2

TWO-VARIABLE REGRESSION ANALYSIS: SOME BASIC IDEAS

In Chapter 1 we discussed the concept of regression in broad terms. In this chapter we approach the subject somewhat formally. Specifically, this and the following two chapters introduce the reader to the theory underlying the simplest possible regression analysis, namely, the **bivariate,** or **two-variable,** regression in which the dependent variable (the regressand) is related to a single explanatory variable (the regressor). This case is considered first, not because of its practical adequacy, but because it presents the fundamental ideas of regression analysis as simply as possible and some of these ideas can be illustrated with the aid of two-dimensional graphs. Moreover, as we shall see, the more general **multiple** regression analysis in which the regressand is related to one or more regressors is in many ways a logical extension of the two-variable case.

2.1 A HYPOTHETICAL EXAMPLE[1]

As noted in Section 1.2, regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable(s).[2] To understand this, consider the data given in Table 2.1. The data

[1]The reader whose statistical knowledge has become somewhat rusty may want to freshen it up by reading the statistical appendix, **App. A,** before reading this chapter.

[2]The *expected value, or expectation, or population mean of a random variable Y* is denoted by the symbol $E(Y)$. On the other hand, the mean value computed from a sample of values from the Y population is denoted as $\bar{Y}$, read as Y bar.

TABLE 2.1 WEEKLY FAMILY INCOME X, $

$Y\downarrow$ / $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

in the table refer to a total **population** of 60 families in a hypothetical community and their weekly income (X) and weekly consumption expenditure (Y), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 *fixed* values of X and the corresponding Y values against each of the X values; so to speak, there are 10 Y subpopulations.

There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly from Figure 2.1. But the general picture that one gets is that, despite the variability of weekly consump-

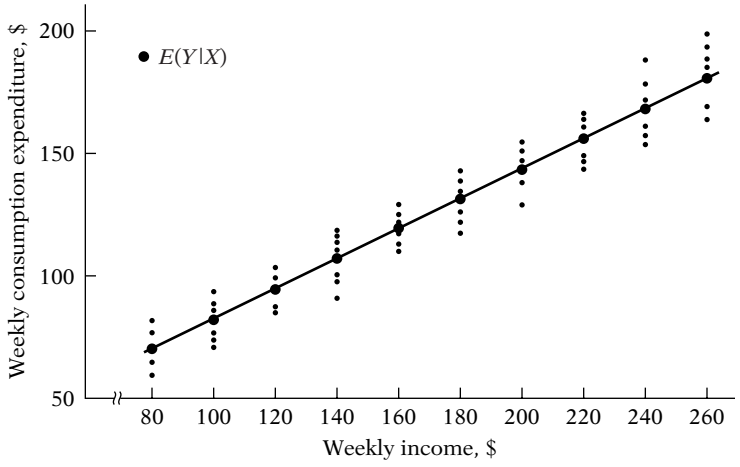

FIGURE 2.1 Conditional distribution of expenditure for various levels of income (data of Table 2.1).

tion expenditure within each income bracket, *on the average,* weekly consumption expenditure increases as income increases. To see this clearly, in Table 2.1 we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of $80, the mean consumption expenditure is $65, while corresponding to the income level of $200, it is $137. In all we have 10 mean values for the 10 subpopulations of *Y*. We call these mean values **conditional expected values,** as they depend on the given values of the (conditioning) variable *X*. Symbolically, we denote them as $E(Y \mid X)$, which is read as the expected value of *Y* given the value of *X* (see also Table 2.2).

It is important to distinguish these conditional expected values from the **unconditional expected value of** weekly consumption expenditure, $E(Y)$. If we add the weekly consumption expenditures for all the 60 families in the *population* and divide this number by 60, we get the number $121.20 ($7272/60), which is the unconditional mean, or expected, value of weekly consumption expenditure, $E(Y)$; it is unconditional in the sense that in arriving at this number we have disregarded the income levels of the various families.[3] Obviously, the various conditional expected values of *Y* given in Table 2.1 are different from the unconditional expected value of *Y* of $121.20. When we ask the question, "What is the *expected value* of weekly consumption expenditure of a family," we get the answer $121.20 (the unconditional mean). But if we ask the question, "What is the *expected value* of weekly consumption expenditure of a family whose monthly income is,

TABLE 2.2 CONDITIONAL PROBABILITIES $p(Y \mid X_i)$ FOR THE DATA OF TABLE 2.1

$p(Y \mid X_i)$ ↓ \ X→	80	100	120	140	160	180	200	220	240	260
Conditional probabilities $p(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	–	–	$\frac{1}{7}$	–	–	–	$\frac{1}{7}$	–	$\frac{1}{7}$
Conditional means of *Y*	65	77	89	101	113	125	137	149	161	173

[3]As shown in **App. A**, in general the conditional and unconditional mean values are different.

say, $140," we get the answer $101 (the conditional mean). To put it differently, if we ask the question, "What is the best (mean) prediction of weekly expenditure of families with a weekly income of $140," the answer would be $101. Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure than if we do not have that knowledge.[4] This probably is the essence of regression analysis, as we shall discover throughout this text.

The dark circled points in Figure 2.1 show the conditional mean values of Y against the various X values. If we join these conditional mean values, we obtain what is known as the **population regression line (PRL),** or more generally, the **population regression curve.**[5] More simply, it is the **regression of *Y* on *X*.** The adjective "population" comes from the fact that we are dealing in this example with the entire population of 60 families. Of course, in reality a population may have many families.

Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More simply, it is the curve connecting the means of the subpopulations of Y corresponding to the given values of the regressor X. It can be depicted as in Figure 2.2.

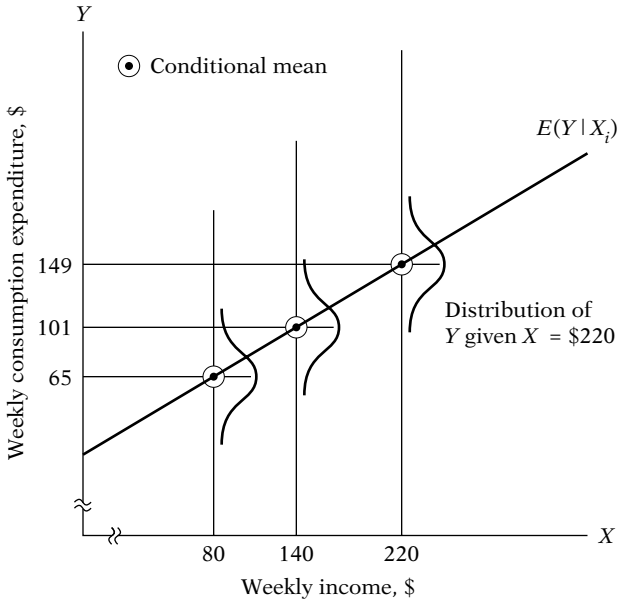

FIGURE 2.2 Population regression line (data of Table 2.1).

[4]I am indebted to James Davidson on this perspective. See James Davidson, *Econometric Theory,* Blackwell Publishers, Oxford, U.K., 2000, p. 11.

[5]In the present example the PRL is a straight line, but it could be a curve (see Figure 2.3).

This figure shows that for each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values. For simplicity, we are assuming that these Y values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

With this background, the reader may find it instructive to reread the definition of regression given in Section 1.2.

2.2 THE CONCEPT OF POPULATION REGRESSION FUNCTION (PRF)

From the preceding discussion and Figures. 2.1 and 2.2, it is clear that each conditional mean $E(Y \mid X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically,

$$E(Y \mid X_i) = f(X_i) \tag{2.2.1}$$

where $f(X_i)$ denotes some function of the explanatory variable X. In our example, $E(Y \mid X_i)$ is a linear function of X_i. Equation (2.2.1) is known as the **conditional expectation function (CEF)** or **population regression function (PRF)** or **population regression (PR)** for short. It states merely that the *expected value* of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X.

What form does the function $f(X_i)$ assume? This is an important question because in real situations we do not have the entire population available for examination. The functional form of the PRF is therefore an empirical question, although in specific cases theory may have something to say. For example, an economist might posit that consumption expenditure is linearly related to income. Therefore, as a first approximation or a working hypothesis, we may assume that the PRF $E(Y \mid X_i)$ is a linear function of X_i, say, of the type

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \tag{2.2.2}$$

where β_1 and β_2 are unknown but fixed parameters known as the **regression coefficients;** β_1 and β_2 are also known as **intercept** and **slope coefficients,** respectively. Equation (2.2.1) itself is known as the **linear population regression function.** Some alternative expressions used in the literature are *linear population regression model* or simply *linear population regression*. In the sequel, the terms **regression, regression equation,** and **regression model** will be used synonymously.

In regression analysis our interest is in estimating the PRFs like (2.2.2), that is, estimating the values of the unknowns β_1 and β_2 on the basis of observations on Y and X. This topic will be studied in detail in Chapter 3.

2.3 THE MEANING OF THE TERM *LINEAR*

Since this text is concerned primarily with linear models like (2.2.2), 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 more "natural" meaning of linearity is that the conditional expectation of Y is a linear function of X_i, such as, for example, (2.2.2).[6] Geometrically, the regression curve in this case is a straight line. In this interpretation, a regression function such as $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is not a linear function because the variable X appears with a power or index of 2.

Linearity in the Parameters

The second interpretation of linearity is that the conditional expectation of Y, $E(Y \mid X_i)$, is a linear function of the parameters, the β's; it may or may not be linear in the variable X.[7] In this interpretation $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is a linear (in the parameter) regression model. To see this, let us suppose X takes the value 3. Therefore, $E(Y \mid X = 3) = \beta_1 + 9\beta_2$, which is obviously linear in β_1 and β_2. All the models shown in Figure 2.3 are thus linear regression models, that is, models linear in the parameters.

Now consider the model $E(Y \mid X_i) = \beta_1 + \beta_2^2 X_i$. Now suppose $X = 3$; then we obtain $E(Y \mid X_i) = \beta_1 + 3\beta_2^2$, which is nonlinear in the parameter β_2. The preceding model is an example of a **nonlinear (in the parameter) regression model.** We will discuss such models in Chapter 14.

Of the two interpretations of linearity, linearity in the parameters is relevant for the development of the regression theory to be presented shortly. Therefore, *from now on the term "linear" regression will always mean a regression that is linear in the parameters; the β's (that is, the parameters are raised to the first power only). It may or may not be linear in the explanatory variables, the X's.* Schematically, we have Table 2.3. Thus, $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$, which is linear both in the parameters and variable, is a LRM, and so is $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$, which is linear in the parameters but nonlinear in variable X.

---

[6] A function $Y = f(X)$ is said to be linear in X if X appears with a power or index of 1 only (that is, terms such as X^2, $\sqrt{X}$, and so on, are excluded) and is not multiplied or divided by any other variable (for example, $X \cdot Z$ or X/Z, where Z is another variable). If Y depends on X alone, another way to state that Y is linearly related to X is that the rate of change of Y with respect to X (i.e., the slope, or derivative, of Y with respect to X, dY/dX) is independent of the value of X. Thus, if $Y = 4X$, $dY/dX = 4$, which is independent of the value of X. But if $Y = 4X^2$, $dY/dX = 8X$, which is not independent of the value taken by X. Hence this function is not linear in X.

[7] A function is said to be linear in the parameter, say, β_1, if β_1 appears with a power of 1 only and is not multiplied or divided by any other parameter (for example, $\beta_1\beta_2$, β_2/β_1, and so on).

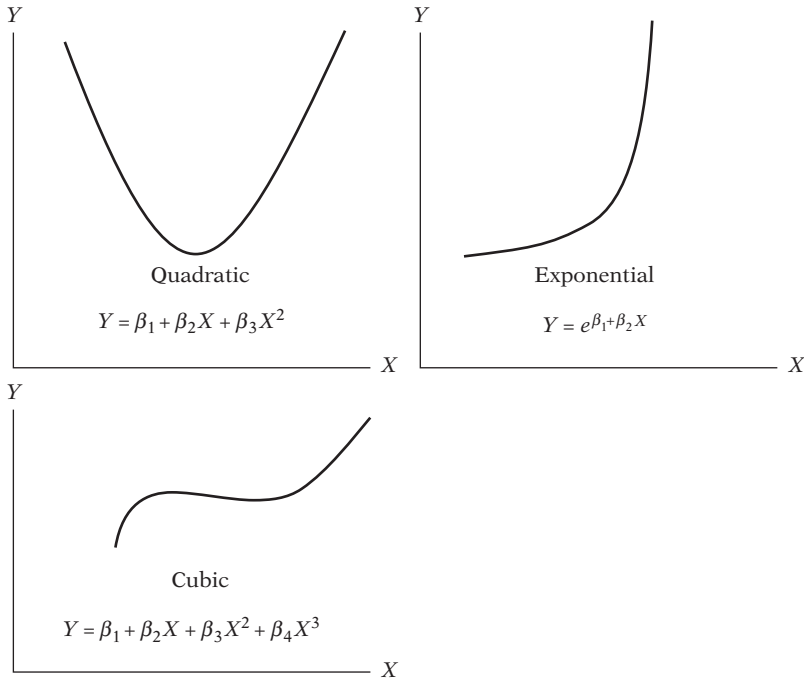

FIGURE 2.3 Linear-in-parameter functions.

TABLE 2.3 LINEAR REGRESSION MODELS

Model linear in parameters?	Model linear in variables?	
	Yes	No
Yes	LRM	LRM
No	NLRM	NLRM

Note: LRM = linear regression model
NLRM = nonlinear regression model

2.4 STOCHASTIC SPECIFICATION OF PRF

It is clear from Figure 2.1 that, as family income increases, family consumption expenditure on the average increases, too. But what about the consumption expenditure of an individual family in relation to its (fixed) level of income? It is obvious from Table 2.1 and Figure 2.1 that an individual family's consumption expenditure does not necessarily increase as the income level increases. For example, from Table 2.1 we observe that corresponding to the income level of $100 there is one family whose consumption expenditure of $65 is less than the consumption expenditures of two families whose weekly income is only $80. But notice that the *average* consumption

expenditure of families with a weekly income of $100 is greater than the average consumption expenditure of families with a weekly income of $80 ($77 versus $65).

What, then, can we say about the relationship between an individual family's consumption expenditure and a given level of income? We see from Figure 2.1 that, given the income level of X_i, an individual family's consumption expenditure is clustered around the average consumption of all families at that X_i, that is, around its conditional expectation. Therefore, we can express the *deviation* of an individual Y_i around its expected value as follows:

$$u_i = Y_i - E(Y \mid X_i)$$

or

$$Y_i = E(Y \mid X_i) + u_i \tag{2.4.1}$$

where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is known as the **stochastic disturbance** or **stochastic error term.**

How do we interpret (2.4.1)? We can say that the expenditure of an individual family, given its income level, can be expressed as the sum of two components: (1) $E(Y \mid X_i)$, which is simply the mean consumption expenditure of all the families with the same level of income. This component is known as the **systematic,** or **deterministic,** component, and (2) u_i, which is the random, or **nonsystematic,** component. We shall examine shortly the nature of the stochastic disturbance term, but for the moment assume that it is a *surrogate or proxy* for all the omitted or neglected variables that may affect Y but are not (or cannot be) included in the regression model.

If $E(Y \mid X_i)$ is assumed to be linear in X_i, as in (2.2.2), Eq. (2.4.1) may be written as

$$\begin{aligned} Y_i &= E(Y \mid X_i) + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned} \tag{2.4.2}$$

Equation (2.4.2) posits that the consumption expenditure of a family is linearly related to its income plus the disturbance term. Thus, the individual consumption expenditures, given $X = \$80$ (see Table 2.1), can be expressed as

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1 \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2 \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3 \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4 \\ Y_5 &= 75 = \beta_1 + \beta_2(80) + u_5 \end{aligned} \tag{2.4.3}$$

Now if we take the expected value of (2.4.1) on both sides, we obtain

$$\begin{aligned} E(Y_i \mid X_i) &= E[E(Y \mid X_i)] + E(u_i \mid X_i) \\ &= E(Y \mid X_i) + E(u_i \mid X_i) \end{aligned} \tag{2.4.4}$$

where use is made of the fact that the expected value of a constant is that constant itself.[8] Notice carefully that in (2.4.4) we have taken the conditional expectation, conditional upon the given X's.

Since $E(Y_i \mid X_i)$ is the same thing as $E(Y \mid X_i)$, Eq. (2.4.4) implies that

$$E(u_i \mid X_i) = 0 \tag{2.4.5}$$

Thus, the assumption that the regression line passes through the conditional means of Y (see Figure 2.2) implies that the conditional mean values of u_i (conditional upon the given X's) are zero.

From the previous discussion, it is clear (2.2.2) and (2.4.2) are equivalent forms if $E(u_i \mid X_i) = 0$.[9] But the stochastic specification (2.4.2) has the advantage that it clearly shows that there are other variables besides income that affect consumption expenditure and that an individual family's consumption expenditure cannot be fully explained only by the variable(s) included in the regression model.

2.5 THE SIGNIFICANCE OF THE STOCHASTIC DISTURBANCE TERM

As noted in Section 2.4, the disturbance term u_i is a surrogate for all those variables that are omitted from the model but that collectively affect Y. The obvious question is: Why not introduce these variables into the model explicitly? Stated otherwise, why not develop a multiple regression model with as many variables as possible? The reasons are many.

1. *Vagueness of theory:* The theory, if any, determining the behavior of Y may be, and often is, incomplete. We might know for certain that weekly income X influences weekly consumption expenditure Y, but we might be ignorant or unsure about the other variables affecting Y. Therefore, u_i may be used as a substitute for all the excluded or omitted variables from the model.

2. *Unavailability of data:* Even if we know what some of the excluded variables are and therefore consider a multiple regression rather than a simple regression, we may not have quantitative information about these

[8] See **App. A** for a brief discussion of the properties of the expectation operator E. Note that $E(Y \mid X_i)$, once the value of X_i is fixed, is a constant.

[9] As a matter of fact, in the method of least squares to be developed in Chap. 3, it is assumed explicitly that $E(u_i \mid X_i) = 0$. See Sec. 3.2.

variables. It is a common experience in empirical analysis that the data we would ideally like to have often are not available. For example, in principle we could introduce family wealth as an explanatory variable in addition to the income variable to explain family consumption expenditure. But unfortunately, information on family wealth generally is not available. Therefore, we may be forced to omit the wealth variable from our model despite its great theoretical relevance in explaining consumption expenditure.

3. *Core variables versus peripheral variables:* Assume in our consumption-income example that besides income X_1, the number of children per family X_2, sex X_3, religion X_4, education X_5, and geographical region X_6 also affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small and at best nonsystematic or random that as a practical matter and for cost considerations it does not pay to introduce them into the model explicitly. One hopes that their combined effect can be treated as a random variable u_i.[10]

4. *Intrinsic randomness in human behavior:* Even if we succeed in introducing all the relevant variables into the model, there is bound to be some "intrinsic" randomness in individual Y's that cannot be explained no matter how hard we try. The disturbances, the u's, may very well reflect this intrinsic randomness.

5. *Poor proxy variables:* Although the classical regression model (to be developed in Chapter 3) assumes that the variables Y and X are measured accurately, in practice the data may be plagued by errors of measurement. Consider, for example, Milton Friedman's well-known theory of the consumption function.[11] He regards *permanent consumption* (Y^p) as a function of *permanent income* (X^p). But since data on these variables are not directly observable, in practice we use proxy variables, such as current consumption (Y) and current income (X), which can be observable. Since the observed Y and X may not equal Y^p and X^p, there is the problem of errors of measurement. The disturbance term u may in this case then also represent the errors of measurement. As we will see in a later chapter, if there are such errors of measurement, they can have serious implications for estimating the regression coefficients, the β's.

6. *Principle of parsimony:* Following Occam's razor,[12] we would like to keep our regression model as simple as possible. If we can explain the behavior of Y "substantially" with two or three explanatory variables and if

[10]A further difficulty is that variables such as sex, education, and religion are difficult to quantify.

[11]Milton Friedman, *A Theory of the Consumption Function,* Princeton University Press, Princeton, N.J., 1957.

[12]"That descriptions be kept as simple as possible until proved inadequate," *The World of Mathematics,* vol. 2, J. R. Newman (ed.), Simon & Schuster, New York, 1956, p. 1247, or, "Entities should not be multiplied beyond necessity," Donald F. Morrison, *Applied Linear Statistical Methods,* Prentice Hall, Englewood Cliffs, N.J., 1983, p. 58.

our theory is not strong enough to suggest what other variables might be included, why introduce more variables? Let u_i represent all other variables. Of course, we should not exclude relevant and important variables just to keep the regression model simple.

7. *Wrong functional form:* Even if we have theoretically correct variables explaining a phenomenon and even if we can obtain data on these variables, very often we do not know the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (invariable) function of income or a nonlinear (invariable) function? If it is the former, $Y_i = \beta_1 + B_2 X_i + u_i$ is the proper functional relationship between Y and X, but if it is the latter, $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ may be the correct functional form. In two-variable models the functional form of the relationship can often be judged from the scattergram. But in a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize scattergrams in multiple dimensions.

For all these reasons, the stochastic disturbances u_i assume an extremely critical role in regression analysis, which we will see as we progress.

2.6 THE SAMPLE REGRESSION FUNCTION (SRF)

By confining our discussion so far to the population of Y values corresponding to the fixed X's, we have deliberately avoided sampling considerations (note that the data of Table 2.1 represent the population, not a sample). But it is about time to face up to the sampling problems, for in most practical situations what we have is but a sample of Y values corresponding to some fixed X's. Therefore, our task now is to estimate the PRF on the basis of the sample information.

As an illustration, pretend that the population of Table 2.1 was not known to us and the only information we had was a randomly selected sample of Y values for the fixed X's as given in Table 2.4. Unlike Table 2.1, we now have only one Y value corresponding to the given X's; each Y (given X_i) in Table 2.4 is chosen randomly from similar Y's corresponding to the same X_i from the population of Table 2.1.

The question is: From the sample of Table 2.4 can we predict the average weekly consumption expenditure Y in the population as a whole corresponding to the chosen X's? In other words, can we estimate the PRF from the sample data? As the reader surely suspects, we may not be able to estimate the PRF "accurately" because of sampling fluctuations. To see this, suppose we draw another random sample from the population of Table 2.1, as presented in Table 2.5.

Plotting the data of Tables 2.4 and 2.5, we obtain the scattergram given in Figure 2.4. In the scattergram two sample regression lines are drawn so as

TABLE 2.4
A RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

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

TABLE 2.5
ANOTHER RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

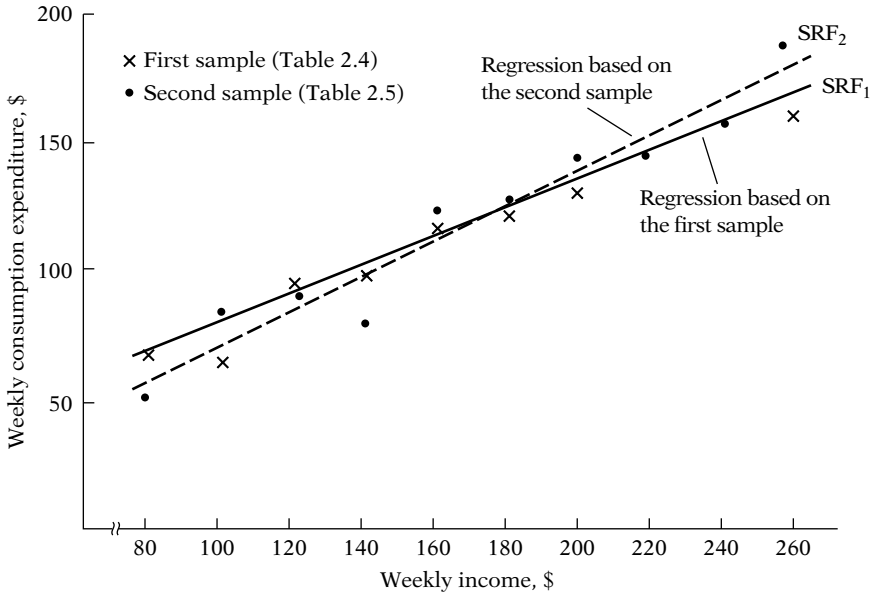

FIGURE 2.4 Regression lines based on two different samples.

to "fit" the scatters reasonably well: SRF_1 is based on the first sample, and SRF_2 is based on the second sample. Which of the two regression lines represents the "true" population regression line? If we avoid the temptation of looking at Figure 2.1, which purportedly represents the PR, there is no way we can be absolutely sure that either of the regression lines shown in Figure 2.4 represents the true population regression line (or curve). The regression lines in Figure 2.4 are known as the **sample regression lines.** Sup-

posedly they represent the population regression line, but because of sampling fluctuations they are at best an approximation of the true PR. In general, we would get N different SRFs for N different samples, and these SRFs are not likely to be the same.

Now, analogously to the PRF that underlies the population regression line, we can develop the concept of the **sample regression function** (SRF) to represent the sample regression line. The sample counterpart of (2.2.2) may be written as

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \tag{2.6.1}$$

where $\hat{Y}$ is read as "Y-hat" or "Y-cap"
$\hat{Y}_i$ = estimator of $E(Y \mid X_i)$
$\hat{\beta}_1$ = estimator of β_1
$\hat{\beta}_2$ = estimator of β_2

Note that an **estimator,** also known as a (sample) **statistic,** is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator in an application is known as an **estimate.**[13]

Now just as we expressed the PRF in two equivalent forms, (2.2.2) and (2.4.2), we can express the SRF (2.6.1) in its stochastic form as follows:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

where, in addition to the symbols already defined, $\hat{u}_i$ denotes the (sample) **residual** term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an *estimate* of u_i. It is introduced in the SRF for the same reasons as u_i was introduced in the PRF.

To sum up, then, we find our primary objective in regression analysis is to estimate the PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

on the basis of the SRF

$$Y_i = \hat{\beta}_1 + \hat{\beta} x_i = \hat{u}_i \tag{2.6.2}$$

because more often than not our analysis is based upon a single sample from some population. But because of sampling fluctuations our estimate of

[13]As noted in the Introduction, a hat above a variable will signify an estimator of the relevant population value.

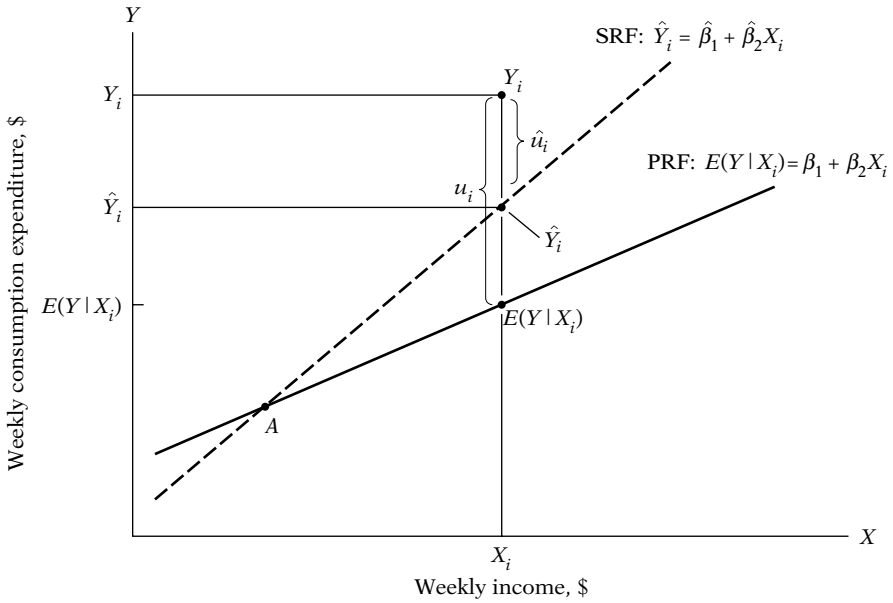

FIGURE 2.5 Sample and population regression lines.

the PRF based on the SRF is at best an approximate one. This approximation is shown diagrammatically in Figure 2.5.

For $X = X_i$, we have one (sample) observation $Y = Y_i$. In terms of the SRF, the observed Y_i can be expressed as

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

and in terms of the PRF, it can be expressed as

$$Y_i = E(Y \mid X_i) + u_i \tag{2.6.4}$$

Now obviously in Figure 2.5 $\hat{Y}_i$ *overestimates* the true $E(Y \mid X_i)$ for the X_i shown therein. By the same token, for any X_i to the left of the point *A*, the SRF will *underestimate* the true PRF. But the reader can readily see that such over- and underestimation is inevitable because of sampling fluctuations.

The critical question now is: Granted that the SRF is but an approximation of the PRF, can we devise a rule or a method that will make this approximation as "close" as possible? In other words, how should the SRF be constructed so that $\hat{\beta}_1$ is as "close" as possible to the true β_1 and $\hat{\beta}_2$ is as "close" as possible to the true β_2 even though we will never know the true β_1 and β_2?

The answer to this question will occupy much of our attention in Chapter 3. We note here that we can develop procedures that tell us how to construct the SRF to 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.

2.7 AN ILLUSTRATIVE EXAMPLE

We conclude this chapter with an example. Table 2.6 gives data on the level of education (measured by the number of years of schooling), the mean hourly wages earned by people at each level of education, and the number of people at the stated level of education. Ernst Berndt originally obtained the data presented in the table, and he derived these data from the current population survey conducted in May 1985.[14] We will explore these data (with additional explanatory variables) in Chapter 3 (Example 3.3, p. 91).

Plotting the (conditional) mean wage against education, we obtain the picture in Figure 2.6. The regression curve in the figure shows how mean wages vary with the level of education; they generally increase with the level of education, a finding one should not find surprising. We will study in a later chapter how variables besides education can also affect the mean wage.

TABLE 2.6
MEAN HOURLY WAGE BY EDUCATION

Years of schooling	Mean 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
	Total	528

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

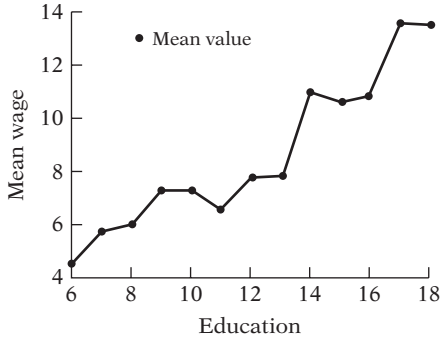

FIGURE 2.6
Relationship between mean wages and education.

[14]Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary,* Addison Wesley, Reading, Mass., 1991. Incidentally, this is an excellent book that the reader may want to read to find out how econometricians go about doing research.

2.8 SUMMARY AND CONCLUSIONS

1. The key concept underlying regression analysis is the concept of the **conditional expectation function (CEF), or population regression function (PRF).** Our objective in regression analysis is to find out how the average value of the dependent variable (or regressand) varies with the given value of the explanatory variable (or regressor).

2. This book largely deals with **linear PRFs,** that is, regressions that are linear in the parameters. They may or may not be linear in the regressand or the regressors.

3. For empirical purposes, it is the **stochastic PRF** that matters. The **stochastic disturbance term** u_i plays a critical role in estimating the PRF.

4. The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Usually, one has a sample of observations from the population. Therefore, one uses the **stochastic sample regression function (SRF)** to estimate the PRF. How this is actually accomplished is discussed in Chapter 3.

EXERCISES

Questions

2.1. What is the conditional expectation function or the population regression function?

2.2. What is the difference between the population and sample regression functions? Is this a distinction without difference?

2.3. What is the role of the stochastic error term u_i in regression analysis? What is the difference between the stochastic error term and the residual, $\hat{u}_i$?

2.4. Why do we need regression analysis? Why not simply use the mean value of the regressand as its best value?

2.5. What do we mean by a *linear* regression model?

2.6. Determine whether the following models are linear in the parameters, or the variables, or both. Which of these models are linear regression models?

Model	Descriptive title
a. $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Reciprocal
b. $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	Semilogarithmic
c. $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	Inverse semilogarithmic
d. $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$	Logarithmic or double logarithmic
e. $\ln Y_i = \beta_1 - \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Logarithmic reciprocal

Note: ln = natural log (i.e., log to the base e); u_i is the stochastic disturbance term. We will study these models in Chapter 6.

2.7. Are the following models linear regression models? Why or why not?

a. $Y_i = e^{\beta_1 + \beta_2 X_i + u_i}$

b. $Y_i = \dfrac{1}{1 + e^{\beta_1 + \beta_2 X_i + u_i}}$

c. $\ln Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$

d. $Y_i = \beta_1 + (0.75 - \beta_1)e^{-\beta_2(X_i - 2)} + u_i$

e. $Y_i = \beta_1 + \beta_2^3 X_i + u_i$

2.8. What is meant by an *intrinsically linear* regression model? If β_2 in exercise 2.7**d** were 0.8, would it be a linear or nonlinear regression model?

***2.9.** Consider the following nonstochastic models (i.e., models without the stochastic error term). Are they linear regression models? If not, is it possible, by suitable algebraic manipulations, to convert them into linear models?

a. $Y_i = \dfrac{1}{\beta_1 + \beta_2 X_i}$

b. $Y_i = \dfrac{X_i}{\beta_1 + \beta_2 X_i}$

c. $Y_i = \dfrac{1}{1 + \exp(-\beta_1 - \beta_2 X_i)}$

2.10. You are given the scattergram in Figure 2.7 along with the regression line. What general conclusion do you draw from this diagram? Is the regression line sketched in the diagram a population regression line or the sample regression line?

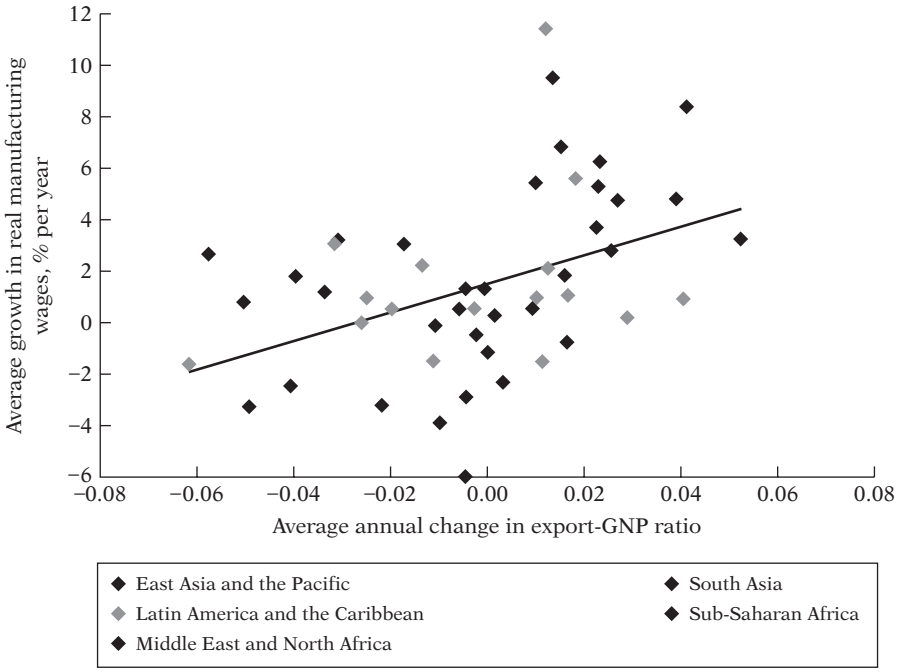

FIGURE 2.7 Growth rates of real manufacturing wages and exports. Data are for 50 developing countries during 1970–90.

Source: The World Bank, *World Development Report 1995,* p. 55. The original source is UNIDO data, World Bank data.

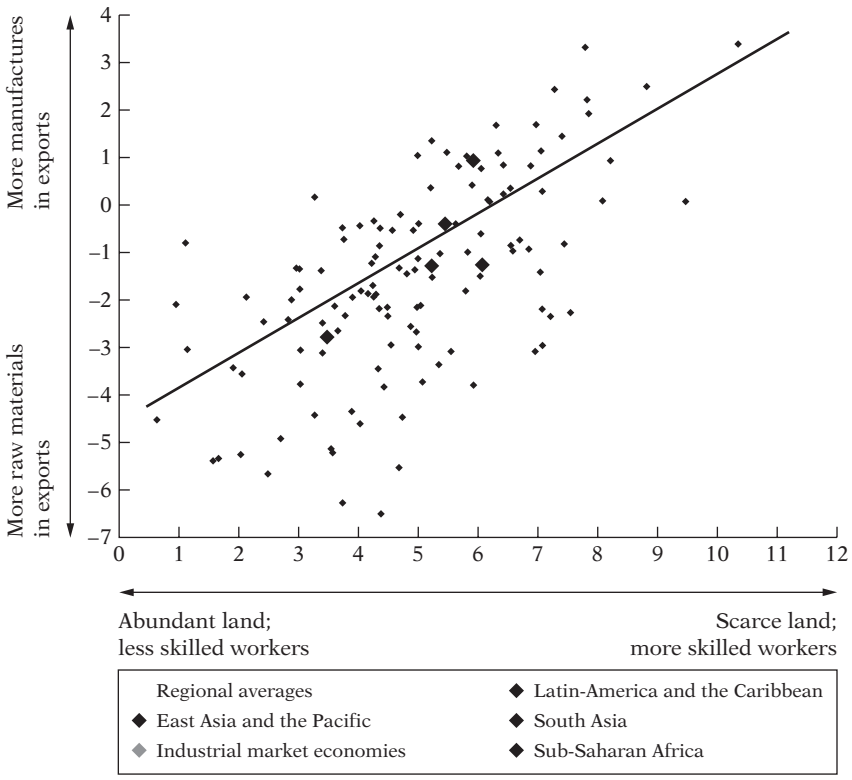

FIGURE 2.8 Skill intensity of exports and human capital endowment. Data are for 126 industrial and developing countries in 1985. Values along the horizontal axis are logarithms of the ratio of the country's average educational attainment to its land area: vertical axis values are logarithms of the ratio of manufactured to primary-products exports.

Source: World Bank, *World Development Report 1995,* p. 59. Original sources: Export data from United Nations Statistical Office COMTRADE data base; education data from UNDP 1990; land data from the World Bank.

2.11. From the scattergram given in Figure 2.8, what general conclusions do you draw? What is the economic theory that underlies this scattergram? (*Hint:* Look up any international economics textbook and read up on the Heckscher–Ohlin model of trade.)

2.12. What does the scattergram in Figure 2.9 reveal? On the basis of this diagram, would you argue that minimum wage laws are good for economic well-being?

2.13. Is the regression line shown in Figure I.3 of the Introduction the PRF or the SRF? Why? How would you interpret the scatterpoints around the regression line? Besides GDP, what other factors, or variables, might determine personal consumption expenditure?

Problems

2.14. You are given the data in Table 2.7 for the United States for years 1980–1996.

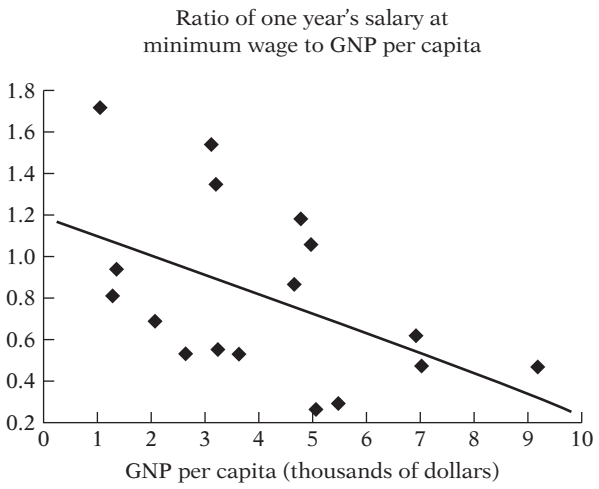

FIGURE 2.9 The minimum wage and GNP per capita. The sample consists of 17 developing countries. Years vary by country from 1988 to 1992. Data are in international prices.

Source: World Bank, *World Development Report 1995,* p. 75.

TABLE 2.7 LABOR FORCE PARTICIPATION DATA

Year	CLFPRM[1]	CLFPRF[2]	UNRM[3]	UNRF[4]	AHE82[5]	AHE[6]
1980	77.4	51.5	6.9	7.4	7.78	6.66
1981	77.0	52.1	7.4	7.9	7.69	7.25
1982	76.6	52.6	9.9	9.4	7.68	7.68
1983	76.4	53.9	9.9	9.2	7.79	8.02
1984	76.4	53.6	7.4	7.6	7.80	8.32
1985	76.3	54.5	7.0	7.4	7.77	8.57
1986	76.3	55.3	6.9	7.1	7.81	8.76
1987	76.2	56.0	6.2	6.2	7.73	8.98
1988	76.2	56.6	5.5	5.6	7.69	9.28
1989	76.4	57.4	5.2	5.4	7.64	9.66
1990	76.4	57.5	5.7	5.5	7.52	10.01
1991	75.8	57.4	7.2	6.4	7.45	10.32
1992	75.8	57.8	7.9	7.0	7.41	10.57
1993	75.4	57.9	7.2	6.6	7.39	10.83
1994	75.1	58.8	6.2	6.0	7.40	11.12
1995	75.0	58.9	5.6	5.6	7.40	11.44
1996	74.9	59.3	5.4	5.4	7.43	11.82

Source: Economic Report of the President, 1997. Table citations below refer to the source document.
[1]*CLFPRM,* Civilian labor force participation rate, male (%), Table B-37, p. 343.
[2]*CLFPRF,* Civilian labor force participation rate, female (%), Table B-37, p. 343.
[3]*UNRM,* Civilian unemployment rate, male (%) Table B-40, p. 346.
[4]*UNRF,* Civilian unemployment rate, female (%) Table B-40, p. 346.
[5]*AHE82,* Average hourly earnings (1982 dollars), Table B-45, p. 352.
[6]*AHE,* Average hourly earnings (current dollars), Table B-45, p. 352.

a. Plot the male civilian labor force participation rate against male civilian unemployment rate. Eyeball a regression line through the scatter points. A priori, what is the expected relationship between the two and what is the underlying economic theory? Does the scattergram support the theory?

b. Repeat part **a** for females.

c. Now plot both the male and female labor participation rates against average hourly earnings (in 1982 dollars). (You may use separate diagrams.) Now what do you find? And how would you rationalize your finding?

d. Can you plot the labor force participation rate against the unemployment rate and the average hourly earnings simultaneously? If not, how would you verbalize the relationship among the three variables?

2.15. Table 2.8 gives data on expenditure on food and total expenditure, measured in rupees, for a sample of 55 rural households from India. (In early 2000, a U.S. dollar was about 40 Indian rupees.)

TABLE 2.8 FOOD AND TOTAL EXPENDITURE (RUPEES)

Observation	Food expenditure	Total expenditure	Observation	Food expenditure	Total expenditure
1	217.0000	382.0000	29	390.0000	655.0000
2	196.0000	388.0000	30	385.0000	662.0000
3	303.0000	391.0000	31	470.0000	663.0000
4	270.0000	415.0000	32	322.0000	677.0000
5	325.0000	456.0000	33	540.0000	680.0000
6	260.0000	460.0000	34	433.0000	690.0000
7	300.0000	472.0000	35	295.0000	695.0000
8	325.0000	478.0000	36	340.0000	695.0000
9	336.0000	494.0000	37	500.0000	695.0000
10	345.0000	516.0000	38	450.0000	720.0000
11	325.0000	525.0000	39	415.0000	721.0000
12	362.0000	554.0000	40	540.0000	730.0000
13	315.0000	575.0000	41	360.0000	731.0000
14	355.0000	579.0000	42	450.0000	733.0000
15	325.0000	585.0000	43	395.0000	745.0000
16	370.0000	586.0000	44	430.0000	751.0000
17	390.0000	590.0000	45	332.0000	752.0000
18	420.0000	608.0000	46	397.0000	752.0000
19	410.0000	610.0000	47	446.0000	769.0000
20	383.0000	616.0000	48	480.0000	773.0000
21	315.0000	618.0000	49	352.0000	773.0000
22	267.0000	623.0000	50	410.0000	775.0000
23	420.0000	627.0000	51	380.0000	785.0000
24	300.0000	630.0000	52	610.0000	788.0000
25	410.0000	635.0000	53	530.0000	790.0000
26	220.0000	640.0000	54	360.0000	795.0000
27	403.0000	648.0000	55	305.0000	801.0000
28	350.0000	650.0000			

Source: Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries,* Routledge, New York, 1998, p. 457.

a. Plot the data, using the vertical axis for expenditure on food and the horizontal axis for total expenditure, and sketch a regression line through the scatterpoints.
b. What broad conclusions can you draw from this example?
c. A priori, would you expect expenditure on food to increase linearly as total expenditure increases regardless of the level of total expenditure? Why or why not? You can use total expenditure as a proxy for total income.

2.16. Table 2.9 gives data on mean Scholastic Aptitude Test (SAT) scores for college-bound seniors for 1967–1990.
a. Use the horizontal axis for years and the vertical axis for SAT scores to plot the verbal and math scores for males and females separately.
b. What general conclusions can you draw?
c. Knowing the verbal scores of males and females, how would you go about predicting their math scores?
d. Plot the female verbal SAT score against the male verbal SAT score. Sketch a regression line through the scatterpoints. What do you observe?

TABLE 2.9 MEAN SCHOLASTIC APTITUDE TEST SCORES FOR COLLEGE-BOUND SENIORS, 1967–1990*

	Verbal			Math		
Year	Males	Females	Total	Males	Females	Total
1967	463	468	466	514	467	492
1968	464	466	466	512	470	492
1969	459	466	463	513	470	493
1970	459	461	460	509	465	488
1971	454	457	455	507	466	488
1972	454	452	453	505	461	484
1973	446	443	445	502	460	481
1974	447	442	444	501	459	480
1975	437	431	434	495	449	472
1976	433	430	431	497	446	472
1977	431	427	429	497	445	470
1978	433	425	429	494	444	468
1979	431	423	427	493	443	467
1980	428	420	424	491	443	466
1981	430	418	424	492	443	466
1982	431	421	426	493	443	467
1983	430	420	425	493	445	468
1984	433	420	426	495	449	471
1985	437	425	431	499	452	475
1986	437	426	431	501	451	475
1987	435	425	430	500	453	476
1988	435	422	428	498	455	476
1989	434	421	427	500	454	476
1990	429	419	424	499	455	476

*Data for 1967–1971 are estimates.
Source: The College Board. The *New York Times,* Aug. 28, 1990, p. B-5.