CHAPTER 2

BASIC IDEAS OF LINEAR REGRESSION: THE TWO-VARIABLE MODEL

In Chapter 1 we noted that in developing a model of an economic phenomenon (e.g., the law of demand) econometricians make heavy use of a statistical technique known as **regression analysis.** The purpose of this chapter and Chapter 3 is to introduce the basics of regression analysis in terms of the simplest possible linear regression model, namely, the two-variable model. Subsequent chapters will consider various modifications and extensions of the two-variable model.

2.1 THE MEANING OF REGRESSION

As noted in Chapter 1, **regression analysis** is concerned with the study of the relationship between one variable called the **explained,** or **dependent, variable** and one or more other variables called **independent,** or **explanatory, variables.**

Thus, we may be interested in studying the relationship between the quantity demanded of a commodity in terms of the price of that commodity, income of the consumer, and prices of other commodities competing with this commodity. Or, we may be interested in finding out how sales of a product (e.g., automobiles) are related to advertising expenditure incurred on that product. Or, we may be interested in finding out how defense expenditures vary in relation to the gross domestic product (GDP). In all these examples there may be some underlying theory that specifies why we would expect one variable to be dependent or related to one or more other variables. In the first example, the *law of demand* provides the rationale for the dependence of the quantity demanded of a product on its own price and several other variables previously mentioned.

For notational uniformity, from here on we will let Y represent the dependent variable and X the independent, or explanatory, variable. If there is more than one explanatory variable, we will show the various X's by the appropriate subscripts (X_1, X_2, X_3, etc.).

It is very important to bear in mind the warning given in Chapter 1 that, although regression analysis deals with the relationship between a dependent variable and one or more independent variables, *it does not necessarily imply causation;* that is, it does not necessarily mean that the independent variables are the *cause* and the dependent variable is the *effect.* If causality between the two exists, it must be justified on the basis of some (economic) theory. As noted earlier, the law of demand suggests that if all other variables are held constant, the quantity demanded of a commodity is (inversely) dependent on its own price. Here microeconomic theory suggests that the price may be the causal force and the quantity demanded the effect. *Always keep in mind that regression does not necessarily imply causation. Causality must be justified, or inferred, from the theory that underlies the phenomenon that is tested empirically.*

Regression analysis may have one of the following objectives:

1. To estimate the *mean,* or *average,* value of the dependent variable, given the values of the independent variables.
2. To test hypotheses about the nature of the dependence—hypotheses suggested by the underlying economic theory. For example, in the demand function mentioned previously, we may want to test the hypothesis that the price elasticity of demand is, say, –1.0; that is, the demand curve has unitary price elasticity. If the price of the commodity goes up by 1 percent, the quantity demanded on the average goes down by 1 percent, assuming all other factors affecting demand are held constant.
3. To predict, or forecast, the mean value of the dependent variable, given the value(s) of the independent variable(s) beyond the sample range. Thus, in the S.A.T. example discussed in Appendix C, we may wish to predict the average score on the critical reasoning part of the S.A.T. for a group of students who know their scores on the math part of the test (see Table 2-15).
4. One or more of the preceding objectives combined.

2.2 THE POPULATION REGRESSION FUNCTION (PRF): A HYPOTHETICAL EXAMPLE

To illustrate what all this means, we will consider a concrete example. In the last two years of high school, most American teenagers take the S.A.T. college entrance examination. The test consists of three sections: critical reasoning (formerly called the verbal section), mathematics, and an essay portion, each scored on a scale of 0 to 800. Since the essay portion is more difficult to score, we will focus primarily on the mathematics section. Suppose we are interested in finding out whether a student's family income is related to how well students score on the mathematics section of the test. Let Y represent the math S.A.T. score and X represent annual family income. The income variable has been broken into 10 classes:

TABLE 2-1 MATHEMATICS S.A.T. SCORES IN RELATION TO ANNUAL FAMILY INCOME

Math S.A.T. Scores

	Family Income									
Student	\$5,000	\$15,000	\$25,000	\$35,000	\$45,000	\$55,000	\$65,000	\$75,000	\$90,000	\$150,000
1	460	480	460	520	500	450	560	530	560	570
2	470	510	450	510	470	540	480	540	500	560
3	460	450	530	440	450	460	530	540	470	540
4	420	420	430	540	530	480	520	500	570	550
5	440	430	520	490	550	530	510	480	580	560
6	500	450	490	460	510	480	550	580	480	510
7	420	510	440	460	530	510	480	560	530	520
8	410	500	480	520	440	540	500	490	520	520
9	450	480	510	490	510	510	520	560	540	590
10	490	520	470	450	470	550	470	500	550	600
Mean	452	475	478	488	496	505	512	528	530	552

(<\$10,000), (\$10,000–\$20,000), (\$20,000–\$30,000), . . . , (\$80,000–\$100,000), and (>\$100,000). For simplicity, we have used the midpoints of each of the classes, estimating the last class midpoint at \$150,000, for the analysis. Assume that a hypothetical *population* of 100 high school students is reported in Table 2-1.

Table 2.1 can be interpreted as follows: For an annual family income of \$5,000, one student scored a 460 on the math section of the S.A.T. Nine other students had similar family incomes, and their scores, together with the first student, averaged to 452. For a family income of \$15,000, one student scored a 480 on the section, and the average of 10 students in that income bracket was 475. The remaining columns are similar.

A **scattergram** of these data is shown in Figure 2-1. For this graph, the horizontal axis represents annual family income and the vertical axis represents the students' math S.A.T. scores. For each income level, there are several S.A.T. scores; in fact, in this instance there are 10 recorded scores.[1] The points connected with the line are the mean values for each income level. It seems as though there is a general, overall upward trend in the math scores; higher income levels tend to be associated with higher math scores. This is especially evident with the connected open circles, representing the average scores per income level. These connected circles are formally called the **conditional mean** or **conditional expected values** (see Appendix B for details). Since we have assumed the data represent the population of score values, the line connecting the conditional means is called the **population regression line (PRL)**. *The PRL gives the average, or mean, value of the dependent variable (math S.A.T. scores in this*

[1]For simplicity, we are assuming there are 10 scores for each income level. In reality, there may be a very large number of scores for each X (income) value, and each income level need not have the same number of observations.

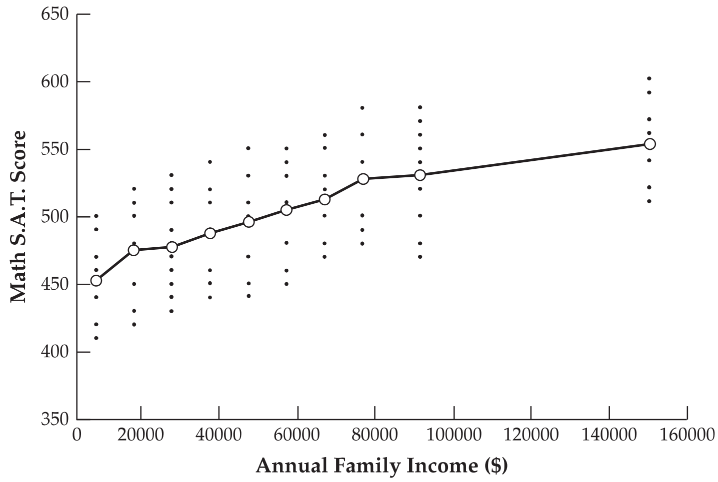

FIGURE 2-1 Annual family income ($) and math S.A.T. score

example) corresponding to each value of the independent variable (here, annual family income) in the population as a whole. Thus, corresponding to an annual income of $25,000, the *average* math S.A.T. score is 478, whereas corresponding to an annual income of $45,000, the *average* math S.A.T. score is 496. In short, the PRL tells us how the mean, or average, value of Y (or any dependent variable) is related to each value of X (or any independent variable) in the whole population.

Since the PRL in Figure 2-1 is approximately linear, we can express it mathematically in the following functional form:

$$E(Y \mid X_i) = B_1 + B_2 X_i \tag{2.1}$$

which is the mathematical equation of a straight line. In Equation (2.1), $E(Y \mid X_i)$ means the mean, or expected value, of Y corresponding to, or *conditional* upon, a given value of X. The subscript i refers to the ith subpopulation. Thus, in Table 2-1, $E(Y \mid X_i = 5000)$ is 452, which is the mean, or expected, value of Y in the first subpopulation (i.e., corresponding to $X = \$5000$).

The last row of Table 2-1 gives the conditional mean values of Y. It is very important to note that $E(Y \mid X_i)$ is a function of X_i (linear in the present example). This means that the dependence of Y on X, technically called the *regression of Y on X*, can be defined simply as the mean of the distribution of Y values (as in Table 2-1), which has the given X. In other words, *the population regression line (PRL) is a line that passes through the conditional means of Y.* The mathematical form in which the PRL is expressed, such as Eq. (2.1), is called the **population regression function (PRF)**, as it represents the regression line in the population as a whole. In the present instance the PRF is linear. (The more technical meaning of linearity is discussed in Section 2.6.)

In Eq. (2.1), B_1 and B_2 are called the **parameters,** also known as the **regression coefficients.** B_1 is also known as the **intercept** (coefficient) and B_2 as the **slope** (coefficient). *The slope coefficient measures the rate of change in the (conditional) mean value of Y per unit change in X.* If, for example, the slope coefficient (B_2) were 0.001, it would suggest that if annual family income were to increase by a dollar, the (conditional) mean value of Y would increase by 0.001 points. Because of the scale of the variables, it is easier to interpret the results for a one-thousand-dollar increase in annual family income; for each one-thousand-dollar increase in annual family income, we would expect to see a 1 point increase in the (conditional) mean value of the math S.A.T. score. B_1 is the (conditional) mean value of Y if X is zero; it gives the average value of the math S.A.T. score if the annual family income were zero. We will have more to say about this interpretation of the intercept later in the chapter.

How do we go about finding the estimates, or numerical values, of the intercept and slope coefficients? We explore this in Section 2.8.

Before moving on, a word about terminology is in order. Since in regression analysis, as noted in Chapter 1, we are concerned with examining the behavior of the dependent variable *conditional upon the given values of the independent variable(s), our approach to regression analysis can be termed* **conditional regression analysis.**[2] As a result, there is no need to use the adjective "conditional" all the time. Therefore, *in the future expressions like* $E(Y \mid X_i)$ *will be simply written as* $E(Y)$, *with the explicit understanding that the latter in fact stands for the former.* Of course, where there is cause for confusion, we will use the more extended notation.

2.3 STATISTICAL OR STOCHASTIC SPECIFICATION OF THE POPULATION REGRESSION FUNCTION

As we just discussed, the PRF gives the average value of the dependent variable corresponding to each value of the independent variable. Let us take another look at Table 2-1. We know, for example, that corresponding to $X = \$75{,}000$, the average Y is 528 points. But if we pick one student at *random* from the 10 students corresponding to this income, we know that the math S.A.T. score for that student will not necessarily be equal to the mean value of 528. To be concrete, take the last student in this group. His or her math S.A.T. score is 500, which is below the mean value. By the same token, if you take the first student in that group, his or her score is 530, which is above the average value.

How do you explain the score of an individual student in relation to income? The best we can do is to say that any individual's math S.A.T. score is equal to

[2]The fact that our analysis is conditional on X does not mean that X causes Y. It is just that we want to see the behavior of Y in relation to an X variable that is of interest to the analyst. For example, when the Federal Reserve Bank (the Fed) changes the Federal funds rate, it is interested in finding out how the economy responds. During the economic crisis of 2008 in the United States, the Fed reduced the Federal Funds rate several times to resuscitate the ailing economy. One of the key determinants of the demand for housing is the mortgage interest rate. It is therefore of great interest to prospective homeowners to track the mortgage interest rates. When the Fed reduces the Federal Funds rate, all other interest rates follow suit.

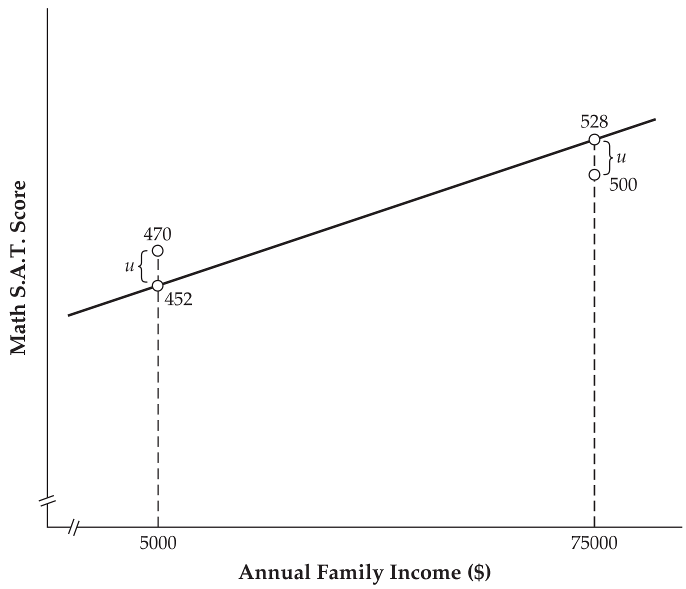

FIGURE 2-2 Math S.A.T. scores in relation to family income

the average for that group plus or minus some quantity. Let us express this mathematically as

$$Y_i = B_1 + B_2X_i + u_i \tag{2.2}$$

where u is known as the **stochastic,** or **random, error term,** or simply the **error term**.[3] We have already encountered this term in Chapter 1. The error term is a *random variable* (r.v.), for its value cannot be controlled or known a priori. As we know from Appendix A, an r.v. is usually characterized by its probability distribution (e.g., the normal or the t distribution).

How do we interpret Equation (2.2)? We can say that a student's math S.A.T. score, say, the ith individual, corresponding to a specific family income can be expressed as the sum of two components. The first component is $(B_1 + B_2X_i)$, which is simply the mean, or average, math score in the ith subpopulation; that is, the point on the PRL corresponding to the family income. This component may be called the *systematic*, or *deterministic,* component. The second component is u_i, which may be called the *nonsystematic,* or random, component (i.e., determined by factors other than income). The error term u_i is also known as the **noise component.**

To see this clearly, consider Figure 2-2, which is based on the data of Table 2-1.

As this figure shows, at annual family income = \$5000, one student scores 470 on the test, whereas the average math score at this income level is 452. Thus, this

[3]The word *stochastic* comes from the Greek word *stokhos* meaning a "bull's eye." The outcome of throwing darts onto a dart board is a stochastic process, that is, a process fraught with misses. In statistics, the word implies the presence of a random variable—a variable whose outcome is determined by a chance experiment.

student's score exceeds the systematic component (i.e., the mean for the group) by 18 points. So his or her u component is +18 units. On the other hand, at income = \$75,000, a randomly chosen second student scores 500 on the math test, whereas the average score for this group is 528. This person's math score is less than the systematic component by 28 points; his or her u component is thus −28.

Eq. (2.2) is called the **stochastic (or statistical) PRF,** whereas Eq. (2.1) is called the **deterministic,** or **nonstochastic, PRF.** The latter represents the means of the various Y values corresponding to the specified income levels, whereas the former tells us how individual math S.A.T. scores vary around their mean values due to the presence of the stochastic error term, u.

What is the nature of the u term?

2.4 THE NATURE OF THE STOCHASTIC ERROR TERM

1. The error term may represent the influence of those variables that are not explicitly included in the model. For example in our math S.A.T. scenario it may very well represent influences, such as a person's wealth, the area where he or she lives, high school GPA, or math courses taken in school.
2. Even if we included all the relevant variables determining the math test score, some intrinsic randomness in the math score is bound to occur that cannot be explained no matter how hard we try. Human behavior, after all, is not totally predictable or rational. Thus, u may reflect this inherent randomness in human behavior.
3. u may also represent errors of measurement. For example, the data on annual family income may be rounded or the data on math scores may be suspect because in some communities few students plan to attend college and therefore don't take the test.
4. The *principle of Ockham's razor*—that descriptions be kept as simple as possible until proved inadequate—would suggest that we keep our regression model as simple as possible. Therefore, even if we know what other variables might affect Y, their combined influence on Y may be so small and nonsystematic that you can incorporate it in the random term, u. Remember that a model is a simplification of reality. If we truly want to build reality into a model it may be too unwieldy to be of any practical use. In model building, therefore, some abstraction from reality is inevitable. By the way, William Ockham (1285–1349) was an English philosopher who maintained that a complicated explanation should not be accepted without good reason and wrote "*Frustra fit per plura, quod fieri potest per pauciora*—It is vain to do with more what can be done with less."

It is for one or more of these reasons that an individual student's math S.A.T. score will deviate from his or her group average (i.e., the systematic component). And as we will soon discover, this error term plays an extremely crucial role in regression analysis.

2.5 THE SAMPLE REGRESSION FUNCTION (SRF)

How do we estimate the PRF of Eq. (2.1), that is, obtain the values of B_1 and B_2? If we have the data from Table 2-1, the whole population, this would be a relatively straightforward task. All we have to do is to find the conditional means of Y corresponding to each X and then join these means. Unfortunately, in practice, we rarely have the entire population at our disposal. Often we have only a *sample* from this population. (Recall from Chapter 1 and Appendix A our discussion regarding the population and the sample.) Our task here is to estimate the PRF on the basis of the sample information. How do we accomplish this?

Pretend that you have never seen Table 2-1 but only had the data given in Table 2-2, which presumably represent a randomly selected sample of Y values corresponding to the X values shown in Table 2-1.

Unlike Table 2-1, we now have only one Y value corresponding to each X. The important question that we now face is: From the sample data of Table 2-2, can we estimate the average S.A.T. math score in the population as a whole corresponding to each X? In other words, can we estimate the PRF from the sample data? As you can well surmise, we may not be able to estimate the PRF accurately because of *sampling fluctuations,* or *sampling error,* a topic we discuss in Appendix C. To see this clearly, suppose another random sample, which is shown in Table 2-3, is drawn from the population of Table 2-1. If we plot the data of Tables 2-2 and 2-3, we obtain the scattergram shown in Figure 2-3.

Through the scatter points we have drawn visually two straight lines that fit the scatter points reasonably well. We will call these lines the **sample regression lines (SRLs).** Which of the two SRLs represents the true PRL? If we avoid the temptation of looking at Figure 2-1, which represents the PRL, there is no way we can be sure that either of the SRLs shown in Figure 2-3 represents the true PRL. For if we had yet another sample, we would obtain a third SRL. Supposedly, each SRL represents the PRL, but because of sampling variation, each is at best an approximation of the true PRL. In general, we would get K different SRLs for K different samples, and all these SRLs are not likely to be the same.

TABLE 2-2 A RANDOM SAMPLE FROM TABLE 2-1

Y	X
410	5000
420	15000
440	25000
490	35000
530	45000
530	55000
550	65000
540	75000
570	90000
590	150000

TABLE 2-3 A RANDOM SAMPLE FROM TABLE 2-1

Y	X
420	5000
520	15000
470	25000
450	35000
470	45000
550	55000
470	65000
500	75000
550	90000
600	150000

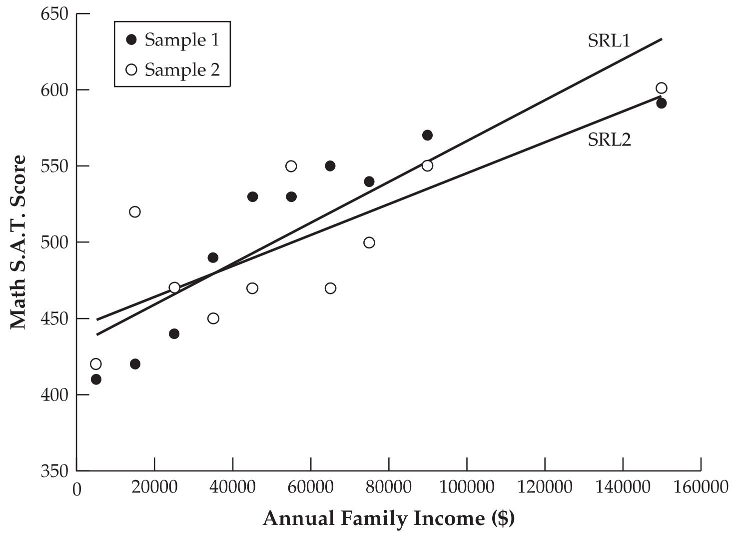

FIGURE 2-3 Sample regression lines based on two independent samples

Now analogous to the PRF that underlies the PRL, we can develop the concept of the **sample regression function (SRF)** to represent the SRL. The sample counterpart of Eq. (2.1) may be written as

$$\hat{Y}_i = b_1 + b_2X_i \tag{2.3}$$

where ^ is read as "hat" or "cap," and

where $\hat{Y}_i$ = estimator of $E(Y \mid X_i)$, the estimator of the population conditional mean
b_1 = estimator of B_1
b_2 = estimator of B_2

As noted in Appendix D, an **estimator,** or a **sample statistic,** is a rule or a formula that suggests how we can estimate the population parameter at hand. A particular numerical value obtained by the estimator in an application, as we know, is an **estimate.** (See Appendix D for the discussion on point and interval estimators.)

If we look at the scattergram in Figure 2-3, we observe that not all the sample data lie exactly on the respective sample regression lines. Therefore, just as we developed the stochastic PRF of Eq. (2.2), we need to develop the stochastic version of Eq. (2.3), which we write as

$$Y_i = b_1 + b_2X_i + e_i \tag{2.4}$$

where e_i = the estimator of u_i.

We call e_i the **residual term,** or simply the **residual.** Conceptually, it is analogous to u_i and can be regarded as the estimator of the latter. It is introduced in the SRF for the same reasons as u_i was introduced in the PRF. *Simply stated, e_i represents the difference between the actual Y values and their estimated values from the sample regression. That is,*

$$e_i = Y_i - \hat{Y}_i \tag{2.5}$$

To summarize, our primary objective in regression analysis is to estimate the (stochastic) PRF

$$Y_i = B_1 + B_2X_i + u_i$$

on the basis of the SRF

$$Y_i = b_1 + b_2X_i + e_i$$

because more often than not our analysis is based on a single sample from some population. But because of sampling variation, our estimate of the PRF based on the SRF is only approximate. This approximation is shown in Figure 2-4. *Keep in mind that we actually do not observe B_1, B_2, and u. What we observe are their proxies, b_1, b_2, and e, once we have a specific sample.*

For a given X_i, shown in this figure, we have one (sample) observation, Y_i. In terms of the SRF, the observed Y_i can be expressed as

$$Y_i = \hat{Y}_i + e_i \tag{2.6}$$

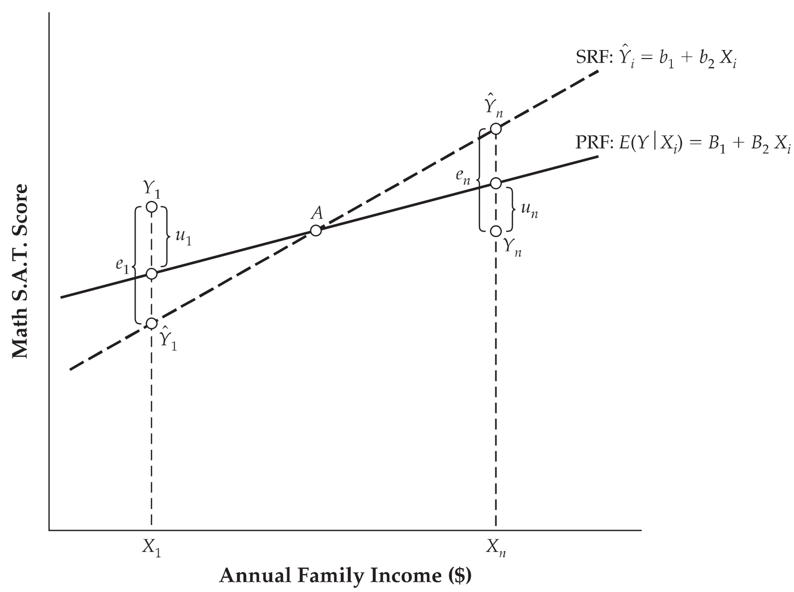

FIGURE 2-4 The population and sample regression lines

and in terms of the PRF it can be expressed as

$$Y_i = E(Y \mid X_i) + u_i \tag{2.7}$$

Obviously, in Figure 2-4, $\hat{Y}_1$ underestimates the true mean value $E(Y \mid X_1)$ for the X_1 shown therein. By the same token, for any Y to the right of point A in Figure 2-4 (e.g., $\hat{Y}_n$), the SRF will *overestimate* the true PRF. But you can readily see that such over- and underestimation is inevitable due to sampling fluctuations.

The important question now is: Granted that the SRF is only an approximation of the PRF, can we find a method or a procedure that will make this approximation as close as possible? In other words, how should we construct the SRF so that b_1 is as close as possible to B_1 and b_2 is as close as possible to B_2, because generally we do not have the entire population at our disposal? As we will show in Section 2.8, we can indeed find a "best-fitting" SRF that will mirror the PRF as faithfully as possible. *It is fascinating to consider that this can be done even though we never actually determine the PRF itself.*

2.6 THE SPECIAL MEANING OF THE TERM "LINEAR" REGRESSION

Since in this text we are concerned primarily with "linear" models like Eq. (2.1), it is essential to know what the term *linear* really means, for it can be interpreted in two different ways.

Linearity in the Variables

The first and perhaps the more "natural" meaning of linearity is that the conditional mean value of the dependent variable is a linear function of the independent variable(s) as in Eq. (2.1) or Eq. (2.2) or in the sample counterparts, Eqs. (2.3) and (2.4).[4] In this interpretation, the following functions are not linear:

$$E(Y) = B_1 + B_2 X_i^2 \tag{2.8}$$

$$E(Y) = B_1 + B_2 \frac{1}{X_i} \tag{2.9}$$

because in Equation (2.8) X appears with a power of 2, and in Eq. (2.9) it appears in the inverse form. For regression models linear in the explanatory variable(s), the rate of change in the dependent variable remains constant for a unit change in the explanatory variable; that is, the slope remains constant. But for a regression

[4]A function $Y = f(X)$ is said to be linear in X if (1) X appears with a power of 1 only; that is, terms such as X^2 and $\sqrt{X}$ are excluded; and (2) X is not multiplied or divided by another variable (e.g., $X \cdot Z$ and X/Z, where Z is another variable).

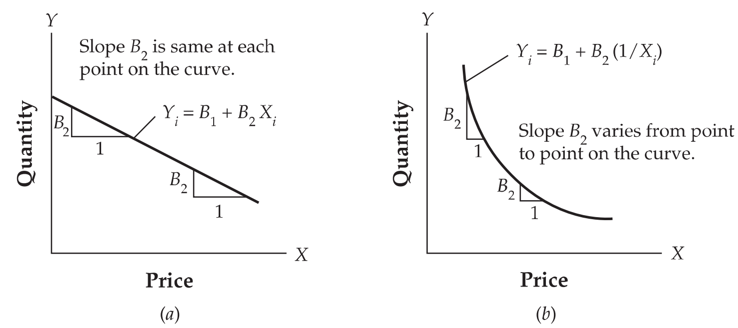

FIGURE 2-5 (*a*) Linear demand curve; (*b*) nonlinear demand curve

model nonlinear in the explanatory variables the slope does not remain constant. This can be seen more clearly in Figure 2-5.

As Figure 2-5 shows, for the regression (2.1), the slope—the rate of change in $E(Y)$—the mean of Y, remains the same, namely, B_2 no matter at what value of X we measure the change. But for regression, say, Eq. (2.8), the rate of change in the mean value of Y varies from point to point on the regression line; it is actually a curve here.[5]

Linearity in the Parameters

The second interpretation of linearity is that the conditional mean of the dependent variable is a linear function of the parameters, the B's; it may or may not be linear in the variables. Analogous to a linear-in-variable function, a function is said to be linear in the parameter, say, B_2, if B_2 appears with a power of 1 only. On this definition, models (2.8) and (2.9) are both linear models because B_1 and B_2 enter the models linearly. It does not matter that the variable X enters nonlinearly in both models. However, a model of the type

$$E(Y) = B_1 + B_2^2 X_i \tag{2.10}$$

is nonlinear in the parameter model since B_2 enters with a power of 2.

In this book we are primarily concerned with models that are linear in the parameters. *Therefore, from now on the term* **linear regression** *will mean a regression that is linear in the parameters, the B's* (i.e., the parameters are raised to the power of 1 only); *it may or may not be linear in the explanatory variables.*[6]

[5]Those who know calculus will recognize that in the linear model the slope, that is, the derivative of Y with respect to X, is constant, equal to B_2, but in the nonlinear model Eq. (2.8) it is equal to $-B_2(1/X_i^2)$, which obviously will depend on the value of X at which the slope is measured, and is therefore not constant.

[6]This is not to suggest that nonlinear (in-the-parameters) models like Eq. (2.10) cannot be estimated or that they are not used in practice. As a matter of fact, in advanced courses in econometrics such models are studied in depth.

2.7 TWO-VARIABLE VERSUS MULTIPLE LINEAR REGRESSION

So far in this chapter we have considered only the **two-variable,** or **simple, regression** models in which the dependent variable is a function of just one explanatory variable. This was done just to introduce the fundamental ideas of regression analysis. But the concept of regression can be extended easily to the case where the dependent variable is a function of more than one explanatory variable. For instance, if the math S.A.T. score is a function of income (X_2), number of math classes taken (X_3), and age of the student (X_4), we can write the extended math S.A.T. function as

$$E(Y) = B_1 + B_2X_{2i} + B_3X_{3i} + B_4X_{4i} \tag{2.11}$$

[*Note:* $E(Y) = E(Y \mid X_{2i}, X_{3i}, X_{4i})$.]

Equation (2.11) is an example of a **multiple linear regression,** a regression in which more than one independent, or explanatory, variable is used to explain the behavior of the dependent variable. Model (2.11) states that the (conditional) mean value of the math S.A.T. score is a linear function of income, number of math classes taken, and age of the student. The score function of a student (i.e., the stochastic PRF) can be expressed as

$$\begin{aligned} Y_i &= B_1 + B_2X_{2i} + B_3X_{3i} + B_4X_{4i} + u_i \\ &= E(Y) + u_i \end{aligned} \tag{2.12}$$

which shows that the individual math S.A.T. score will differ from the group mean by the factor u, which is the stochastic error term. As noted earlier, even in a multiple regression we introduce the error term because we cannot take into account all the forces that might affect the dependent variable.

Notice that both Eqs. (2.11) and (2.12) are linear in the parameters and are therefore *linear regression models.* The explanatory variables themselves do not need to enter the model linearly, although in the present example they do.

2.8 ESTIMATION OF PARAMETERS: THE METHOD OF ORDINARY LEAST SQUARES

As noted in Section 2.5, we estimate the population regression function (PRF) on the basis of the sample regression function (SRF), since in practice we only have a sample (or two) from a given population. How then do we estimate the PRF? And how do we find out whether the estimated PRF (i.e., the SRF) is a "good" estimate of the true PRF? We will answer the first question in this chapter and take up the second question—of the "goodness" of the estimated PRF—in Chapter 3.

To introduce the fundamental ideas of estimation of the PRF, we consider the simplest possible linear regression model, namely, the two-variable linear regression in which we study the relationship of the dependent variable Y to a single explanatory variable X. In Chapter 4 we extend the analysis to the multiple regression, where we will study the relationship of the dependent variable Y to more than one explanatory variable.

The Method of Ordinary Least Squares

Although there are several methods of obtaining the SRF as an estimator of the true PRF, in regression analysis the method that is used most frequently is that of *least squares (LS)*, more popularly known as the **method of ordinary least squares (OLS).**[7] We will use the terms LS and OLS methods interchangeably. To explain this method, we first explain the **least squares principle.**

The Least Squares Principle Recall our two-variable PRF, Eq. (2.2):

$$Y_i = B_1 + B_2 X_i + u_i$$

Since the PRF is not directly observable (Why?), we estimate it from the SRF

$$Y_i = b_1 + b_2 X_1 + e_i$$

which we can write as

$$\begin{aligned} e_i &= \text{actual } Y_i - \text{predicted } Y_i \\ &= Y_i - \hat{Y}_i \\ &= Y_i - b_1 - b_2 X_i \text{ [using Eq. (2.3)]} \end{aligned}$$

which shows that the residuals are simply the differences between the actual and estimated Y values, the latter obtained from the SRF, Eq. (2.3). This can be seen more vividly in Figure 2-4.

Now the best way to estimate the PRF is to choose b_1 and b_2, the estimators of B_1 and B_2, in such a way that the residuals e_i are as small as possible. The method of **ordinary least squares (OLS)** states that b_1 and b_2 should be chosen in such a way that the **residual sum of squares (RSS),** $\sum e_i^2$, is as small as possible.[8]

Algebraically, the least squares principle states

$$\begin{aligned} \textit{Minimize} \sum e_i^2 &= \sum (Y_i - \hat{Y})^2 \\ &= \sum (Y_i - b_1 - b_2 X_i)^2 \end{aligned} \qquad \textbf{(2.13)}$$

[7]Despite the name, there is nothing ordinary about this method. As we will show, this method has several desirable statistical properties. It is called OLS because there is another method, called the *generalized least squares* (GLS) method, of which OLS is a special case.

[8]Note that the smaller the e_i is, the smaller their sum of squares will be. The reason for considering the squares of e_i and not the e_i themselves is that this procedure avoids the problem of the sign of the residuals. Note that e_i can be positive as well as negative.

As you can observe from Eq. (2.13), once the sample values of Y and X are given, RSS is a function of the estimators b_1 and b_2. Choosing different values of b_1 and b_2 will yield different e's and hence different values of RSS. To see this, just rotate the SRF shown in Figure 2-4 any way you like. For each rotation, you will get a different intercept (i.e., b_1) and a different slope (i.e., b_2). We want to choose the values of these estimators that will give the smallest possible RSS.

How do we actually determine these values? This is now simply a matter of arithmetic and involves the technique of differential calculus. Without going into detail, it can be shown that the values of b_1 and b_2 that actually minimize the RSS given in Eq. (2.13) are obtained by solving the following two simultaneous equations. (The details are given in Appendix 2A at the end of this chapter.)

$$\sum Y_i = nb_1 + b_2 \sum X_i \tag{2.14}$$

$$\sum Y_i X_i = b_1 \sum X_i + b_2 \sum X_i^2 \tag{2.15}$$

where n is the sample size. These simultaneous equations are known as the (least squares) **normal equations.**

In Equations (2.14) and (2.15) the unknowns are the b's and the knowns are the quantities involving sums, squared sums, and the sum of the cross-products of the variables Y and X, which can be easily obtained from the sample at hand. Now solving these two equations simultaneously (using any high school algebra trick you know), we obtain the following solutions for b_1 and b_2.

$$b_1 = \overline{Y} - b_2\overline{X} \tag{2.16}$$

which is the estimator of the population intercept, B_1. The sample intercept is thus the sample mean value of Y minus the estimated slope times the sample mean value of X.

$$\begin{aligned} b_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \\ &= \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum X_i^2 - n\overline{X}^2} \end{aligned} \tag{2.17}$$

which is the estimator of the population slope coefficient B_2. Note that

$$x_i = (X_i - \overline{X}) \quad \text{and} \quad y_i = (Y_i - \overline{Y})$$

that is, *the small letters denote deviations from the sample mean values, a convention that we will adopt in this book*. As you can see from the formula for b_2, it is simpler to write the estimator using the deviation form. *Expressing the values of a variable from its mean value does not change the ranking of the values, since we are subtracting the same constant from each value*. Note that b_1 and b_2 are solely expressed in terms of quantities that can be readily computed from the sample at hand. Of course, these days the computer will do all the calculations for you.

The estimators given in Equations (2.16) and (2.17) are known as **OLS estimators,** since they are obtained by the method of OLS.

Before proceeding further, we should note a few interesting features of the OLS estimators given in Eqs. (2.16) and (2.17):

1. The SRF obtained by the method of OLS passes through the sample mean values of X and Y, which is evident from Eq. (2.16), for it can be written as

$$\overline{Y} = b_1 + b_2\overline{X} \tag{2.18}$$

2. The mean value of the residuals, $\bar{e}(=\sum e_i/n)$ is always zero, which provides a check on the arithmetical accuracy of the calculations (see Table 2-4).
3. The sum of the product of the residuals e and the values of the explanatory variable X is zero; that is, these two variables are uncorrelated (on the definition of correlation, see Appendix B). Symbolically,

$$\sum e_i X_i = 0 \tag{2.19}$$

   This provides yet another check on the least squares calculations.
4. The sum of the product of the residuals e_i and the estimated $Y_i(=\hat{Y}_i)$ is zero; that is, $\sum e_i\hat{Y}_i$ is zero (see Question 2.25).

2.9 PUTTING IT ALL TOGETHER

Let us use the sample data given in Table 2-2 to compute the values of b_1 and b_2. The necessary computations involved in implementing formulas (2.16) and (2.17) are laid out in Table 2-4. Keep in mind that the data given in Table 2-2 are a random sample from the population given in Table 2-1.

From the computations shown in Table 2-4, we obtain the following sample math S.A.T. score regression:

$$\hat{Y}_i = 432.4138 + 0.0013X_i \tag{2.20}$$

where Y represents math S.A.T. score and X represents annual family income. *Note that we have put a cap on Y to remind us that it is an estimator of the true population mean corresponding to the given level of X (recall Eq. 2.3).* The estimated regression line is shown in Figure 2-6.

TABLE 2-4 RAW DATA (FROM TABLE 2-2) FOR MATH S.A.T. SCORES

Y_i	X_i	$\sum Y_iX_i$	X_i^2	x_i	y_i	x_i^2	y_i^2	$\sum y_ix_i$	$\hat{Y}_i$	e_i	e_i^2	$\sum e_ix_i$
410	5000	2050000	25000000	−51000	−97	2601000000	9409	4947000	439.073	−29.0733	845.255	1482737.069
420	15000	6300000	225000000	−41000	−87	1681000000	7569	3567000	452.392	−32.3922	1049.257	1328081.897
440	25000	11000000	625000000	−31000	−67	961000000	4489	2077000	465.711	−25.7112	661.066	797047.4138
490	35000	17150000	1225000000	−21000	−17	441000000	289	357000	479.030	10.9698	120.337	−230366.3793
530	45000	23850000	2025000000	−11000	23	121000000	529	−253000	492.349	37.6509	1417.587	−414159.4828
530	55000	29150000	3025000000	−1000	23	1000000	529	−23000	505.668	24.3319	592.0412	−24331.89655
550	65000	35750000	4225000000	9000	43	81000000	1849	387000	518.987	31.0129	961.8019	279116.3793
540	75000	40500000	5625000000	19000	33	361000000	1089	627000	532.306	7.69397	59.1971	146185.3448
570	90000	51300000	8100000000	34000	63	1156000000	3969	2142000	552.284	17.7155	313.8396	602327.5862
590	150000	88500000	22500000000	94000	83	8836000000	6889	7802000	632.198	−42.1982	1780.694	−3966637.931
5070	560000	305550000	47600000000	0	0	16240000000	36610	21630000	5070	0	7801.0776	0

Note: $x_i = (X_i - \overline{X})$; $y_i = (Y_i - \overline{Y})$; $\overline{X} = 56000$; $\overline{Y} = 507$.

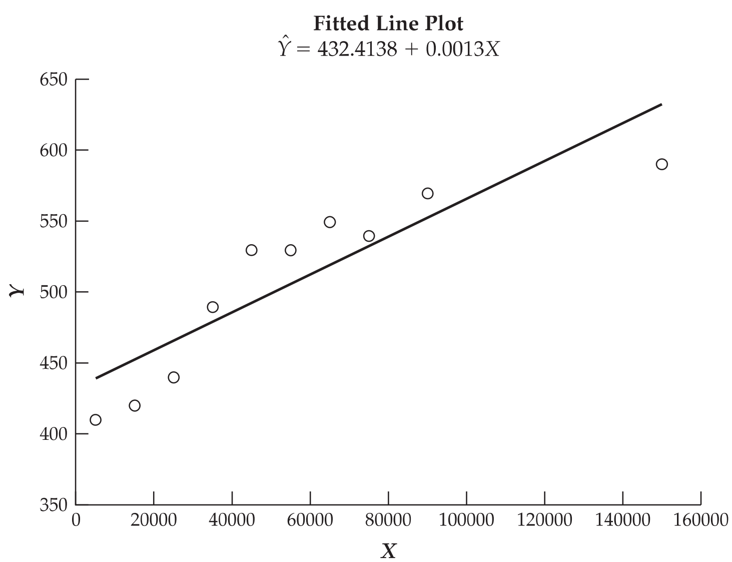

FIGURE 2-6 Regression line based on data from Table 2-4

Interpretation of the Estimated Math S.A.T. Score Function

The interpretation of the estimated math S.A.T. score function is as follows: The slope coefficient of 0.0013 means that, other things remaining the same, if annual family income goes up by a dollar, the *mean* or *average* math S.A.T. score goes up by about 0.0013 points. The intercept value of 432.4138 means that if family income is zero, the mean math score will be about 432.4138. Very often such an interpretation has no economic meaning. For example, we have no data where an annual family income is zero. *As we will see throughout the book, often the intercept has no particular economic meaning.* In general you have to use common sense in interpreting the intercept term, for very often the sample range of the X values (family income in our example) may not include zero as one of the observed values. *Perhaps it is best to interpret the intercept term as the mean or average effect on Y of all the variables omitted from the regression model.*

2.10 SOME ILLUSTRATIVE EXAMPLES

Now that we have discussed the OLS method and learned how to estimate a PRF, let us provide some concrete applications of regression analysis.

Example 2.1. Years of Schooling and Average Hourly Earnings

Based on a sample of 528 observations, Table 2-5 gives data on average hourly wage Y($) and years of schooling (X).

Suppose we want to find out how Y behaves in relation to X. From human capital theories of labor economics, we would expect average wage to increase with years of schooling. That is, we expect a positive relationship between the two variables; it would be bad news if such were not the case.

The regression results based on the data in Table 2-5 are as follows:

$$\hat{Y}_i = -0.0144 + 0.7241X_i \tag{2.21}$$

TABLE 2-5 AVERAGE HOURLY WAGE BY EDUCATION

Years of schooling	Average hourly wage ($)	Number of people
6	4.4567	3
7	5.7700	5
8	5.9787	15
9	7.3317	12
10	7.3182	17
11	6.5844	27
12	7.8182	218
13	7.8351	37
14	11.0223	56
15	10.6738	13
16	10.8361	70
17	13.6150	24
18	13.5310	31

Source: Arthur S. Goldberger, *Introductory Econometrics*, Harvard University Press, Cambridge, Mass., 1998, Table 1.1, p. 5 (adapted).

As these results show, there is a positive association between education and earnings, which accords with prior expectations. For every additional year of schooling, the mean wage rate goes up by about 72 cents per hour.[9] The negative intercept in the present instance has no particular economic meaning.

Example 2.2. Okun's Law

Based on the U.S. data for 1947 to 1960, the late Arthur Okun of the Brookings Institution and a former chairman of the President's Council of Economic Advisers obtained the following regression, known as Okun's law:

$$Y_t = -0.4(X_t - 2.5) \tag{2.22}$$

where Y_t = change in the unemployment rate, percentage points
X_t = percent growth rate in real output, as measured by real GDP
2.5 = the long-term, or trend, rate of growth of output historically observed in the United States

In this regression the intercept is zero and the slope coefficient is −0.4. Okun's law says that for every percentage point of growth in real GDP above 2.5 percent, the unemployment rate declines by 0.4 percentage points.

Okun's law has been used to predict the required growth in real GDP to reduce the unemployment rate by a given percentage point. Thus, a growth rate of 5 percent in real GDP will reduce the unemployment rate by 1 percentage point, or a growth rate of 7.5 percent is required to reduce the unemployment rate by 2 percentage points. In Problem 2.17, which gives comparatively more recent data, you are asked to find out if Okun's law still holds.

This example shows how sometimes a simple (i.e., two-variable) regression model can be used for policy purposes.

Example 2.3. Stock Prices and Interest Rates

Stock prices and interest rates are key economic indicators. Investors in stock markets, individual or institutional, watch very carefully the movements in the interest rates. Since interest rates represent the cost of borrowing money, they have a vast effect on investment and hence on the profitability of a company. Macroeconomic theory would suggest an inverse relationship between stock prices and interest rates.

As a measure of stock prices, let us use the S&P 500 composite index ($1941–1943 = 10), and as a measure of interest rates, let us use the three-month

[9]Since the data in Table 2-5 refer to the mean wage for the various categories, the slope coefficient here should strictly be interpreted as the average increase in the mean hourly earnings.

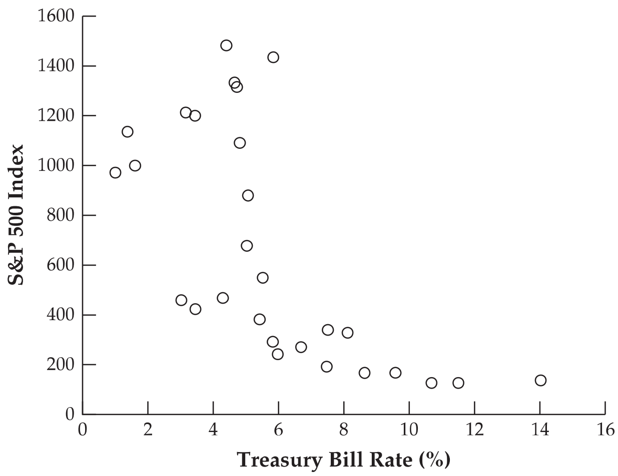

FIGURE 2-7 S&P 500 composite index and three-month Treasury bill rate, 1980–2007

Treasury bill rate (%). Table 2-6, found on the textbook's Web site, gives data on these variables for the period 1980–2007.

Plotting these data, we obtain the scattergram as shown in Figure 2-7. The scattergram clearly shows that there is an inverse relationship between the two variables, as per theory. But the relationship between the two is not linear (i.e., straight line); it more closely resembles Figure 2-5(*b*). Therefore, let us maintain that the true relationship is:

$$Y_t = B_1 + B_2(1/X_i) + u_i \tag{2.23}$$

Note that Eq. (2.23) is a linear regression model, as the parameters in the model are linear. It is, however, nonlinear in the variable X. If you let $Z = 1/X$, then the model is linear in the parameters as well as the variables Y and Z.

Using the EViews statistical package, we estimate Eq. (2.23) by OLS, giving the following results:

$$\hat{Y}_t = 404.4067 + 996.866(1/X_t) \tag{2.24}$$

How do we interpret these results? The value of the intercept has no practical economic meaning. The interpretation of the coefficient of $(1/X)$ is rather tricky. Literally interpreted, it suggests that if the reciprocal of the three-month Treasury bill rate goes up by one unit, the average value of the S&P 500 index will go up by about 997 units. This is, however, not a very enlightening interpretation. If you want to measure the rate of change of (mean) Y with respect to X (i.e., the derivative of Y with respect to X), then as footnote 5 shows, this rate of change is given by $-B_2(1/X_i^2)$, which depends on the value taken by X. Suppose $X = 2$. Knowing that the estimated B_2 is 996.866, we find the rate of change at this X value as -249.22 (approx). That is, starting with a Treasury bill rate of about 2 percent, if that rate goes up by one percentage point, on average, the S&P 500 index will decline by about 249 units. Of course, an increase in the Treasury bill rate from 2 percent to 3 percent is a substantial increase.

Interestingly, if you had disregarded Figure 2-5 and had simply fitted the straight line regression to the data in Table 2-6, (found on the textbook's Web site), you would obtain the following regression:

$$\hat{Y}_t = 1229.3414 - 99.4014X_t \quad \textbf{(2.25)}$$

Here the interpretation of the intercept term is that if the Treasury bill rate were zero, the average value of the S&P index would be about 1229. Again, this may not have any concrete economic meaning. The slope coefficient here suggests that if the Treasury bill rate were to increase by one unit, say, one percentage point, the average value of the S&P index would go down by about 99 units.

Regressions (2.24) and (2.25) bring out the practical problems in choosing an appropriate model for empirical analysis. Which is a better model? How do we know? What tests do we use to choose between the two models? We will provide answers to these questions as we progress through the book (see Chapter 5). *A question to ponder*: In Eq. (2.24) the sign of the slope coefficient is positive, whereas in Eq. (2.25) it is negative. Are these findings conflicting?

Example 2.4. Median Home Price and Mortgage Interest Rate in the United States, 1980–2007

Over the past several years there has been a surge in home prices across the United States. It is believed that this surge is due to sharply falling mortgage interest rates. To see the impact of mortgage interest rates on home prices, Table 2-7 (found on the textbook's Web site) gives data on median home prices (1000 $) and 30-year fixed rate mortgage (%) in the United States for the period 1980–2007.

These data are plotted in Figure 2-8.

As a first approximation, if you fit a straight line regression model, you will obtain the following results, where Y = median home price (1000 $) and X = 30-year fixed rate mortgage (%):

$$\hat{Y}_t = 329.0041 - 17.3694X_t \quad \textbf{(2.26)}$$

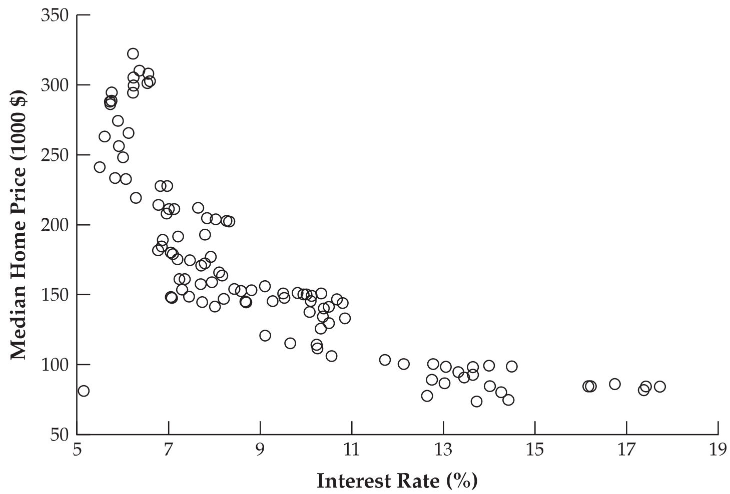

FIGURE 2-8 Median home prices and interest rates, 1980–2007

These results show that if the mortgage interest rate goes up by 1 percentage point,[10] on average, the median home price goes down by about 17.4 units or about $17,400. (*Note*: *Y* is measured in thousands of dollars.) Literally interpreted, the intercept coefficient of about 329 would suggest that if the mortgage interest rate were zero, the median home price on average would be about $329,000, an interpretation that may stretch our credulity.

It seems that falling interest rates do have a substantial impact on home prices. *A question*: If we had taken median family income into account, would this conclusion still stand?

Example 2.5. Antique Clocks and Their Prices

The Triberg Clock Company of Schonachbach, Germany, holds an annual antique clock auction. Data on about 32 clocks (the age of the clock, the number of bidders, and the price of the winning bid in marks) are given in Table 2-14 in Problem 2.19. Note that this auction took place about 25 years ago.

If we believe that the price of the winning bid depends on the age of the clock—the older the clock, the higher the price, *ceteris paribus*—we would expect a positive relationship between the two. Similarly, the higher the number of bidders, the higher the auction price because a large number of bidders for a particular clock would suggest that that clock is more valuable, and hence we would expect a positive relationship between the two variables.

[10]Note that there is a difference between a 1 percentage point increase and a 1 percent increase. For example, if the current interest rate is 6 percent but then goes to 7 percent, this represents a 1 percentage point increase; the percentage increase is, however, $\left(\frac{7-6}{6}\right) \times 100 = 16.6\%$.

Using the data given in Table 2-14, we obtained the following OLS regressions:

$$\text{Price} = -191.6662 + 10.4856\ \text{Age} \quad \textbf{(2.27)}$$

$$\text{Price} = 807.9501 + 54.5724\ \text{Bidders} \quad \textbf{(2.28)}$$

As these results show, the auction price is positively related to the age of the clock, as well as to the number of bidders present at the auction.

In Chapter 4 on multiple regression we will see what happens when we regress price on age and number of bidders together, rather than individually, as in the preceding two regressions.

The regression results presented in the preceding examples can be obtained easily by applying the OLS formulas Eq. (2.16) and Eq. (2.17) to the data presented in the various tables. Of course, this would be very tedious and very time-consuming to do manually. Fortunately, there are several statistical software packages that can estimate regressions in practically no time. In this book we will use the EViews and MINITAB software packages to estimate several regression models because these packages are comprehensive, easy to use, and readily available. (Excel can also do simple and multiple regressions.) *Throughout this book, we will reproduce the computer output obtained from these packages*. But keep in mind that there are other software packages that can estimate all kinds of regression models. Some of these packages are LIMDEP, MICROFIT, PC-GIVE, RATS, SAS, SHAZAM, SPSS, and STATA.

2.11 SUMMARY

In this chapter we introduced some fundamental ideas of regression analysis. Starting with the key concept of the population regression function (PRF), we developed the concept of linear PRF. This book is primarily concerned with linear PRFs, that is, regressions that are *linear in the parameters* regardless of whether or not they are linear in the variables. We then introduced the idea of the stochastic PRF and discussed in detail the nature and role of the stochastic error term u. PRF is, of course, a theoretical or idealized construct because, in practice, all we have is a sample(s) from some population. This necessitated the discussion of the sample regression function (SRF).

We then considered the question of how we actually go about obtaining the SRF. Here we discussed the popular method of ordinary least squares (OLS) and presented the appropriate formulas to estimate the parameters of the PRF. We illustrated the OLS method with a fully worked-out numerical example as well as with several practical examples.

Our next task is to find out how good the SRF obtained by OLS is as an estimator of the true PRF. We undertake this important task in Chapter 3.

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this chapter are

Regression analysis
- **a)** explained, or dependent, variable
- **b)** independent, or explanatory, variable

Scattergram; scatter diagram
Population regression line (PRL)
- **a)** conditional mean, or conditional expected, values

Population regression function (PRF)
Regression coefficients; parameters
- **a)** intercept
- **b)** slope

Conditional regression analysis
Stochastic, or random, error term; error term
- **a)** noise component
- **b)** stochastic, or statistical, PRF
- **c)** deterministic, or nonstochastic, PRF

Sample regression line (SRL)
Sample regression function (SRF)
Estimator; sample statistic
Estimate
Residual term e; residual
Linearity in variables
Linearity in parameters
- **a)** linear regression

Two-variable, or simple, regression vs. multiple linear regression
Estimation of parameters
- **a)** the method of ordinary least squares (OLS)
- **b)** the least squares principle
- **c)** residual sum of squares (RSS)
- **d)** normal equations
- **e)** OLS estimators

QUESTIONS

2.1. Explain carefully the meaning of each of the following terms:
- **a.** Population regression function (PRF).
- **b.** Sample regression function (SRF).
- **c.** Stochastic PRF.
- **d.** Linear regression model.
- **e.** Stochastic error term (u_i).
- **f.** Residual term (e_i).
- **g.** Conditional expectation.
- **h.** Unconditional expectation.
- **i.** Regression coefficients or parameters.
- **j.** Estimators of regression coefficients.

2.2. What is the difference between a stochastic population regression function (PRF) and a stochastic sample regression function (SRF)?

2.3. Since we do not observe the PRF, why bother studying it? Comment on this statement.

2.4. State whether the following statements are true, false, or uncertain. Give your reasons. Be precise.
- **a.** The stochastic error term u_i and the residual term e_i mean the same thing.
- **b.** The PRF gives the value of the dependent variable corresponding to each value of the independent variable.
- **c.** A linear regression model means a model linear in the variables.

d. In the linear regression model the explanatory variable is the cause and the dependent variable is the effect.
e. The conditional and unconditional mean of a random variable are the same thing.
f. In Eq. (2.2) the regression coefficients, the B's, are random variables, whereas the b's in Eq. (2.4) are the parameters.
g. In Eq. (2.1) the slope coefficient B_2 measures the slope of Y per unit change in X.
h. In practice, the two-variable regression model is useless because the behavior of a dependent variable can never be explained by a single explanatory variable.
i. The sum of the deviation of a random variable from its mean value is *always* equal to zero.

2.5. What is the relationship between
a. B_1 and b_1; **b.** B_2 and b_2; and **c.** u_i and e_i? Which of these entities can be observed and how?

2.6. Can you rewrite Eq. (2.22) to express X as a function of Y? How would you interpret the converted equation?

2.7. The following table gives pairs of dependent and independent variables. In each case state whether you would expect the relationship between the two variables to be positive, negative, or uncertain. In other words, tell whether the slope coefficient will be positive, negative, or neither. Give a brief justification in each case.

Dependent variable	Independent variable
(*a*) GDP	Rate of interest
(*b*) Personal savings	Rate of interest
(*c*) Yield of crop	Rainfall
(*d*) U.S. defense expenditure	Soviet Union's defense expenditure
(*e*) Number of home runs hit by a star baseball player	Annual salary
(*f*) A president's popularity	Length of stay in office
(*g*) A student's first-year grade-point average	S.A.T. score
(*h*) A student's grade in econometrics	Grade in statistics
(*i*) Imports of Japanese cars	U.S. per capita income

PROBLEMS

2.8. State whether the following models are linear regression models:
a. $Y_i = B_1 + B_2(1/X_i)$
b. $Y_i = B_1 + B_2 \ln X_i + u_i$
c. $\ln Y_i = B_1 + B_2 X_i + u_i$
d. $\ln Y_i = B_1 + B_2 \ln X_i + u_i$
e. $Y_i = B_1 + B_2 B_3 X_i + u_i$
f. $Y_i = B_1 + B_2^3 X_i + u_i$
Note: ln stands for the natural log, that is, log to the base e. (More on this in Chapter 4.)

2.9. Table 2-8 gives data on weekly family consumption expenditure (Y) (in dollars) and weekly family income (X) (in dollars).

TABLE 2-8 HYPOTHETICAL DATA ON WEEKLY CONSUMPTION EXPENDITURE AND WEEKLY INCOME

Weekly income ($)($X$)	Weekly consumption expenditure ($) ($Y$)
80	55, 60, 65, 70, 75
100	65, 70, 74, 80, 85, 88
120	79, 84, 90, 94, 98
140	80, 93, 95, 103, 108, 113, 115
160	102, 107, 110, 116, 118, 125
180	110, 115, 120, 130, 135, 140
200	120, 136, 140, 144, 145
220	135, 137, 140, 152, 157, 160, 162
240	137, 145, 155, 165, 175, 189
260	150, 152, 175, 178, 180, 185, 191

a. For each income level, compute the mean consumption expenditure, $E(Y \mid X_i)$, that is, the conditional expected value.
b. Plot these data in a scattergram with income on the horizontal axis and consumption expenditure on the vertical axis.
c. Plot the conditional means derived in part (*a*) in the same scattergram created in part (*b*).
d. What can you say about the relationship between Y and X and between mean Y and X?
e. Write down the PRF and the SRF for this example.
f. Is the PRF linear or nonlinear?

2.10. From the data given in the preceding problem, a random sample of Y was drawn against each X. The result was as follows:

Y	70	65	90	95	110	115	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260

a. Draw the scattergram with Y on the vertical axis and X on the horizontal axis.
b. What can you say about the relationship between Y and X?
c. What is the SRF for this example? Show all your calculations in the manner of Table 2-4.
d. On the same diagram, show the SRF and PRF.
e. Are the PRF and SRF identical? Why or why not?

2.11. Suppose someone has presented the following regression results for your consideration:

$$\hat{Y}_t = 2.6911 - 0.4795X_t$$

where Y = coffee consumption in the United States (cups per person per day)
X = retail price of coffee ($ per pound)
t = time period

a. Is this a time series regression or a cross-sectional regression?
b. Sketch the regression line.

c. What is the interpretation of the intercept in this example? Does it make economic sense?
d. How would you interpret the slope coefficient?
e. Is it possible to tell what the true PRF is in this example?
f. The *price elasticity* of demand is defined as the percentage change in the quantity demanded for a percentage change in the price. Mathematically, it is expressed as

$$\text{Elasticity} = \text{Slope}\left(\frac{X}{Y}\right)$$

That is, elasticity is equal to the product of the slope and the ratio of X to Y, where X = the price and Y = the quantity. From the regression results presented earlier, can you tell what the price elasticity of demand for coffee is? If not, what additional information would you need to compute the price elasticity?

2.12. Table 2-9 gives data on the Consumer Price Index (CPI) for all items (1982–1984 = 100) and the Standard & Poor's (S&P) index of 500 common stock prices (base of index: 1941–1943 = 10).

TABLE 2-9 CONSUMER PRICE INDEX (CPI) AND S&P 500 INDEX (S&P), UNITED STATES, 1978–1989

Year	CPI	S&P
1978	65.2	96.02
1979	72.6	103.01
1980	82.4	118.78
1981	90.9	128.05
1982	96.5	119.71
1983	99.6	160.41
1984	103.9	160.46
1985	107.6	186.84
1986	109.6	236.34
1987	113.6	286.83
1988	118.3	265.79
1989	124.0	322.84

Source: Economic Report of the President, 1990, Table C-58, for CPI and Table C-93 for the S&P index.

a. Plot the data on a scattergram with the S&P index on the vertical axis and CPI on the horizontal axis.
b. What can you say about the relationship between the two indexes? What does economic theory have to say about this relationship?
c. Consider the following regression model:

$$(\text{S\&P})_t = B_1 + B_2\text{CPI}_t + u_t$$

Use the method of least squares to estimate this equation from the preceding data and interpret your results.
d. Do the results obtained in part (*c*) make economic sense?
e. Do you know why the S&P index dropped in 1988?

2.13. Table 2-10 gives data on the nominal interest rate (Y) and the inflation rate (X) for the year 1988 for nine industrial countries.

TABLE 2-10 NOMINAL INTEREST RATE (Y) AND INFLATION (X) IN NINE INDUSTRIAL COUNTRIES FOR THE YEAR 1988

Country	Y (%)	X (%)
Australia	11.9	7.7
Canada	9.4	4.0
France	7.5	3.1
Germany	4.0	1.6
Italy	11.3	4.8
Mexico	66.3	51.7
Switzerland	2.2	2.0
United Kingdom	10.3	6.8
United States	7.6	4.4

Source: Rudiger Dornbusch and Stanley Fischer, *Macroeconomics,* 5th ed., McGraw-Hill, New York, 1990, p. 652. The original data are from various issues of *International Financial Statistics,* published by the International Monetary Fund (IMF).

a. Plot these data with the interest rate on the vertical axis and the inflation rate on the horizontal axis. What does the scattergram reveal?
b. Do an OLS regression of Y on X. Present all your calculations.
c. If the real interest rate is to remain constant, what must be the relationship between the nominal interest rate and the inflation rate? That is, what must be the value of the slope coefficient in the regression of Y on X and that of the intercept? Do your results suggest that this is the case? For a theoretical discussion of the relationship among the nominal interest rate, the inflation rate, and the real interest rate, see any textbook on macroeconomics and look up the topic of the Fisher equation, named after the famous American economist, Irving Fisher.

2.14. The real exchange rate (RE) is defined as the nominal exchange rate (NE) times the ratio of the domestic price to foreign price. Thus, RE for the United States against UK is

$$\text{RE}_{\text{US}} = \text{NE}_{\text{US}}(\text{US}_{\text{CPI}}/\text{UK}_{\text{CPI}})$$

a. From the data given in Table 1-3 of Problem 1.7, compute RE_{US}.
b. Using a regression package you are familiar with, estimate the following regression:

$$\text{NE}_{\text{US}} = B_1 + B_2\,\text{RE}_{\text{US}} + u \tag{1}$$

c. A priori, what do you expect the relationship between the nominal and real exchange rates to be? You may want to read up on the purchasing power parity (PPP) theory from any text on international trade or macroeconomics.
d. Are the a priori expectations supported by your regression results? If not, what might be the reason?

*e. Run regression (1) in the following alternative form:

$$\ln NE_{US} = A_1 + A_2 \ln RE_{US} + u \qquad \textbf{(2)}$$

where ln stands for the natural logarithm, that is, log to the base *e*. Interpret the results of this regression. Are the results from regressions (1) and (2) qualitatively the same?

2.15. Refer to problem 2.12. In Table 2-11 we have data on CPI and the S&P 500 index for the years 1990 to 2007.

TABLE 2-11 CONSUMER PRICE INDEX (CPI) AND S&P 500 INDEX (S&P), UNITED STATES, 1990–2007

Year	CPI	S&P
1990	130.7	334.59
1991	136.2	376.18
1992	140.3	415.74
1993	144.5	451.41
1994	148.2	460.42
1995	152.4	541.72
1996	156.9	670.50
1997	160.5	873.43
1998	163.0	1085.50
1999	166.6	1327.33
2000	172.2	1427.22
2001	177.1	1194.18
2002	179.9	993.94
2003	184.0	965.23
2004	188.9	1130.65
2005	195.3	1207.23
2006	201.6	1310.46
2007	207.3	1477.19

Source: Economic Report of the President, 2008.

a. Repeat questions (*a*) to (*e*) from problem 2.12.
b. Do you see any difference in the estimated regressions?
c. Now combine the two sets of data and estimate the regression of the S&P index on the CPI.
d. Are there noticeable differences in the three regressions?

2.16. Table 2-12, found on the textbook's Web site, gives data on average starting pay (ASP), grade point average (GPA) scores (on a scale of 1 to 4), GMAT scores, annual tuition, percent of graduates employed at graduation, recruiter assessment score (5.0 highest), and percent of applicants accepted in the graduate business school for 47 well-regarded business schools in the United States for the year 2007–2008. *Note:* Northwestern University ranked 4th (in a tie with MIT and University of Chicago) but was removed from the data set because there was no information available about percent of applicants accepted.
a. Using a bivariate regression model, find out if GPA has any effect on ASP.
b. Using a suitable regression model, find out if GMAT scores have any relationship to ASP.

*Optional.

c. Does annual tuition have any relationship to ASP? How do you know? If there is a positive relationship between the two, does that mean it pays to go to the most expensive business school? Can you argue that a high-tuition business school means a high-quality MBA program? Why or why not?

d. Does the recruiter perception have any bearing on ASP?

2.17. Table 2-13 (found on the textbook's Web site) gives data on real GDP (Y) and civilian unemployment rate (X) for the United States for period 1960 to 2006.

a. Estimate Okun's law in the form of Eq. (2.22). Are the regression results similar to the ones shown in (2.22)? Does this suggest that Okun's law is universally valid?

b. Now regress percentage change in real GDP on change in the civilian unemployment rate and interpret your regression results.

c. If the unemployment rate remains unchanged, what is the expected (percent) rate of growth in real GDP? (Use the regression in [*b*]). How would you interpret this growth rate?

2.18. Refer to Example 2.3, for which the data are as shown in Table 2-6 (on the textbook's Web site).

a. Using a statistical package of your choice, confirm the regression results given in Eq. (2.24) and Eq. (2.25).

b. For both regressions, get the estimated values of Y (i.e., $\hat{Y}_i$) and compare them with the actual Y values in the sample. Also obtain the residual values, e_i. From this can you tell which is a better model, Eq. (2.24) or Eq. (2.25)?

2.19. Refer to Example 2.5 on antique clock prices. Table 2-14 gives the underlying data.

a. Plot clock prices against the age of the clock and against the number of bidders. Does this plot suggest that the linear regression models shown in Eq. (2.27) and Eq. (2.28) may be appropriate?

TABLE 2-14 AUCTION DATA ON PRICE, AGE OF CLOCK, AND NUMBER OF BIDDERS

Observations	Price	Age	Number of bidders	Observations	Price	Age	Number of bidders
1	1235	127	13	17	854	143	6
2	1080	115	12	18	1483	159	9
3	845	127	7	19	1055	108	14
4	1552	150	9	20	1545	175	8
5	1047	156	6	21	729	108	6
6	1979	182	11	22	1792	179	9
7	1822	156	12	23	1175	111	15
8	1253	132	10	24	1593	187	8
9	1297	137	9	25	1147	137	8
10	946	113	9	26	1092	153	6
11	1713	137	15	27	1152	117	13
12	1024	117	11	28	1336	126	10
13	2131	170	14	29	785	111	7
14	1550	182	8	30	744	115	7
15	1884	162	11	31	1356	194	5
16	2041	184	10	32	1262	168	7

b. Would it make any sense to plot the number of bidders against the age of the clock? What would such a plot reveal?

2.20. Refer to the math S.A.T. score example discussed in the text. Table 2-4 gives the necessary raw calculations to obtain the OLS estimators. Look at the columns Y (actual Y) and $\hat{Y}$ (estimated Y) values. Plot the two in a scattergram. What does the scattergram reveal? If you believe that the fitted model [Eq. (2.20)] is a "good" model, what should be the shape of the scattergram? In the next chapter we will see what we mean by a "good" model.

2.21. Table 2-15 (on the textbook's Web site) gives data on verbal and math S.A.T. scores for both males and females for the period 1972–2007.

a. You want to predict the male math score (Y) on the basis of the male verbal score (X). Develop a suitable linear regression model and estimate its parameters.

b. Interpret your regression results.

c. Reverse the roles of Y and X and regress the verbal score on the math score. Interpret this regression

d. Let a_2 be the slope coefficient in the regression of the math score on the verbal score and let b_2 be the slope coefficient of the verbal score on the math score. Multiply these two values. Compare the resulting value with the r^2 obtained from the regression of math score on verbal score or the r^2 value obtained from the regression of verbal score on math score. What conclusion can you draw from this exercise?

2.22. Table 2-16 (on the textbook's Web site) gives data on investment rate (ipergdp) and savings rate (spergdp), both measured as percent of GDP, for a cross-section of countries. These rates are averages for the period 1960–1974.*

a. Plot the investment rate on the vertical axis and the savings rate on the horizontal axis.

b. Eyeball a suitable curve from the scatter diagram in (*a*).

c. Now estimate the following model

$$\text{ipergdp}_i = B_1 + B_2\,\text{spergdp}_i + u_i$$

d. Interpret the estimated coefficients.

e. What general conclusion do you draw from your analysis?

Note: Save your results for further analysis in the next chapter.

OPTIONAL QUESTIONS

2.23. Prove that $\sum e_i = 0$, and hence show that $\bar{e} = 0$.

2.24. Prove that $\sum e_i x_i = 0$.

2.25. Prove that $\sum e_i \hat{Y}_i = 0$, that is, that the sum of the product of residuals e_i and the estimated Y_i is always zero.

2.26. Prove that $\bar{Y} = \bar{\hat{Y}}$, that is, that the means of the actual Y values and the estimated Y values are the same.

*Source of data: Martin Feldstein and Charles Horioka, "Domestic Savings and International Capital Flows," *Economic Journal*, vol. 90, June 1980, pp. 314–329.

2.27. Prove that $\sum x_i y_i = \sum x_i Y_i = \sum X_i y_i$, where $x_i = (X_i - \overline{X})$ and $y_i = (Y_i - \overline{Y})$.
2.28. Prove that $\sum x_i = \sum y_i = 0$, where x_i and y_i are as defined in Problem 2.27.
2.29. For the math S.A.T. score example data given in Table 2-4, verify that statements made in Question 2.23 hold true (save the rounding errors).

APPENDIX 2A: Derivation of Least-Squares Estimates

We start with Eq. (2.13):

$$\sum e_i^2 = \sum (Y_i - b_1 - b_2 X_1)^2 \qquad \textbf{(2A.1)}$$

Using the technique of *partial differentiation* from calculus, we obtain:

$$\partial \sum e_i^2 / \partial b_1 = 2 \sum (Y_i - b_1 - b_2 X_i)(-1) \qquad \textbf{(2A.2)}$$

$$\partial \sum e_i^2 / \partial b_2 = 2 \sum (Y_i - b_1 - b_2 X_i)(-X_i) \qquad \textbf{(2A.3)}$$

By the first order condition of optimization, we set these two derivations to zero and simplify, which will give

$$\sum Y_i = n b_1 + b_2 \sum X_i \qquad \textbf{(2A.4)}$$

$$\sum Y_i X_i = b_1 \sum X_i + b_2 \sum X_i^2 \qquad \textbf{(2A.5)}$$

which are Eqs. (2.14) and (2.15), respectively, given in the text.

Solving these two equations simultaneously, we get the formulas given in Eqs. (2.16) and (2.17).