top} = \boldsymbol{P}_1.$$

The middle equality follows by noting that $\boldsymbol{P_X X}_1 = \boldsymbol{X}_1$, because all the columns of $\boldsymbol{X}_1$ are in $\mathcal{S}(\boldsymbol{X})$, and so are left unchanged by $\boldsymbol{P_X}$. The other equality in (2.36), namely $\boldsymbol{P}_1\boldsymbol{P_X} = \boldsymbol{P}_1$, is obtained directly by transposing $\boldsymbol{P_X P}_1 = \boldsymbol{P}_1$ and using the symmetry of $\boldsymbol{P_X}$ and $\boldsymbol{P}_1$. The two results in (2.36) tell us that the product of two orthogonal projections, where one projects on to a subspace of the image of the other, is the projection on to that subspace. See also Exercise 2.14, for the application of this result to the complementary projections $\boldsymbol{M_X}$ and $\boldsymbol{M}_1$.

The general result corresponding to the one shown in Figure 2.12 can be stated as follows. If we transform the regressor matrix in (2.34) by adding $\boldsymbol{X}_1\boldsymbol{A}$ to $\boldsymbol{X}_2$, where $\boldsymbol{A}$ is a $k_1 \times k_2$ matrix, and leaving $\boldsymbol{X}_1$ as it is, we have the regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + (\boldsymbol{X}_2 + \boldsymbol{X}_1\boldsymbol{A})\boldsymbol{\alpha}_2 + \boldsymbol{u}. \tag{2.38}$$

Then $\hat{\boldsymbol{\alpha}}_2$ from (2.38) is the same as $\hat{\boldsymbol{\beta}}_2$ from (2.34). This can be seen immediately by expressing the right-hand side of (2.38) as a linear combination of the columns of $\boldsymbol{X}_1$ and of $\boldsymbol{X}_2$.

In the present general context, there is an operation analogous to that of centering. The result of centering a variable $\boldsymbol{x}$ is a variable $\boldsymbol{z}$ that is orthogonal to $\boldsymbol{\iota}$, the constant. We can create from $\boldsymbol{X}_2$ a set of variables orthogonal to $\boldsymbol{X}_1$ by acting on $\boldsymbol{X}_2$ with the orthogonal projection $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{P}_1$, so as to obtain $\boldsymbol{M}_1\boldsymbol{X}_2$. This allows us to run the regression

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\alpha}_2 + \boldsymbol{u}\\ &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \big(\boldsymbol{X}_2 - \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\big)\boldsymbol{\alpha}_2 + \boldsymbol{u}.\end{aligned}$$

The first line above is a regression model with two groups of regressors, $\boldsymbol{X}_1$ and $\boldsymbol{M}_1\boldsymbol{X}_2$, which are mutually orthogonal. Therefore, $\hat{\boldsymbol{\alpha}}_2$ will be unchanged if we omit $\boldsymbol{X}_1$. The second line makes it clear that this regression is a special case of (2.38), which implies that $\hat{\boldsymbol{\alpha}}_2$ is equal to $\hat{\boldsymbol{\beta}}_2$ from (2.34). Consequently, we see that the two regressions

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u} \quad \text{and} \tag{2.39}$$

$$\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{v} \tag{2.40}$$

must yield the same estimates of $\boldsymbol{\beta}_2$.

Although regressions (2.34) and (2.40) give the same estimates of $\boldsymbol{\beta}_2$, they do not give the same residuals, as we have indicated by writing $\boldsymbol{u}$ for one regression and $\boldsymbol{v}$ for the other. We can see why the residuals are not the same by looking again at Figure 2.13, in which the constant $\boldsymbol{\iota}$ plays the role of $\boldsymbol{X}_1$, and the centered variable $\boldsymbol{z}$ plays the role of $\boldsymbol{M}_1\boldsymbol{X}_2$. The point corresponding to $\boldsymbol{y}$ can be thought of as lying somewhere on a line through the point $\hat{\boldsymbol{y}}$ and sticking perpendicularly out from the page. The residual vector from regressing $\boldsymbol{y}$ on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$ is thus represented by the line segment from $\hat{\boldsymbol{y}}$, in the page, to $\boldsymbol{y}$, vertically above the page. However, if $\boldsymbol{y}$ is regressed on $\boldsymbol{z}$ alone, the residual vector is the sum of this line segment and the segment from $\hat{\alpha}_2\boldsymbol{z}$ and $\hat{\boldsymbol{y}}$, that is, the top side of the rectangle in the figure. If we want the same residuals in regression (2.34) and a regression like (2.40), we need to purge the dependent variable of the second segment, which can be seen from the figure to be equal to $\hat{\alpha}_1\boldsymbol{\iota}$.

This suggests replacing $\boldsymbol{y}$ by what we get by projecting $\boldsymbol{y}$ off $\boldsymbol{\iota}$. This projection would be the line segment perpendicular to the page, translated in the horizontal direction so that it intersected the page at the point $\hat{\alpha}_2\boldsymbol{z}$ rather than $\hat{\boldsymbol{y}}$. In the general context, the analogous operation replaces $\boldsymbol{y}$ by $\boldsymbol{M}_1\boldsymbol{y}$, the projection off $\boldsymbol{X}_1$ rather than off $\boldsymbol{\iota}$. When we perform this projection, (2.40) is replaced by the regression

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}, \tag{2.41}$$

which will yield the same vector of OLS estimates $\hat{\boldsymbol{\beta}}_2$ as (2.34), and also the same vector of residuals. This regression is sometimes called the **FWL regression**. We used the notation "+ residuals" instead of "+ $\boldsymbol{u}$" in (2.41) because, in general, the difference between $\boldsymbol{M}_1\boldsymbol{y}$ and $\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$ is not the same thing as the vector $\boldsymbol{u}$ in (2.34). If $\boldsymbol{u}$ is interpreted as an error vector, then (2.41) would not be true if "residuals" were replaced by $\boldsymbol{u}$.

We can now formally state the FWL Theorem. Although the conclusions of the theorem have been established gradually in this section, we also provide a short formal proof.

Theorem 2.1. (Frisch-Waugh-Lovell Theorem)

1. The OLS estimates of $\boldsymbol{\beta}_2$ from regressions (2.34) and (2.41) are numerically identical.
2. The residuals from regressions (2.34) and (2.41) are numerically identical.

Proof: By the standard formula (1.46), the estimate of $\boldsymbol{\beta}_2$ from (2.41) is

$$(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}. \tag{2.42}$$

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ denote the two vectors of OLS estimates from (2.34). Then

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}. \tag{2.43}$$

Premultiplying the leftmost and rightmost expressions in (2.43) by $\boldsymbol{X}_2^\top\boldsymbol{M}_1$, we obtain

$$\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2. \tag{2.44}$$

The first term on the right-hand side of (2.43) has dropped out because $\boldsymbol{M}_1$ annihilates $\boldsymbol{X}_1$. To see that the last term also drops out, observe that

$$\boldsymbol{M_X M}_1\boldsymbol{X}_2 = \boldsymbol{M_X X}_2 = \mathbf{O}. \tag{2.45}$$

The first equality follows from (2.36) (see also Exercise 2.14), and the second from (2.24), which shows that $\boldsymbol{M_X}$ annihilates all the columns of $\boldsymbol{X}$, in particular those of $\boldsymbol{X}_2$. Premultiplying $\boldsymbol{y}$ by the transpose of (2.45) shows that $\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{M_X y} = \mathbf{0}$. We can now solve (2.44) for $\hat{\boldsymbol{\beta}}_2$ to obtain

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y},$$

which is expression (2.42). This proves the first part of the theorem.

If we had premultiplied (2.43) by $\boldsymbol{M}_1$ instead of by $\boldsymbol{X}_2^\top\boldsymbol{M}_1$, we would have obtained

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M_X y}, \tag{2.46}$$

where the last term is unchanged from (2.43) because $\boldsymbol{M}_1\boldsymbol{M_X} = \boldsymbol{M_X}$. The regressand in (2.46) is the regressand from regression (2.41). Because $\hat{\boldsymbol{\beta}}_2$ is the estimate of $\boldsymbol{\beta}_2$ from (2.41), by the first part of the theorem, the first term on the right-hand side of (2.46) is the vector of fitted values from that regression. Thus the second term must be the vector of residuals from regression (2.41). But $\boldsymbol{M_X y}$ is also the vector of residuals from regression (2.34), and this therefore proves the second part of the theorem. ■

2.5 Applications of the FWL Theorem

A regression like (2.34), in which the regressors are broken up into two groups, can arise in many situations. In this section, we will study three of these. The first two, seasonal dummy variables and time trends, are obvious applications of the FWL Theorem. The third, measures of goodness of fit that take the constant term into account, is somewhat less obvious. In all cases, the FWL Theorem allows us to obtain explicit expressions based on (2.42) for subsets of the parameter estimates of a linear regression.

Seasonal Dummy Variables

For a variety of reasons, it is sometimes desirable to include among the explanatory variables of a regression model variables that can take on only two possible values, which are usually 0 and 1. Such variables are called **indicator variables**, because they indicate a subset of the observations, namely, those for which the value of the variable is 1. Indicator variables are a special case of **dummy variables**, which can take on more than two possible values.

Seasonal variation provides a good reason to employ dummy variables. It is common for economic data that are indexed by time to take the form of

quarterly data, where each year in the sample period is represented by four observations, one for each quarter, or season, of the year. Many economic activities are strongly affected by the season, for obvious reasons like Christmas shopping, or summer holidays, or the difficulty of doing outdoor work during very cold weather. This seasonal variation, or **seasonality**, in economic activity is likely to be reflected in the economic **time series** that are used in regression models. The term "time series" is used to refer to any variable the observations of which are indexed by the time. Of course, time-series data are sometimes annual, in which case there is no seasonal variation to worry about, and sometimes monthly, in which case there are twelve "seasons" instead of four. For simplicity, we consider only the case of quarterly data.

Since there are four seasons, there may be four **seasonal dummy variables**, each taking the value 1 for just one of the four seasons. Let us denote these variables as $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, $\boldsymbol{s}_3$, and $\boldsymbol{s}_4$. If we consider a sample the first observation of which corresponds to the first quarter of some year, these variables look like

$$\boldsymbol{s}_1 = \begin{bmatrix} 1\\0\\0\\0\\1\\0\\0\\0\\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_2 = \begin{bmatrix} 0\\1\\0\\0\\0\\1\\0\\0\\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_3 = \begin{bmatrix} 0\\0\\1\\0\\0\\0\\1\\0\\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_4 = \begin{bmatrix} 0\\0\\0\\1\\0\\0\\0\\1\\ \vdots \end{bmatrix}. \tag{2.47}$$

An important property of these variables is that, since every observation must correspond to some season, the sum of the seasonal dummies must indicate every season. This means that this sum is a vector every component of which equals 1. Algebraically,

$$\boldsymbol{s}_1 + \boldsymbol{s}_2 + \boldsymbol{s}_3 + \boldsymbol{s}_4 = \boldsymbol{\iota}, \tag{2.48}$$

as is clear from (2.47). Since $\boldsymbol{\iota}$ represents the constant in a regression, (2.48) means that the five-variable set consisting of all four seasonal dummies plus the constant is linearly dependent. Consequently, one of the five variables must be dropped if all the regressors are to be linearly independent.

Just which one of the five is dropped makes no difference to the fitted values and residuals of a regression, because it is easy to check that

$$\mathcal{S}(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4) = \mathcal{S}(\boldsymbol{\iota}, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4) = \mathcal{S}(\boldsymbol{\iota}, \boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_4),$$

and so on. However the parameter estimates associated with the set of four variables that we choose to keep have different interpretations depending on that choice. Suppose first that we drop the constant and run the regression

$$\boldsymbol{y} = \alpha_1\boldsymbol{s}_1 + \alpha_2\boldsymbol{s}_2 + \alpha_3\boldsymbol{s}_3 + \alpha_4\boldsymbol{s}_4 + \boldsymbol{X\beta} + \boldsymbol{u}, \tag{2.49}$$

where the $n \times k$ matrix $\boldsymbol{X}$ contains other explanatory variables. Consider a single observation, indexed by t, that corresponds to the first season. The t^{th} observations of $\boldsymbol{s}_2$, $\boldsymbol{s}_3$, and $\boldsymbol{s}_4$ are all 0, and that of $\boldsymbol{s}_1$ is 1. Thus, if we write out the t^{th} observation of (2.49), we get

$$y_t = \alpha_1 + \boldsymbol{X}_t\boldsymbol{\beta} + u_t.$$

From this it is clear that, for all t belonging to the first season, the constant term in the regression is α_1. If we repeat this exercise for t in the second, third, or fourth season, we see at once that α_i is the constant for season i. Thus the introduction of the seasonal dummies gives us a different constant for every season.

An alternative is to retain the constant and drop $\boldsymbol{s}_1$. This yields

$$\boldsymbol{y} = \alpha_0\boldsymbol{\iota} + \gamma_2\boldsymbol{s}_2 + \gamma_3\boldsymbol{s}_3 + \gamma_4\boldsymbol{s}_4 + \boldsymbol{X\beta} + \boldsymbol{u}.$$

It is clear that, in this specification, the overall constant α_0 is really the constant for season 1. For an observation belonging to season 2, the constant is $\alpha_0 + \gamma_2$, for an observation belonging to season 3, it is $\alpha_0 + \gamma_3$, and so on. The easiest way to interpret this is to think of season 1 as the reference season. The coefficients γ_i, $i = 2, 3, 4$, measure the difference between α_0, the constant for the reference season, and the constant for season i. Since we could have dropped any of the seasonal dummies, the reference season is, of course, entirely arbitrary.

Another alternative is to retain the constant and use the three dummy variables defined by

$$\boldsymbol{s}_1' = \boldsymbol{s}_1 - \boldsymbol{s}_4, \quad \boldsymbol{s}_2' = \boldsymbol{s}_2 - \boldsymbol{s}_4, \quad \boldsymbol{s}_3' = \boldsymbol{s}_3 - \boldsymbol{s}_4. \tag{2.50}$$

These new dummy variables are not actually indicator variables, because their components for season 4 are equal to -1, but they have the advantage that, for each complete year, the sum of their components for that year is 0. Thus, for any sample whose size is a multiple of 4, each of the $\boldsymbol{s}_i'$, $i = 1, 2, 3$, is orthogonal to the constant. We can write the regression as

$$\boldsymbol{y} = \delta_0\boldsymbol{\iota} + \delta_1\boldsymbol{s}_1' + \delta_2\boldsymbol{s}_2' + \delta_3\boldsymbol{s}_3' + \boldsymbol{X\beta} + \boldsymbol{u}. \tag{2.51}$$

It is easy to see that, for t in season i, $i = 1, 2, 3$, the constant term is $\delta_0 + \delta_i$. For t belonging to season 4, it is $\delta_0 - \delta_1 - \delta_2 - \delta_3$. Thus the average of the constants for all four seasons is just δ_0, the coefficient of the constant, $\boldsymbol{\iota}$. Accordingly, the δ_i, $i = 1, 2, 3$, measure the difference between the average constant δ_0 and the constant specific to season i. Season 4 is a bit of a mess, because of the arithmetic needed to ensure that the average does indeed work out to δ_0.

Let $\boldsymbol{S}$ denote whatever $n \times 4$ matrix we choose to use in order to span the constant and the four seasonal variables $\boldsymbol{s}_i$. Then any of the regressions we have considered so far can be written as

$$\boldsymbol{y} = \boldsymbol{S\delta} + \boldsymbol{X\beta} + \boldsymbol{u}. \tag{2.52}$$

This regression has two groups of regressors, as required for the application of the FWL Theorem. That theorem implies that the estimates $\hat{\boldsymbol{\beta}}$ and the residuals $\hat{\boldsymbol{u}}$ can also be obtained by running the FWL regression

$$\boldsymbol{M_S y} = \boldsymbol{M_S X\beta} + \text{residuals}, \tag{2.53}$$

where, as the notation suggests, $\boldsymbol{M_S} \equiv \mathbf{I} - \boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{S})^{-1}\boldsymbol{S}^\top$.

The effect of the projection $\boldsymbol{M_S}$ on $\boldsymbol{y}$ and on the explanatory variables in the matrix $\boldsymbol{X}$ can be considered as a form of **seasonal adjustment**. By making $\boldsymbol{M_S y}$ orthogonal to all the seasonal variables, we are, in effect, purging it of its seasonal variation. Consequently, $\boldsymbol{M_S y}$ can be called a **seasonally adjusted**, or **deseasonalized**, version of $\boldsymbol{y}$, and similarly for the explanatory variables. In practice, such seasonally adjusted variables can be conveniently obtained as the residuals from regressing $\boldsymbol{y}$ and each of the columns of $\boldsymbol{X}$ on the variables in $\boldsymbol{S}$. The FWL Theorem tells us that we get the same results in terms of estimates of $\boldsymbol{\beta}$ and residuals whether we run (2.52), in which the variables are unadjusted and seasonality is explicitly accounted for, or run (2.53), in which all the variables are seasonally adjusted by regression. This was, in fact, the subject of the famous paper by Lovell (1963).

The equivalence of (2.52) and (2.53) is sometimes used to claim that, in estimating a regression model with time-series data, it does not matter whether one uses "raw" data, along with seasonal dummies, or seasonally adjusted data. Such a conclusion is completely unwarranted. Official seasonal adjustment procedures are almost never based on regression; using official seasonally adjusted data is therefore *not* equivalent to using residuals from regression on a set of seasonal variables. Moreover, if (2.52) is not a sensible model (and it would not be if, for example, the seasonal pattern were more complicated than that given by $\boldsymbol{S\alpha}$), then (2.53) is not a sensible specification either. Seasonality is actually an important practical problem in applied work with time-series data. We will discuss it further in Chapter 13. For more detailed treatments, see Hylleberg (1986, 1992) and Ghysels and Osborn (2001).

The deseasonalization performed by the projection $\boldsymbol{M_S}$ makes all variables orthogonal to the constant as well as to the seasonal dummies. Thus the effect of $\boldsymbol{M_S}$ is not only to deseasonalize, but also to center, the variables on which it acts. Sometimes this is undesirable; if so, we may use the three variables $\boldsymbol{s}'_i$ given in (2.50). Since they are themselves orthogonal to the constant, no centering takes place if only these three variables are used for seasonal adjustment. An explicit constant should normally be included in any regression that uses variables seasonally adjusted in this way.

Time Trends

Another sort of constructed, or artificial, variable that is often encountered in models of time-series data is a **time trend**. The simplest sort of time trend is the **linear time trend**, represented by the vector $\boldsymbol{T}$, with typical element $T_t \equiv t$. Thus $\boldsymbol{T} = [1 \,\vdots\, 2 \,\vdots\, 3 \,\vdots\, 4 \,\vdots\, \ldots]$. Imagine that we have a regression with a constant and a linear time trend:

$$\boldsymbol{y} = \gamma_1 \boldsymbol{\iota} + \gamma_2 \boldsymbol{T} + \boldsymbol{X\beta} + \boldsymbol{u}.$$

For observation t, y_t is equal to $\gamma_1 + \gamma_2 t + \boldsymbol{X}_t \boldsymbol{\beta} + u_t$. Thus the overall level of y_t increases or decreases steadily as t increases. Instead of just a constant, we now have the linear (strictly speaking, affine) function of time, $\gamma_1 + \gamma_2 t$. An increasing time trend might be appropriate, for instance, in a model of a production function where technical progress is taking place. An explicit model of technical progress might well be difficult to construct, in which case a linear time trend could serve as a simple way to take account of the phenomenon.

It is often desirable to make the time trend orthogonal to the constant by centering it, that is, operating on it with $\boldsymbol{M_\iota}$. If we do this with a sample with an odd number of elements, the result is a variable that looks like

$$[\,\cdots \,\vdots\, -3 \,\vdots\, -2 \,\vdots\, -1 \,\vdots\, 0 \,\vdots\, 1 \,\vdots\, 2 \,\vdots\, 3 \,\vdots\, \cdots\,].$$

If the sample size is even, the variable is made up of the half integers $\pm{}^1\!/_2$, $\pm{}^3\!/_2$, $\pm{}^5\!/_2, \ldots$. In both cases, the coefficient of $\boldsymbol{\iota}$ is the average value of the linear function of time over the whole sample.

Sometimes it is appropriate to use constructed variables that are more complicated than a linear time trend. A simple case would be a quadratic time trend, with typical element t^2. In fact, any deterministic function of the time index t can be used, including the trigonometric functions $\sin t$ and $\cos t$, which could be used to account for oscillatory behavior. With such variables, it is again usually preferable to make them orthogonal to the constant by centering them.

The FWL Theorem applies just as well with time trends of various sorts as it does with seasonal dummy variables. It is possible to project all the other variables in a regression model off the time trend variables, thereby obtaining **detrended** variables. The parameter estimates and residuals will be same as if the trend variables were explicitly included in the regression. This was in fact the type of situation dealt with by Frisch and Waugh (1933).

Goodness of Fit of a Regression

In equations (2.18) and (2.19), we showed that the total sum of squares (TSS) in the regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ can be expressed as the sum of the explained sum of squares (ESS) and the sum of squared residuals (SSR).

This was really just an application of Pythagoras' Theorem. In terms of the orthogonal projection matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$, the relation between TSS, ESS, and SSR can be written as

$$\text{TSS} = \|\boldsymbol{y}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2 = \text{ESS} + \text{SSR}.$$

This allows us to introduce a measure of **goodness of fit** for a regression model. This measure is formally called the **coefficient of determination**, but it is universally referred to as the $\boldsymbol{R^2}$. The R^2 is simply the ratio of ESS to TSS. It can be written as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\|\boldsymbol{P_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\boldsymbol{M_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\text{SSR}}{\text{TSS}} = \cos^2\theta, \tag{2.54}$$

where θ is the angle between $\boldsymbol{y}$ and $\boldsymbol{P_X y}$; see Figure 2.10. For any angle θ, we know that $-1 \leq \cos\theta \leq 1$. Consequently, $0 \leq R^2 \leq 1$. If the angle θ were zero, $\boldsymbol{y}$ and $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ would coincide, the residual vector $\hat{\boldsymbol{u}}$ would vanish, and we would have what is called a **perfect fit**, with $R^2 = 1$. At the other extreme, if $R^2 = 0$, the fitted value vector would vanish, and $\boldsymbol{y}$ would coincide with the residual vector $\hat{\boldsymbol{u}}$.

As we will see shortly, (2.54) is not the only measure of goodness of fit. It is known as the **uncentered** R^2, and, to distinguish it from other versions of R^2, it is sometimes denoted as R^2_u. Because R^2_u depends on $\boldsymbol{y}$ only through the residuals and fitted values, it is invariant under nonsingular linear transformations of the regressors. In addition, because it is defined as a ratio, the value of R^2_u is invariant to changes in the scale of $\boldsymbol{y}$. For example, we could change the units in which the regressand is measured from dollars to thousands of dollars without affecting the value of R^2_u.

However, R^2_u is not invariant to changes of units that change the angle θ. An example of such a change is given by the conversion between the Celsius and Fahrenheit scales of temperature, where a constant is involved; see (2.29). To see this, let us consider a very simple change of measuring units, whereby a constant α, analogous to the constant 32 used in converting from Celsius to Fahrenheit, is added to each element of $\boldsymbol{y}$. In terms of these new units, the regression of $\boldsymbol{y}$ on a regressor matrix $\boldsymbol{X}$ becomes

$$\boldsymbol{y} + \alpha\boldsymbol{\iota} = \boldsymbol{X\beta} + \boldsymbol{u}. \tag{2.55}$$

If we assume that the matrix $\boldsymbol{X}$ includes a constant, it follows that $\boldsymbol{P_X\iota} = \boldsymbol{\iota}$ and $\boldsymbol{M_X\iota} = \boldsymbol{0}$, and so we find that

$$\boldsymbol{y} + \alpha\boldsymbol{\iota} = \boldsymbol{P_X}(\boldsymbol{y} + \alpha\boldsymbol{\iota}) + \boldsymbol{M_X}(\boldsymbol{y} + \alpha\boldsymbol{\iota}) = \boldsymbol{P_X y} + \alpha\boldsymbol{\iota} + \boldsymbol{M_X y}.$$

This allows us to compute R^2_u as

$$R^2_u = \frac{\|\boldsymbol{P_X y} + \alpha\boldsymbol{\iota}\|^2}{\|\boldsymbol{y} + \alpha\boldsymbol{\iota}\|^2},$$

which is clearly different from (2.54). By choosing α sufficiently large, we can in fact make R_u^2 as close as we wish to 1, because, for very large α, the term $\alpha\boldsymbol{\iota}$ will completely dominate the terms $\boldsymbol{P_X y}$ and $\boldsymbol{y}$ in the numerator and denominator respectively. But a large R_u^2 in such a case would be entirely misleading, since the "good fit" would be accounted for almost exclusively by the constant.

It is easy to see how to get around this problem, at least for regressions that include a constant term. An elementary consequence of the FWL Theorem is that we can express all variables as deviations from their means, by the operation of the projection $\boldsymbol{M_\iota}$, without changing parameter estimates or residuals. The ordinary R^2 from the regression that uses centered variables is called the **centered** R^2. It is defined as

$$R_c^2 \equiv \frac{\|\boldsymbol{P_X M_\iota y}\|^2}{\|\boldsymbol{M_\iota y}\|^2} = 1 - \frac{\|\boldsymbol{M_X y}\|^2}{\|\boldsymbol{M_\iota y}\|^2}, \tag{2.56}$$

and it is clearly unaffected by the addition of a constant to the regressand, as in equation (2.55).

The centered R^2 is much more widely used than the uncentered R^2. When $\boldsymbol{\iota}$ is contained in the span $\mathcal{S}(\boldsymbol{X})$ of the regressors, R_c^2 certainly makes far more sense than R_u^2. However, R_c^2 does not make sense for regressions without a constant term or its equivalent in terms of dummy variables. If a statistical package reports a value for R^2 in such a regression, one needs to be very careful. Different ways of computing R_c^2, all of which would yield the same, correct, answer for regressions that include a constant, may yield quite different answers for regressions that do not. It is even possible to obtain values of R_c^2 that are less than 0 or greater than 1, depending on how the calculations are carried out.

Either version of R^2 is a valid measure of goodness of fit only when the least squares estimates $\hat{\boldsymbol{\beta}}$ are used. If we used some other estimates of $\boldsymbol{\beta}$, say $\tilde{\boldsymbol{\beta}}$, the triangle in Figure 2.10 would no longer be a right-angled triangle, and Pythagoras' Theorem would no longer apply. As a consequence, (2.54) would no longer hold, and the different definitions of R^2 would no longer be the same:

$$1 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y}\|^2} \neq \frac{\|\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y}\|^2}.$$

If we chose to define R^2 in terms of the residuals, using the first of these expressions, we could not guarantee that it would be positive, and if we chose to define it in terms of the fitted values, using the second, we could not guarantee that it would be less than 1. Thus, when anything other than least squares is used to estimate a regression, one should be very cautious about interpreting a reported R^2. It is not a sensible measure of fit in such a case, and, depending on how it is actually computed, it may be seriously misleading.

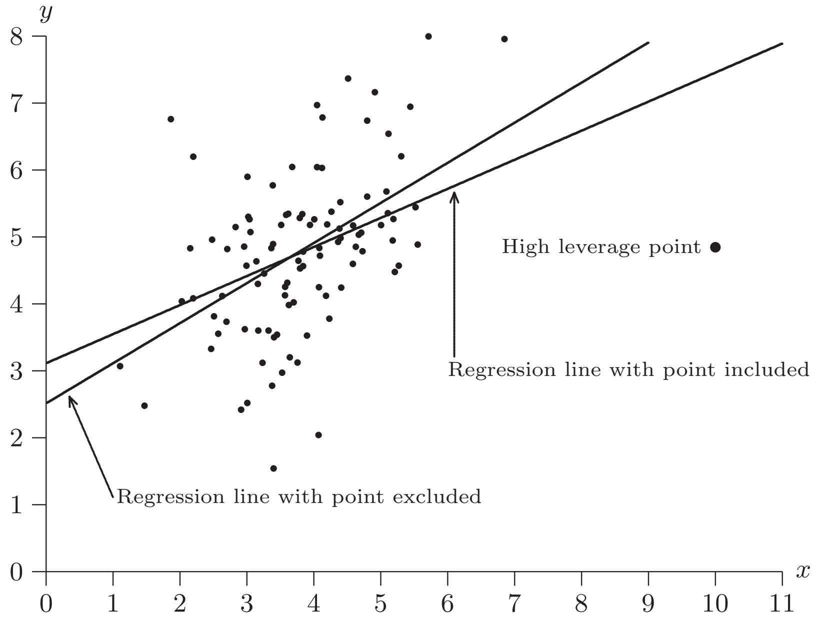


Figure 2.14 An influential observation

2.6 Influential Observations and Leverage

One important feature of OLS estimation, which we have not stressed up to this point, is that each element of the vector of parameter estimates $\hat{\boldsymbol{\beta}}$ is simply a weighted average of the elements of the vector $\boldsymbol{y}$. To see this, define $\boldsymbol{c}_i$ as the i^{th} row of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and observe from (2.02) that $\hat{\beta}_i = \boldsymbol{c}_i\boldsymbol{y}$. This fact will prove to be of great importance when we discuss the statistical properties of least squares estimation in the next chapter.

Because each element of $\hat{\boldsymbol{\beta}}$ is a weighted average, some observations may affect the value of $\hat{\boldsymbol{\beta}}$ much more than others do. Consider Figure 2.14. This figure is an example of a **scatter diagram**, a long-established way of graphing the relation between two variables. Each point in the figure has Cartesian coordinates (x_t, y_t), where x_t is a typical element of a vector $\boldsymbol{x}$, and y_t of a vector $\boldsymbol{y}$. One point, drawn with a larger dot than the rest, is indicated, for reasons to be explained, as a high leverage point. Suppose that we run the regression

$$\boldsymbol{y} = \beta_1\boldsymbol{\iota} + \beta_2\boldsymbol{x} + \boldsymbol{u}$$

twice, once with, and once without, the high leverage observation. For each regression, the fitted values all lie on the so-called **regression line**, which is the straight line with equation

$$y = \hat{\beta}_1 + \hat{\beta}_2 x.$$

The slope of this line is just $\hat{\beta}_2$, which is why β_2 is sometimes called the **slope coefficient**; see Section 1.1. Similarly, because $\hat{\beta}_1$ is the intercept that the regression line makes with the y axis, the constant term β_1 is sometimes called the **intercept**. The regression line is entirely determined by the estimated coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$.

The regression lines for the two regressions in Figure 2.14 are substantially different. The high leverage point is quite distant from the regression line obtained when it is excluded. When that point is included, it is able, by virtue of its position well to the right of the other observations, to exert a good deal of **leverage** on the regression line, pulling it down toward itself. If the y coordinate of this point were greater, making the point closer to the regression line excluding it, then it would have a smaller **influence** on the regression line including it. If the x coordinate were smaller, putting the point back into the main cloud of points, again there would be a much smaller influence. Thus it is the x coordinate that gives the point its position of high leverage, but it is the y coordinate that determines whether the high leverage position will actually be exploited, resulting in substantial influence on the regression line. In a moment, we will generalize these conclusions to regressions with any number of regressors.

If one or a few observations in a regression are highly **influential**, in the sense that deleting them from the sample would change some elements of $\hat{\boldsymbol{\beta}}$ substantially, the prudent econometrician will normally want to scrutinize the data carefully. It may be that these **influential observations** are erroneous, or at least untypical of the rest of the sample. Since a single erroneous observation can have an enormous effect on $\hat{\boldsymbol{\beta}}$, it is important to ensure that any influential observations are not in error. Even if the data are all correct, the interpretation of the regression results may change if it is known that a few observations are primarily responsible for them, especially if those observations differ systematically in some way from the rest of the data.

Leverage

The effect of a single observation on $\hat{\boldsymbol{\beta}}$ can be seen by comparing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}^{(t)}$, the estimate of $\boldsymbol{\beta}$ that would be obtained if the t^{th} observation were omitted from the sample. Rather than actually omit the t^{th} observation, it is easier to remove its effect by using a dummy variable. The appropriate dummy variable is $\boldsymbol{e}_t$, an n–vector which has t^{th} element 1 and all other elements 0. The vector $\boldsymbol{e}_t$ is called a **unit basis vector**, unit because its norm is 1, basis because the set of all the $\boldsymbol{e}_t$, for $t = 1, \ldots, n$, span, or constitute a **basis** for, the full space E^n; see Exercise 2.20. Considered as an indicator variable, $\boldsymbol{e}_t$ indexes the singleton subsample that contains only observation t.

Including $\boldsymbol{e}_t$ as a regressor leads to a regression of the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \alpha\boldsymbol{e}_t + \boldsymbol{u}, \tag{2.57}$$

and, by the FWL Theorem, this gives the same parameter estimates and residuals as the FWL regression

$$\boldsymbol{M}_t\boldsymbol{y} = \boldsymbol{M}_t\boldsymbol{X\beta} + \text{residuals}, \tag{2.58}$$

where $\boldsymbol{M}_t \equiv \boldsymbol{M}_{\boldsymbol{e}_t} = \mathbf{I} - \boldsymbol{e}_t(\boldsymbol{e}_t^\top\boldsymbol{e}_t)^{-1}\boldsymbol{e}_t^\top$ is the orthogonal projection off the vector $\boldsymbol{e}_t$. It is easy to see that $\boldsymbol{M}_t\boldsymbol{y}$ is just $\boldsymbol{y}$ with its $t^{\rm th}$ component replaced by 0. Since $\boldsymbol{e}_t^\top\boldsymbol{e}_t = 1$, and since $\boldsymbol{e}_t^\top\boldsymbol{y}$ can easily be seen to be the $t^{\rm th}$ component of $\boldsymbol{y}$,

$$\boldsymbol{M}_t\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{e}_t\boldsymbol{e}_t^\top\boldsymbol{y} = \boldsymbol{y} - y_t\boldsymbol{e}_t.$$

Thus y_t is subtracted from $\boldsymbol{y}$ for the $t^{\rm th}$ observation only. Similarly, $\boldsymbol{M}_t\boldsymbol{X}$ is just $\boldsymbol{X}$ with its $t^{\rm th}$ row replaced by zeros. Running regression (2.58) will give the same parameter estimates as those that would be obtained if we deleted observation t from the sample. Since the vector $\hat{\boldsymbol{\beta}}$ is defined exclusively in terms of scalar products of the variables, replacing the $t^{\rm th}$ elements of these variables by 0 is tantamount to simply leaving observation t out when computing those scalar products.

Let us denote by $\boldsymbol{P}_{\boldsymbol{Z}}$ and $\boldsymbol{M}_{\boldsymbol{Z}}$, respectively, the orthogonal projections on to and off $\mathcal{S}(\boldsymbol{X}, \boldsymbol{e}_t)$. The fitted values and residuals from regression (2.57) are then given by

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{e}_t + \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}. \tag{2.59}$$

Now premultiply (2.59) by $\boldsymbol{P}_{\boldsymbol{X}}$ to obtain

$$\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t, \tag{2.60}$$

where we have used the fact that $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{P}_{\boldsymbol{X}} = \mathbf{O}$, because $\boldsymbol{M}_{\boldsymbol{Z}}$ annihilates both $\boldsymbol{X}$ and $\boldsymbol{e}_t$. But $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and so (2.60) gives

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}) = -\hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t. \tag{2.61}$$

We can compute the difference between $\hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\beta}}$ from this if we can compute the value of $\hat{\alpha}$.

In order to calculate $\hat{\alpha}$, we once again use the FWL Theorem, which tells us that the estimate of α from (2.57) is the same as the estimate from the FWL regression

$$\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \hat{\alpha}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{e}_t + \text{residuals}.$$

Therefore, using (2.02) and the idempotency of $\boldsymbol{M}_{\boldsymbol{X}}$,

$$\hat{\alpha} = \frac{\boldsymbol{e}_t^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{\boldsymbol{e}_t^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{e}_t}. \tag{2.62}$$

Now $\boldsymbol{e}_t^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ is the $t^{\rm th}$ element of $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$, the vector of residuals from the regression including all observations. We may denote this element as $\hat{u}_t$. In like manner, $\boldsymbol{e}_t^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{e}_t$, which is just a scalar, is the $t^{\rm th}$ diagonal element of $\boldsymbol{M}_{\boldsymbol{X}}$. Substituting these into (2.62), we obtain

$$\hat{\alpha} = \frac{\hat{u}_t}{1 - h_t}, \tag{2.63}$$

where h_t denotes the $t^{\rm th}$ diagonal element of $\boldsymbol{P_X}$, which is equal to 1 minus the $t^{\rm th}$ diagonal element of $\boldsymbol{M_X}$. The rather odd notation h_t comes from the fact that $\boldsymbol{P_X}$ is sometimes referred to as the **hat matrix**, because the vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P_X y}$ is sometimes written as $\hat{\boldsymbol{y}}$, and $\boldsymbol{P_X}$ is therefore said to "put a hat on" $\boldsymbol{y}$.

Finally, if we premultiply (2.61) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and use (2.63), we find that

$$\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}} = -\hat{\alpha}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_X e}_t = \frac{-1}{1-h_t}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_t^\top\hat{u}_t. \tag{2.64}$$

The second equality uses the facts that $\boldsymbol{X}^\top\boldsymbol{P_X} = \boldsymbol{X}^\top$ and that the final factor of $\boldsymbol{e}_t$ selects the $t^{\rm th}$ column of $\boldsymbol{X}^\top$, which is the transpose of the $t^{\rm th}$ row, $\boldsymbol{X}_t$. Expression (2.64) makes it clear that, when either $\hat{u}_t$ is large or h_t is large, or both, the effect of the $t^{\rm th}$ observation on at least some elements of $\hat{\boldsymbol{\beta}}$ is likely to be substantial. Such an observation is said to be influential.

From (2.64), it is evident that the influence of an observation depends on both $\hat{u}_t$ and h_t. It will be greater if the observation has a large residual, which, as we saw in Figure 2.14, is related to its y coordinate. On the other hand, h_t is related to the x coordinate of a point, which, as we also saw in the figure, determines the leverage, or potential influence, of the corresponding observation. We say that observations for which h_t is large have **high leverage** or are **leverage points**. A leverage point is not necessarily influential, but it has the potential to be influential.

The Diagonal Elements of the Hat Matrix

Since the leverage of the $t^{\rm th}$ observation depends on h_t, the $t^{\rm th}$ diagonal element of the hat matrix, it is worth studying the properties of these diagonal elements in a little more detail. We can express h_t as

$$h_t = \boldsymbol{e}_t^\top\boldsymbol{P_X e}_t = \|\boldsymbol{P_X e}_t\|^2. \tag{2.65}$$

Since the rightmost expression here is a square, $h_t \geq 0$. Moreover, since $\|\boldsymbol{e}_t\| = 1$, we obtain from (2.28) applied to $\boldsymbol{e}_t$ that $h_t = \|\boldsymbol{P_X e}_t\|^2 \leq 1$. Thus

$$0 \leq h_t \leq 1. \tag{2.66}$$

The geometrical reason for these bounds on the value of h_t can be found in Exercise 2.26.

The lower bound in (2.66) can be strengthened when there is a constant term. In that case, none of the h_t can be less than $1/n$. This follows from (2.65), because if $\boldsymbol{X}$ consisted only of a constant vector $\boldsymbol{\iota}$, $\boldsymbol{e}_t^\top\boldsymbol{P_\iota e}_t$ would equal $1/n$. If other regressors are present, then we have

$$1/n = \|\boldsymbol{P_\iota e}_t\|^2 = \|\boldsymbol{P_\iota P_X e}_t\|^2 \leq \|\boldsymbol{P_X e}_t\|^2 = h_t.$$

Here we have used the fact that $\boldsymbol{P_\iota P_X} = \boldsymbol{P_\iota}$ since $\boldsymbol{\iota}$ is in $\mathcal{S}(\boldsymbol{X})$ by assumption, and, for the inequality, we have used (2.28). Although h_t cannot be 0 in normal circumstances, there is a special case in which it equals 1. If one column of $\boldsymbol{X}$ is the dummy variable $\boldsymbol{e}_t$, $h_t = \boldsymbol{e}_t^\top \boldsymbol{P}_X \boldsymbol{e}_t = \boldsymbol{e}_t^\top \boldsymbol{e}_t = 1$.

In a regression with n observations and k regressors, the average of the h_t is equal to k/n. In order to demonstrate this, we need to use some properties of the **trace** of a square matrix. If $\boldsymbol{A}$ is an $n \times n$ matrix, its trace, denoted $\mathrm{Tr}(\boldsymbol{A})$, is the sum of the elements on its principal diagonal. Thus

$$\mathrm{Tr}(\boldsymbol{A}) \equiv \sum_{i=1}^{n} A_{ii}.$$

A convenient property is that the trace of a product of two not necessarily square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is unaffected by the order in which the two matrices are multiplied together. If the dimensions of $\boldsymbol{A}$ are $n \times m$, then, in order for the product $\boldsymbol{AB}$ to be square, those of $\boldsymbol{B}$ must be $m \times n$. This implies further that the product $\boldsymbol{BA}$ exists and is $m \times m$. We have

$$\mathrm{Tr}(\boldsymbol{AB}) = \sum_{i=1}^{n} (\boldsymbol{AB})_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}B_{ji} = \sum_{j=1}^{m} (\boldsymbol{BA})_{jj} = \mathrm{Tr}(\boldsymbol{BA}). \tag{2.67}$$

The result (2.67) can be extended. If we consider a (square) product of several matrices, the trace is invariant under what is called a **cyclic permutation** of the factors. Thus, as can be seen by successive applications of (2.67),

$$\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{CAB}) = \mathrm{Tr}(\boldsymbol{BCA}). \tag{2.68}$$

We now return to the h_t. Their sum is

$$\begin{aligned}\sum_{t=1}^{n} h_t &= \mathrm{Tr}(\boldsymbol{P_X}) = \mathrm{Tr}\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) \\ &= \mathrm{Tr}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big) = \mathrm{Tr}(\mathbf{I}_k) = k.\end{aligned} \tag{2.69}$$

The first equality in the second line makes use of (2.68). Then, because we are multiplying a $k \times k$ matrix by its inverse, we get a $k \times k$ identity matrix, the trace of which is obviously just k. It follows from (2.69) that the average of the h_t equals k/n. When, for a given regressor matrix $\boldsymbol{X}$, the diagonal elements of $\boldsymbol{P_X}$ are all close to their average value, no observation has very much leverage. Such an $\boldsymbol{X}$ matrix is sometimes said to have a **balanced design**. On the other hand, if some of the h_t are much larger than k/n, and others consequently smaller, the $\boldsymbol{X}$ matrix is said to have an **unbalanced design**.

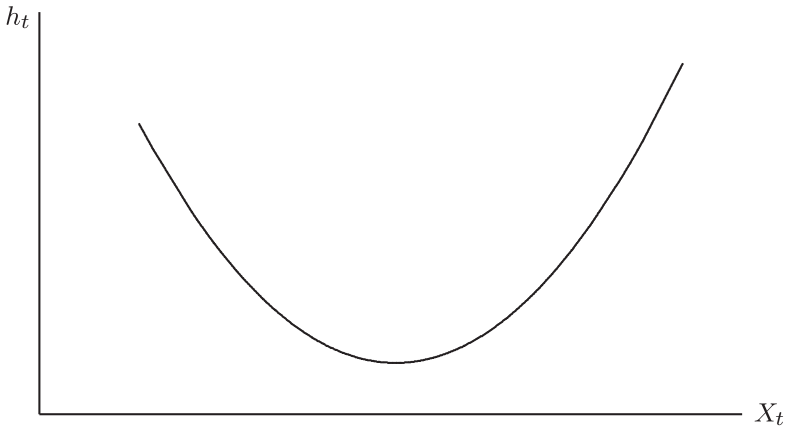


Figure 2.15 h_t as a function of X_t

The h_t tend to be larger for values of the regressors that are farther away from their average over the sample. As an example, Figure 2.15 plots them as a function of X_t for a particular sample of 100 observations for the model

$$y_t = \beta_1 + \beta_2 X_t + u_t.$$

The elements X_t of the regressor are perfectly well behaved, being drawings from the standard normal distribution. Although the average value of the h_t is $2/100 = 0.02$, h_t varies from 0.0100 for values of X_t near the sample mean to 0.0695 for the largest value of X_t, which is about 2.4 standard deviations above the sample mean. Thus, even in this very typical case, some observations have a great deal more leverage than others. Those observations with the greatest amount of leverage are those for which x_t is farthest from the sample mean, in accordance with the intuition of Figure 2.14.

2.7 Final Remarks

In this chapter, we have discussed the numerical properties of OLS estimation of linear regression models from a geometrical point of view. This perspective often provides a much simpler way to understand such models than does a purely algebraic approach. For example, the fact that certain matrices are idempotent becomes quite clear as soon as one understands the notion of an orthogonal projection. Most of the results discussed in this chapter are thoroughly fundamental, and many of them will be used again and again throughout the book. In particular, the FWL Theorem will turn out to be extremely useful in many contexts.

The use of geometry as an aid to the understanding of linear regression has a long history; see Herr (1980). One valuable reference on linear models that takes the geometric approach is Seber (1980). A good expository paper that is reasonably accessible is Bryant (1984), and a detailed treatment is provided by Ruud (2000).

It is strongly recommended that readers attempt the exercises which follow this chapter before starting Chapter 3, in which we turn our attention to the statistical properties of OLS estimation. Many of the results of this chapter will be useful in establishing these properties, and the exercises are designed to enhance understanding of these results.

2.8 Exercises

2.1 Consider two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in E^2. Let $\boldsymbol{x} = [x_1 \,\vdots\, x_2]$ and $\boldsymbol{y} = [y_1 \,\vdots\, y_2]$. Show trigonometrically that $\boldsymbol{x}^\top\boldsymbol{y} \equiv x_1y_1 + x_2y_2$ is equal to $\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|\cos\theta$, where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

2.2 A vector in E^n can be **normalized** by multiplying it by the reciprocal of its norm. Show that, for any $\boldsymbol{x} \in E^n$ with $\boldsymbol{x} \neq \boldsymbol{0}$, the norm of $\boldsymbol{x}/\|\boldsymbol{x}\|$ is 1.

Now consider two vectors $\boldsymbol{x}, \boldsymbol{y} \in E^n$. Compute the norm of the sum and of the difference of $\boldsymbol{x}$ normalized and $\boldsymbol{y}$ normalized, that is, of

$$\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} + \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \quad \text{and} \quad \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}.$$

By using the fact that the norm of any nonzero vector is positive, prove the Cauchy-Schwartz inequality (2.08):

$$|\boldsymbol{x}^\top\boldsymbol{y}| \leq \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|. \tag{2.08}$$

Show that this inequality becomes an equality when $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel. **Hint:** Show first that $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel if and only if $\boldsymbol{x}/\|\boldsymbol{x}\| = \pm\,\boldsymbol{y}/\|\boldsymbol{y}\|$.

2.3 The **triangle inequality** states that, for $\boldsymbol{x}, \boldsymbol{y} \in E^n$,

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. \tag{2.70}$$

Draw a 2–dimensional picture to illustrate this result. Prove the result algebraically by computing the squares of both sides of the above inequality, and then using (2.08). In what circumstances will (2.70) hold with equality?

2.4 Suppose that $\boldsymbol{x} = [1.0 \,\vdots\, 1.5 \,\vdots\, 1.2 \,\vdots\, 0.7]$ and $\boldsymbol{y} = [3.2 \,\vdots\, 4.4 \,\vdots\, 2.5 \,\vdots\, 2.0]$. What are $\|\boldsymbol{x}\|$, $\|\boldsymbol{y}\|$, and $\boldsymbol{x}^\top\boldsymbol{y}$? Use these quantities to calculate θ, the angle θ between $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\cos\theta$.

2.5 Show explicitly that the left-hand sides of (2.11) and (2.12) are the same. This can be done either by comparing typical elements or by using the results in Section 2.3 on partitioned matrices.

2.6 Prove that, if the k columns of $\boldsymbol{X}$ are linearly independent, each vector $\boldsymbol{z}$ in $\mathcal{S}(\boldsymbol{X})$ can be expressed as $\boldsymbol{X}\boldsymbol{b}$ for one and only one k–vector $\boldsymbol{b}$. **Hint:** Suppose that there are two different vectors, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, such that $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{b}_i$, $i = 1, 2$, and show that this implies that the columns of $\boldsymbol{X}$ are linearly dependent.

2.7 Consider the vectors $\boldsymbol{x}_1 = [1 \vdots 2 \vdots 4]$, $\boldsymbol{x}_2 = [2 \vdots 3 \vdots 5]$, and $\boldsymbol{x}_3 = [3 \vdots 6 \vdots 12]$. What is the dimension of the subspace that these vectors span?

2.8 Consider the example of the three vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ defined in (2.16). Show that any vector $\boldsymbol{z} \equiv b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$ in $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ also belongs to $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_3)$ and $\mathcal{S}(\boldsymbol{x}_2, \boldsymbol{x}_3)$. Give explicit formulas for $\boldsymbol{z}$ as a linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$, and of $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$.

2.9 Prove algebraically that $\boldsymbol{P_X M_X} = \mathbf{O}$. This is equation (2.26). Use only the requirement (2.25) that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ be complementary projections, and the idempotency of $\boldsymbol{P_X}$.

2.10 Prove algebraically that equation (2.27), which is really Pythagoras' Theorem for linear regression, holds. Use the facts that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric, idempotent, and orthogonal to each other.

2.11 Show algebraically that, if $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are complementary orthogonal projections, then $\boldsymbol{M_X}$ annihilates all vectors in $\mathcal{S}(\boldsymbol{X})$, and $\boldsymbol{P_X}$ annihilates all vectors in $\mathcal{S}^\perp(\boldsymbol{X})$.

2.12 Consider the two regressions

$$\begin{aligned}\boldsymbol{y} &= \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \beta_3\boldsymbol{x}_3 + \boldsymbol{u}, \text{ and}\\ \boldsymbol{y} &= \alpha_1\boldsymbol{z}_1 + \alpha_2\boldsymbol{z}_2 + \alpha_3\boldsymbol{z}_3 + \boldsymbol{u},\end{aligned}$$

where $\boldsymbol{z}_1 = \boldsymbol{x}_1 - 2\boldsymbol{x}_2$, $\boldsymbol{z}_2 = \boldsymbol{x}_2 + 4\boldsymbol{x}_3$, and $\boldsymbol{z}_3 = 2\boldsymbol{x}_1 - 3\boldsymbol{x}_2 + 5\boldsymbol{x}_3$. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \boldsymbol{x}_3]$ and $\boldsymbol{Z} = [\boldsymbol{z}_1 \;\; \boldsymbol{z}_2 \;\; \boldsymbol{z}_3]$. Show that the columns of $\boldsymbol{Z}$ can be expressed as linear combinations of the columns of $\boldsymbol{X}$, that is, that $\boldsymbol{Z} = \boldsymbol{XA}$, for some 3×3 matrix $\boldsymbol{A}$. Find the elements of this matrix $\boldsymbol{A}$.

Show that the matrix $\boldsymbol{A}$ is invertible, by showing that the columns of $\boldsymbol{X}$ are linear combinations of the columns of $\boldsymbol{Z}$. Give the elements of $\boldsymbol{A}^{-1}$. Show that the two regressions give the same fitted values and residuals.

Precisely how is the OLS estimate $\hat\beta_1$ related to the OLS estimates $\hat\alpha_i$, for $i = 1, \ldots, 3$? Precisely how is $\hat\alpha_1$ related to the $\hat\beta_i$, for $i = 1, \ldots, 3$?

2.13 Let $\boldsymbol{X}$ be an $n \times k$ matrix of full rank. Consider the $n \times k$ matrix $\boldsymbol{XA}$, where $\boldsymbol{A}$ is a *singular* $k \times k$ matrix. Show that the columns of $\boldsymbol{XA}$ are linearly dependent, and that $\mathcal{S}(\boldsymbol{XA}) \subset \mathcal{S}(\boldsymbol{X})$.

2.14 Use the result (2.36) to show that $\boldsymbol{M_X M}_1 = \boldsymbol{M}_1\boldsymbol{M_X} = \boldsymbol{M_X}$, where $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$.

2.15 Consider the following linear regression:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}_1$ is $n \times k_1$, and $\boldsymbol{X}_2$ is $n \times k_2$. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the OLS parameter estimates from running this regression.

Now consider the following regressions, all to be estimated by OLS:

$$\begin{aligned}&\text{a)} \qquad \boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u};\\ &\text{b)} \qquad \boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u};\\ &\text{c)} \qquad \boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u};\\ &\text{d)} \qquad \boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u};\end{aligned}$$

e) $\boldsymbol{P_X y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;

f) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;

g) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;

h) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;

i) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;

j) $\boldsymbol{P_X y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$.

Here $\boldsymbol{P}_1$ projects orthogonally on to the span of $\boldsymbol{X}_1$, and $\boldsymbol{M}_1 = \mathbf{I} - \boldsymbol{P}_1$. For which of the above regressions will the estimates of $\boldsymbol{\beta}_2$ be the same as for the original regression? Why? For which will the residuals be the same? Why?

2.16 Consider the linear regression

$$\boldsymbol{y} = \beta_1\boldsymbol{\iota} + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{\iota}$ is an n–vector of 1s, and $\boldsymbol{X}_2$ is an $n \times (k-1)$ matrix of observations on the remaining regressors. Show, using the FWL Theorem, that the OLS estimators of β_1 and $\boldsymbol{\beta}_2$ can be written as

$$\begin{bmatrix} \hat\beta_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} n & \boldsymbol{\iota}^\top\boldsymbol{X}_2 \\ 0 & \boldsymbol{X}_2^\top\boldsymbol{M_\iota}\boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\iota}^\top\boldsymbol{y} \\ \boldsymbol{X}_2^\top\boldsymbol{M_\iota}\boldsymbol{y} \end{bmatrix},$$

where, as usual, $\boldsymbol{M_\iota}$ is the matrix that takes deviations from the sample mean.

2.17 Show, preferably using (2.36), that $\boldsymbol{P_X} - \boldsymbol{P}_1$ is an orthogonal projection matrix. That is, show that $\boldsymbol{P_X} - \boldsymbol{P}_1$ is symmetric and idempotent. Show further that

$$\boldsymbol{P_X} - \boldsymbol{P}_1 = \boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2},$$

where $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}$ is the projection on to the span of $\boldsymbol{M}_1\boldsymbol{X}_2$. This can be done most easily by showing that any vector in $\mathcal{S}(\boldsymbol{M}_1\boldsymbol{X}_2)$ is invariant under the action of $\boldsymbol{P_X} - \boldsymbol{P}_1$, and that any vector orthogonal to this span is annihilated by $\boldsymbol{P_X} - \boldsymbol{P}_1$.

2.18 Let $\boldsymbol{\iota}$ be a vector of 1s, and let $\boldsymbol{X}$ be an $n \times 3$ matrix, with full rank, of which the first column is $\boldsymbol{\iota}$. What can you say about the matrix $\boldsymbol{M_\iota X}$? What can you say about the matrix $\boldsymbol{P_\iota X}$? What is $\boldsymbol{M_\iota M_X}$ equal to? What is $\boldsymbol{P_\iota M_X}$ equal to?

2.19 Express the four seasonal variables, $\boldsymbol{s}_i$, $i = 1, 2, 3, 4$, defined in (2.47), as functions of the constant $\boldsymbol{\iota}$ and the three variables $\boldsymbol{s}'_i$, $i = 1, 2, 3$, defined in (2.50).

2.20 Show that the full n–dimensional space E^n is the span of the set of **unit basis vectors** $\boldsymbol{e}_t$, $t = 1, \ldots, n$, where all the components of $\boldsymbol{e}_t$ are zero except for the t^{th}, which is equal to 1.

2.21 The file **tbrate.data** contains data for 1950:1 to 1996:4 for three series: r_t, the interest rate on 90-day treasury bills, π_t, the rate of inflation, and y_t, the logarithm of real GDP. For the period 1950:4 to 1996:4, run the regression

$$\Delta r_t = \beta_1 + \beta_2\pi_{t-1} + \beta_3\Delta y_{t-1} + \beta_4\Delta r_{t-1} + \beta_5\Delta r_{t-2} + u_t, \tag{2.71}$$

where Δ is the **first-difference operator**, defined so that $\Delta r_t = r_t - r_{t-1}$. Plot the residuals and fitted values against time. Then regress the residuals on the fitted values and on a constant. What do you learn from this second regression? Now regress the fitted values on the residuals and on a constant. What do you learn from this third regression?

2.22 For the same sample period, regress Δr_t on a constant, Δy_{t-1}, Δr_{t-1}, and Δr_{t-2}. Save the residuals from this regression, and call them $\hat{e}_t$. Then regress π_{t-1} on a constant, Δy_{t-1}, Δr_{t-1}, and Δr_{t-2}. Save the residuals from this regression, and call them $\hat{v}_t$. Now regress $\hat{e}_t$ on $\hat{v}_t$. How are the estimated coefficient and the residuals from this last regression related to anything that you obtained when you estimated regression (2.71)?

2.23 Calculate the diagonal elements of the hat matrix for regression (2.71) and use them to calculate a measure of leverage. Plot this measure against time. On the basis of this plot, which observations seem to have unusually high leverage?

2.24 Show that the t^{th} residual from running regression (2.57) is 0. Use this fact to demonstrate that, as a result of omitting observation t, the t^{th} residual from the regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ changes by an amount

$$\hat{u}_t \frac{h_t}{1-h_t}.$$

2.25 Calculate a vector of "omit 1" residuals $\hat{\boldsymbol{u}}^{(\cdot)}$ for regression (2.71). The t^{th} element of $\hat{\boldsymbol{u}}^{(\cdot)}$ is the residual for the t^{th} observation calculated from a regression that uses data for every observation except the t^{th}. Try to avoid running 185 regressions in order to do this! Regress $\hat{\boldsymbol{u}}^{(\cdot)}$ on the ordinary residuals $\hat{\boldsymbol{u}}$. Is the estimated coefficient roughly the size you expected it to be? Would it be larger or smaller if you were to omit some of the high-leverage observations?

2.26 Show that the leverage measure h_t is the square of the cosine of the angle between the unit basis vector $\boldsymbol{e}_t$ and its projection on to the span $\mathcal{S}(\boldsymbol{X})$ of the regressors.

2.27 Suppose the matrix $\boldsymbol{X}$ is 150×5 and has full rank. Let $\boldsymbol{P_X}$ be the matrix that projects on to $\mathcal{S}(\boldsymbol{X})$ and let $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{P_X}$. What is $\mathrm{Tr}(\boldsymbol{P_X})$? What is $\mathrm{Tr}(\boldsymbol{M_X})$? What would these be if $\boldsymbol{X}$ did not have full rank but instead had rank 3?

2.28 Generate a figure like Figure 2.15 for yourself. Begin by drawing 100 observations of a regressor x_t from the $N(0,1)$ distribution. Then compute and save the h_t for a regression of any regressand on a constant and x_t. Plot the points (x_t, h_t), and you should obtain a graph similar to the one in Figure 2.15.

Now add one more observation, x_{101}. Start with $x_{101} = \bar{x}$, the average value of the x_t, and then increase x_{101} progressively until $x_{101} = \bar{x} + 20$. For each value of x_{101}, compute the leverage measure h_{101}. How does h_{101} change as x_{101} gets larger? Why is this in accord with the result that $h_t = 1$ if the regressors include the dummy variable $\boldsymbol{e}_t$?

Chapter 3

The Statistical Properties of Ordinary Least Squares

3.1 Introduction

In the previous chapter, we studied the numerical properties of ordinary least squares estimation, properties that hold no matter how the data may have been generated. In this chapter, we turn our attention to the **statistical** properties of OLS, ones that depend on how the data were actually generated. These properties can never be shown to hold numerically for any actual data set, but they can be proven to hold if we are willing to make certain assumptions. Most of the properties that we will focus on concern the first two moments of the least squares estimator.

In Section 1.5, we introduced the concept of a **data-generating process**, or **DGP**. For any data set that we are trying to analyze, the DGP is simply the mechanism that actually generated the data. Most real DGPs for economic data are probably very complicated, and economists do not pretend to understand every detail of them. However, for the purpose of studying the statistical properties of estimators, it is almost always necessary to assume that the DGP is quite simple. For instance, when we are studying the (multiple) linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2), \tag{3.01}$$

we may wish to assume that the data were actually generated by the DGP

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta}_0 + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma_0^2). \tag{3.02}$$

The symbol "$\sim$" in (3.01) and (3.02) means "is distributed as." We introduced the abbreviation IID, which means "independently and identically distributed," in Section 1.3. In the model (3.01), the notation $\mathrm{IID}(0, \sigma^2)$ means that the u_t are statistically independent and all follow the same distribution, with mean 0 and variance σ^2. Similarly, in the DGP (3.02), the notation $\mathrm{NID}(0, \sigma_0^2)$ means that the u_t are *normally*, independently, and identically distributed, with mean 0 and variance σ_0^2. In both cases, it is implicitly being assumed that the distribution of u_t is in no way dependent on $\boldsymbol{X}_t$.