2
THE LINEAR REGRESSION MODEL

2.1 INTRODUCTION

Econometrics is concerned with *model building*. An intriguing point to begin the inquiry is to consider the question, "What is the model?" The statement of a "model" typically begins with an observation or a proposition that movement of one variable "is caused by" movement of another, or "a variable varies with another," or some qualitative statement about a relationship between a variable and one or more **covariates** that are expected to be related to the interesting variable in question. The model might make a broad statement about behavior, such as the suggestion that individuals' usage of the health care system depends on, for example, perceived health status, demographics (e.g., income, age, and education), and the amount and type of insurance they have. It might come in the form of a verbal proposition, or even a picture (e.g., a flowchart or **path diagram** that suggests directions of influence). The econometric model rarely springs forth in full bloom as a set of equations. Rather, it begins with an *idea* of some kind of relationship. The natural next step for the econometrician is to translate that idea into a set of equations, with a notion that some feature of that set of equations will answer interesting questions about the variable of interest. To continue our example, a more definite statement of the relationship between insurance and health care demanded might be able to answer *how* does health care system utilization depend on insurance coverage? Specifically, is the relationship "positive"—all else equal, is an insured consumer more likely to demand more health care than an uninsured one—or is it "negative"? And, ultimately, one might be interested in a more precise statement, "How much more (or less)?" This and the next several chapters will build the framework that model builders use to pursue questions such as these using data and econometric methods.

From a purely statistical point of view, the researcher might have in mind a variable, y, broadly "demand for health care, H," and a vector of covariates, $\mathbf{x}$ (income, I, insurance, T), and a joint probability distribution of the three, $p(H,I,T)$. Stated in this form, the "relationship" is not posed in a particularly interesting fashion—what is the statistical process that produces health care demand, income, and insurance coverage? However, it is true that $p(H,I,T) = p(H|I, T)p(I,T)$, which decomposes the probability model for the joint process into two outcomes, the joint distribution of income and insurance coverage in the population, $p(I,T)$, and the distribution of "demand for health care" for a specific income and insurance coverage, $p(H|I,T)$. From this perspective, the conditional distribution, $p(H|I,T)$, holds some particular interest, while $p(I,T)$, the distribution of income and insurance coverage in the population, is perhaps of secondary, or no interest. (On the other hand, from the same perspective, the conditional "demand" for insurance coverage, given income, $p(T|I)$, might also be interesting.) Continuing this line of thinking,

the model builder is often interested not in joint variation of all the variables in the model, but in **conditional variation** of one of the variables related to the others.

The idea of the conditional distribution provides a useful starting point for thinking about a relationship between a variable of interest, a "y," and a set of variables, "$\mathbf{x}$," that we think might bear some relationship to it. There is a question to be considered now that returns us to the issue of "What is the model?" What feature of the conditional distribution is of interest? The model builder, thinking in terms of features of the conditional distribution, often gravitates to the expected value, focusing attention on $E[y|\mathbf{x}]$, that is, the **regression function**, which brings us to the subject of this chapter. For the preceding example, this might be natural if y were "number of doctor visits" as in an application examined at several points in the chapters to follow. If we were studying incomes, I, however, which often have a highly skewed distribution, then the mean might not be particularly interesting. Rather, the **conditional median**, for given ages, $M[I|\mathbf{x}]$, might be a more interesting statistic. Still considering the distribution of incomes (and still conditioning on age), other quantiles, such as the 20th percentile, or a poverty line defined as, say, the 5th percentile, might be more interesting yet. Finally, consider a study in finance, in which the variable of interest is asset returns. In at least some contexts, means are not interesting at all—it is variances, and conditional variances in particular, that are most interesting.

The point is that we begin the discussion of the regression model with an understanding of what we mean by "the model." For the present, we will focus on the conditional mean, which is usually the feature of interest. Once we establish how to analyze the regression function, we will use it as a useful departure point for studying other features, such as quantiles and variances. The **linear regression model** is the single most useful tool in the econometrician's kit. Although to an increasing degree in contemporary research it is often only the starting point for the full investigation, it remains the device used to begin almost all empirical research. And it is the lens through which relationships among variables are usually viewed. This chapter will develop the linear regression model in detail. Here, we will detail the fundamental assumptions of the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

2.2 THE LINEAR REGRESSION MODEL

The **multiple linear regression model** is used to study the relationship between a **dependent variable** and one or more **independent variables**. The generic form of the linear regression model is

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_K) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \ \ldots \ + x_K\beta_K + \varepsilon, \end{aligned} \tag{2-1}$$

where y is the dependent or **explained variable** and $x_1, \ldots, x_K$ are the independent or **explanatory variables**. (We will return to the meaning of "independent" shortly.) One's theory will specify $f(x_1, x_2, \ldots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \ldots, x_K$. In this setting, y is the **regressand** and x_k, $k = 1, \ldots, K$ are the **regressors** or covariates. The underlying theory will specify the dependent and independent variables in the model. It is not always obvious which is

appropriately defined as each of these—for example, a demand equation, $quantity = \beta_1 + price \times \beta_2 + income \times \beta_3 + \varepsilon$, and an inverse demand equation, $price = \gamma_1 + quantity \times \gamma_2 + income \times \gamma_3 + u$ are equally valid representations of a market. For modeling purposes, it will often prove useful to think in terms of "autonomous variation." One can conceive of movement of the independent variables outside the relationships defined by the model while movement of the dependent variable is considered in response to some independent or exogenous stimulus.[1]

The term ε is a random **disturbance**, so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate. The net effect, which can be positive or negative, of these omitted factors is captured in the disturbance. There are many other contributors to the disturbance in an empirical model. Probably the most significant is errors of measurement. It is easy to theorize about the relationships among precisely defined variables; it is quite another matter to obtain accurate measures of these variables. For example, the difficulty of obtaining reasonable measures of profits, interest rates, capital stocks, or, worse yet, flows of services from capital stocks, is a recurrent theme in the empirical literature. At the extreme, there may be no observable counterpart to the theoretical variable. The literature on the permanent income model of consumption [e.g., Friedman (1957)] provides an interesting example.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$, $i = 1, \ldots, n$, is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ \ldots \ + x_{iK}\beta_K + \varepsilon_i.$$

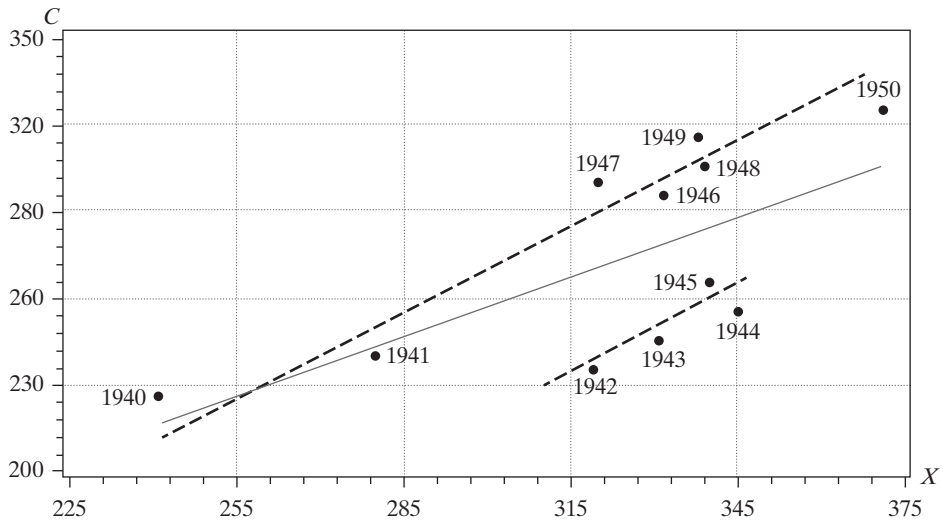
The observed value of y_i is the sum of two parts, the regression function and the disturbance, ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

Example 2.1 Keynes's Consumption Function

Example 1.2 discussed a model of consumption proposed by Keynes in his *General Theory* (1936). The theory that consumption, C, and income, X, are related certainly seems consistent with the observed "facts" in Figures 1.1 and 2.1. (These data are in Data Table F2.1.) Of course, the linear function is only approximate. Even ignoring the anomalous wartime years, consumption and income cannot be connected by any simple **deterministic relationship**. The linear part of the model, $C = \alpha + \beta X$, is intended only to represent the salient features of this part of the economy. It is hopeless to attempt to capture every influence in the relationship. The next step is to incorporate the inherent randomness in its real-world counterpart. Thus, we write $C = f(X, \varepsilon)$, where ε is a stochastic element. It is important not to view ε as a catchall for the inadequacies of the model. The model including ε appears adequate for the data not including the war years, but for 1942–1945, something systematic clearly seems to be missing. Consumption in these years could not rise to rates historically consistent with these levels of income because of wartime rationing. A model meant to describe consumption in this period would have to accommodate this influence.

[1] By this definition, it would seem that in our demand relationship, only income would be an independent variable while both price and quantity would be dependent. That makes sense—in a market, equilibrium price and quantity *are* determined at the same time, and do change only when something outside the market equilibrium changes.

FIGURE 2.1 Consumption Data, 1940–1950.



It remains to establish how the stochastic element will be incorporated in the equation. The most frequent approach is to assume that it is *additive.* Thus, we recast the equation in stochastic terms: $C = \alpha + \beta X + \varepsilon$. This equation is an empirical counterpart to Keynes's theoretical model. But, what of those anomalous years of rationing? If we were to ignore our intuition and attempt to fit a line to all these data—the next chapter will discuss at length how we should do that—we might arrive at the solid line in the figure as our best guess. This line, however, is obviously being distorted by the rationing. A more appropriate specification for these data that accommodates both the stochastic nature of the data and the special circumstances of the years 1942–1945 might be one that shifts straight down in the war years, $C = \alpha + \beta X + d_{waryears}\delta_w + \varepsilon$, where the new variable, $d_{waryears}$, equals one in 1942–1945 and zero in other years, and $\delta_w < 0$. This more detailed model is shown by the parallel dashed lines.

One of the most useful aspects of the multiple regression model is its ability to identify the separate effects of a set of variables on a dependent variable. Example 2.2 describes a common application.

Example 2.2 Earnings and Education

Many studies have analyzed the relationship between earnings and education. We would expect, on average, higher levels of education to be associated with higher incomes. The simple regression model

$$earnings = \beta_1 + \beta_2\, education + \varepsilon,$$

however, neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Thus, β_2 will overstate the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increases in income with increases in education and none with, say, experience. A better specification would account for the effect of age, as in

$$earnings = \gamma_1 + \gamma_2\, education + \gamma_3\, age + \varepsilon.$$

It is often observed that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this possibility, we might further extend the model to

$$earnings = \delta_1 + \delta_2\, education + \delta_3\, age + \delta_4\, age^2 + \varepsilon.$$

We would expect δ_3 to be positive and δ_4 to be negative.

The crucial feature of this model is that it allows us to carry out a conceptual experiment that might not be observed in the actual data. In the example, we might like to (and could) compare the earnings of two individuals of the same age with different amounts of education even if the data set does not actually contain two such individuals. How education should be measured in this setting is a difficult problem. The study of the earnings of twins by Ashenfelter and Krueger (1994), which uses precisely this specification of the earnings equation, presents an interesting approach. [Studies of twins and siblings have provided an interesting thread of research on the education and income relationship. Two other studies are Ashenfelter and Zimmerman (1997) and Bonjour, Cherkas, Haskel, Hawkes, and Spector (2003).] The experiment embodied in the earnings model thus far suggested is a comparison of two otherwise identical individuals who have different years of education. Under this interpretation, the impact of education would be $\partial E[Earnings \mid Age, Education]/\partial Education = \beta_2$. But, one might suggest that the experiment the analyst really has in mind is the truly unobservable impact of the additional year of education on a particular individual. To carry out the experiment, it would be necessary to observe the individual twice, once under circumstances that actually occur, $Education_i$, and a second time under the hypothetical (**counterfactual**) *circumstance,* $Education_i + 1$. It is convenient to frame this in a **potential outcomes model** [Rubin (1974)] for individual i:

$$Potential\ Earning = \begin{cases} y_{i0} & \text{if Education} = \mathrm{E}_i, \\ y_{i1} & \text{if Education} = \mathrm{E}_i + 1. \end{cases}$$

By this construction, all other effects would indeed be held constant, and $(y_{i1} - y_{i0})$ could reasonably be labeled the **causal effect** of the additional year of education. If we consider Education in this example as a **treatment**, then the real objective of the experiment is to measure the ***effect of the treatment on the treated***. The ability to infer this result from nonexperimental data that essentially compares "otherwise similar individuals" will be examined in Chapters 8 and19.

A large literature has been devoted to another intriguing question on this subject. Education is not truly independent in this setting. Highly motivated individuals will choose to pursue more education (e.g., by going to college or graduate school) than others. By the same token, highly motivated individuals may do things that, on average, lead them to have higher incomes. If so, does a positive β_2 that suggests an association between income and education really measure the causal effect of education on income, or does it reflect the result of some underlying effect on both variables that we have not included in the regression model? We will revisit the issue in Chapter 19.[2]

2.3 ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

The linear regression model consists of a set of assumptions about how a data set will be produced by an underlying "data-generating process." The theory will specify a relationship between a dependent variable and a set of independent variables. The

[2] This model lays yet another trap for the practitioner. In a cross section, the higher incomes of the older individuals in the sample might tell an entirely different, perhaps macroeconomic story (a cohort effect) from the lower incomes of younger individuals as time and their incomes evolve. It is not necessarily possible to deduce the characteristics of incomes of younger people in the sample *if they were older* by comparing the older individuals in the sample to the younger ones. A parallel problem arises in the analysis of treatment effects that we will examine in Chapter 8.

assumptions that describe the form of the model and relationships among its parts and imply appropriate estimation and inference procedures are listed in Table 2.1.

2.3.1 LINEARITY OF THE REGRESSION MODEL

Let the column vector $\mathbf{x}_k$ be the n observations on variable x_k, $k = 1, \ldots, K$, in a random sample of n observations, and assemble these data in an $n \times K$ data matrix, $\mathbf{X}$. In most contexts, the first column of $\mathbf{X}$ is assumed to be a column of 1s so that β_1 is

TABLE 2.1 Assumptions of the Linear Regression Model

A1. Linearity: We list the assumptions as a description of the joint distribution of y and a set of independent variables, $(x_1, x_2, \ldots, x_K) = \mathbf{x}$. The model specifies a linear relationship between y and $\mathbf{x}$; $y = x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. We will be more specific and assume that this is the regression function, $E[y|x_1, x_2, \ldots, x_K] = E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. The difference between y and $E[y|\mathbf{x}]$ is the disturbance, ε.

A2. Full rank: There is no exact *linear* relationship among any of the independent variables in the model. One way to formulate this is to assume that $E[\mathbf{x}\mathbf{x}'] = \mathbf{Q}$, a $K \times K$ matrix that has full rank K. In practical terms, we wish to be sure that for a random sample of n observations drawn from this process, $(y_1,\mathbf{x}_1'), \ldots, (y_i,\mathbf{x}_i'), \ldots, (y_n,\mathbf{x}_n')$, that the $n \times K$ matrix $\mathbf{X}$ with n rows $\mathbf{x}_i'$ always has rank K if $n \geq K$. This assumption will be necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables: $E[\varepsilon|x_1, x_2, \ldots, x_K] = E[\varepsilon|\mathbf{x}] = 0$. This states that the expected value of the disturbance in the regression is not a function of the independent variables observed. This means that the independent variables will not carry useful information for prediction of ε. The assumption is labeled **mean independence**. By the Law of Iterated Expectations (Theorem B.1), it follows that $E[\varepsilon] = 0$. An implication of the exogeneity assumption is that $E[y|x_1, x_2, \ldots, x_K] = \Sigma_{k=1}^{K} x_k\beta_k$. That is, the linear function in A1 is the **conditional mean function**, or **regression** of y on $x_1, \ldots, x_K$. In the setting of a random sample, we will also begin from an assumption that observations on ε in the sample are uncorrelated with information in other observations—that is, $E[\varepsilon_i|\mathbf{x}_1, \ldots, \mathbf{x}_n] = 0$. This is labeled **strict exogeneity**. An implication will be, for each observation in a sample of observations, $E[\varepsilon_i|\mathbf{X}] = 0$, and for the sample as a whole, $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$.

A4. Homoscedasticity: The disturbance in the regression has **conditional variance**, $\mathrm{Var}[\varepsilon|\mathbf{x}] = \mathrm{Var}[\varepsilon] = \sigma^2$. (The second equality follows from Theorem B.4.) This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow. Once again, considering a random sample, we will assume that the observations ε_i and ε_j are uncorrelated for $i \neq j$. With reference to a times-series setting, this will be labeled **nonautocorrelation**. The implication will be $\mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{x}_i,\mathbf{x}_j]$. We will strengthen this to $E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0$ for $i \neq j$ and $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{I}$.

A5. Data generation: The data in $(x_1, x_2, \ldots, x_K)$ (that is, the process by which $\mathbf{x}$ is generated) may be any mixture of constants and random variables. The crucial elements for present purposes are the exogeneity assumption, A3, and the variance and covariance assumption, A4. Analysis can be done conditionally on the observed $\mathbf{X}$, so whether the elements in $\mathbf{X}$ are fixed constants or random draws from a stochastic process will not influence the results. In later, more advanced treatments, we will want to be more specific about the possible relationship between ε_i and $\mathbf{x}_j$. Nothing is lost by assuming that the n observations in hand are a **random sample** of independent, identically distributed draws from a joint distribution of $(y,\mathbf{x})$. In some treatments to follow, such as panel data, some observations will be correlated by construction. It will be necessary to revisit the assumptions at that point, and revise them as necessary.

A6. Normal distribution: The disturbances are normally distributed. This is a convenience that we will dispense with after some analysis of its implications. The normality assumption is useful for defining the computations behind statistical inference about the regression, such as confidence intervals and hypothesis tests. For practical purposes, it will be useful then to extend those results and in the process develop a more flexible approach that does not rely on this specific assumption.

the constant term in the model. Let $\mathbf{y}$ be the n observations, $y_1, \ldots, y_n$, and let $\boldsymbol{\varepsilon}$ be the column vector containing the n disturbances. The model in (2-1) as it applies to each of and all n observations can now be written

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + \boldsymbol{\varepsilon}, \tag{2-2}$$

or in the form of Assumption A1,

$$\boxed{\text{ASSUMPTION A1:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.} \tag{2-3}$$

A NOTATIONAL CONVENTION

Henceforth, to avoid a possibly confusing and cumbersome notation, we will use a boldface $\mathbf{x}$ to denote a column *or* a row of $\mathbf{X}$. Which of these applies will be clear from the context. In (2-2), $\mathbf{x}_k$ is the kth column of $\mathbf{X}$. Subscript k will usually be used to denote columns (variables). It will often be convenient to refer to a single observation in (2-3), which we would write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{2-4}$$

Subscripts i, j, and t will generally be used to denote rows (observations) of $\mathbf{X}$. In (2-4), $\mathbf{x}_i'$ is a row vector that is the ith $1 \times K$ row of $\mathbf{X}$.

Our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. Note that the simple regression model in Example 2.1 is a special case in which $\mathbf{X}$ has only two columns, the first of which is a column of 1s. The assumption of linearity of the regression model includes the additive disturbance. For the regression to be linear in the sense described here, it must be of the form in (2-1) either in the original variables or after some suitable transformation. For example, the model

$$y = Ax^{\beta}e^{\varepsilon}$$

is linear (after taking logs on both sides of the equation), whereas

$$y = Ax^{\beta} + \varepsilon$$

is not. The observed dependent variable is thus the sum of two components, a deterministic element $\alpha + \beta x$ and a random variable ε. It is worth emphasizing that neither of the two parts is directly observed because α and β are unknown.

The linearity assumption is not so narrow as it might first appear. In the regression context, *linearity* refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables. For example, the equations $y = \alpha + \beta x + \varepsilon$, $y = \alpha + \beta \cos(x) + \varepsilon$, $y = \alpha + \beta/x + \varepsilon$, and $y = \alpha + \beta \ln x + \varepsilon$ are all linear in some function of x by the definition we have used here. In the examples, only x has been transformed, but y could have been as well, as in $y = Ax^{\beta}e^{\varepsilon}$, which is a linear relationship in the logs of x and y; $\ln y = \alpha + \beta \ln x + \varepsilon$. The variety of functions is unlimited. This aspect of the model is used in a number of commonly used functional forms. For example, the **loglinear model** is

$$\ln y = \beta_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \cdots + \beta_K \ln x_K + \varepsilon.$$

This equation is also known as the **constant elasticity** form, as in this equation, the elasticity of y with respect to changes in x_k is $\partial \ln y / \partial \ln x_k = \beta_k$, which does not vary with x_k. The loglinear form is often used in models of demand and production. Different values of β_k produce widely varying functions.

Example 2.3 The U.S. Gasoline Market

Data on the U.S. gasoline market for the years 1953–2004 are given in Table F2.2 in Appendix F. We will use these data to obtain, among other things, estimates of the income, own price, and cross-price elasticities of demand in this market. These data also present an interesting question on the issue of holding "all other things constant," that was suggested in Example 2.2. In particular, consider a somewhat abbreviated model of per capita gasoline consumption:

$$\ln(G/pop) = \beta_1 + \beta_2 \ln(Income/pop) + \beta_3 \ln price_G + \beta_4 \ln P_{newcars} + \beta_5 \ln P_{usedcars} + \varepsilon.$$

This model will provide estimates of the income and price elasticities of demand for gasoline and an estimate of the elasticity of demand with respect to the prices of new and used cars. What should we expect for the sign of β_4? Cars and gasoline are complementary goods, so if the prices of new cars rise, *ceteris paribus*, gasoline consumption should fall. Or should it? If the prices of new cars rise, then consumers will buy fewer of them; they will keep their used cars longer and buy fewer new cars. If older cars use more gasoline than newer ones, then the rise in the prices of new cars would lead to higher gasoline consumption than otherwise, not lower. We can use the multiple regression model and the gasoline data to attempt to answer the question.

A **semilog** model is often used to model growth rates:

$$\ln y_t = \mathbf{x}_t'\boldsymbol{\beta} + \delta t + \varepsilon t.$$

In this model, the autonomous (at least not explained by the model itself) proportional, per period growth rate is $\partial \ln y / \partial t = \delta$. Other variations of the general form

$$f(y_t) = g(\mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t)$$

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

The linear regression model is sometimes interpreted as an approximation to some unknown, underlying function. (See Section A.8.1 for discussion.) By this interpretation, however, the linear model, even with quadratic terms, is fairly limited in that such an approximation is likely to be useful only over a small range of variation of the independent variables. The translog model discussed in Example 2.4, in contrast, has proven more effective as an approximating function.

Example 2.4 The Translog Model

Modern studies of demand and production are usually done with a **flexible functional form**. Flexible functional forms are used in econometrics because they allow analysts to model complex features of the production function, such as elasticities of substitution, which are functions of the second derivatives of production, cost, or utility functions. The linear model restricts these to equal zero, whereas the loglinear model (e.g., the Cobb–Douglas model) restricts the interesting elasticities to the uninteresting values of -1 or $+1$. The most popular flexible functional form is the **translog model**, which is often interpreted as a second-order approximation to an unknown functional form. [See Berndt and Christensen (1973).] One way to derive it is as follows. We first write $y = g(x_1, \ldots, x_K)$. Then, $\ln y = \ln g(\ldots) = f(\ldots)$. Since by a trivial transformation $x_k = \exp(\ln x_k)$, we interpret the function as a function of the logarithms of the x's. Thus, $\ln y = f(\ln x_1, \ldots, \ln x_K)$.

Now, expand this function in a second-order Taylor series around the point $\mathbf{x} = [1, 1, \ldots, 1]'$ so that at the expansion point, the log of each variable is a convenient zero. Then

$$\ln y = f(\mathbf{0}) + \sum_{k=1}^{K} [\partial f(\cdot)/\partial \ln x_k]_{|\ \ln \mathbf{x} = 0} \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} [\partial^2 f(\cdot)/\partial \ln x_k \partial \ln x_l]_{|\ \ln \mathbf{x} = 0} \ln x_k \ln x_l + \varepsilon.$$

The disturbance in this model is assumed to embody the familiar factors and the error of approximation to the unknown function. Because the function and its derivatives evaluated at the fixed value **0** are constants, we interpret them as the coefficients and write

$$\ln y = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} \ln x_k \ln x_l + \varepsilon.$$

This model is linear by our definition but can, in fact, mimic an impressive amount of curvature when it is used to approximate another function. An interesting feature of this formulation is that the loglinear model is a special case, when $\gamma_{kl} = 0$. Also, there is an interesting test of the underlying theory possible because if the underlying function were assumed to be continuous and twice continuously differentiable, then by Young's theorem it must be true that $\gamma_{kl} = \gamma_{lk}$. We will see in Chapter 10 how this feature is studied in practice.

Despite its great flexibility, the linear model will not accommodate all the situations we will encounter in practice. In Example 14.13 and Chapter 18, we will examine the regression model for doctor visits that was suggested in the introduction to this chapter. An appropriate model that describes the number of visits has conditional mean function $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. It is tempting to linearize this directly by taking logs, because $\ln E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. But $\ln E[y|\mathbf{x}]$ is not equal to $E[\ln y|\mathbf{x}]$. In that setting, y can equal zero (and does for most of the sample), so $\mathbf{x}'\boldsymbol{\beta}$ (which can be negative) is not an appropriate model for $\ln y$ (which does not exist) or for y which cannot be negative. The methods we consider in this chapter are not appropriate for estimating the parameters of such a model. Relatively straightforward techniques have been developed for nonlinear models such as this, however. We shall treat them in detail in Chapter 7.

2.3.2 FULL RANK

Assumption A2 is that there are no exact *linear* relationships among the variables.

ASSUMPTION A2: $\mathbf{X}$ is an $n \times K$ matrix with rank K. **(2-5)**

Hence, $\mathbf{X}$ has full column rank; the columns of $\mathbf{X}$ are linearly independent and there are at least K observations. [See (A-42) and the surrounding text.] This assumption is known as an **identification condition**. To see the need for this assumption, consider an example.

Example 2.5 Short Rank

Suppose that a cross-section model specifies that consumption, C, relates to income as follows:

$$C = \beta_1 + \beta_2\, \textit{nonlabor income} + \beta_3\, \textit{salary} + \beta_4\, \textit{total income} + \varepsilon,$$

where *total income* is exactly equal to *salary* plus *nonlabor income*. Clearly, there is an exact linear relationship among the variables in the model. Now, let

$$\beta'_2 = \beta_2 + a,$$
$$\beta'_3 = \beta_3 + a,$$

and

$$\beta'_4 = \beta_4 - a,$$

where a is any number. Then the exact same value appears on the right-hand side of C if we substitute β'_2, β'_3, and β'_4 for β_2, β_3, and β_4. Obviously, there is no way to estimate the parameters of this model.

If there are fewer than K observations, then $\mathbf{X}$ cannot have **full rank**. Hence, we make the assumption that n is at least as large as K.

In the simple linear model with a constant term and a single x, the full rank assumption means that there must be variation in the regressor, x. If there is no variation in x, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could* have drawn a sample in which there was variation in x, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand.

Example 2.6 An Inestimable Model

In Example 3.4, we will consider a model for the sale price of Monet paintings. Theorists and observers have different models for how prices of paintings at auction are determined. One (naïve) student of the subject suggests the model

$$\begin{aligned}\ln \textit{Price} &= \beta_1 + \beta_2 \ln \textit{Size} + \beta_3 \ln \textit{Aspect Ratio} + \beta_4 \ln \textit{Height} + \varepsilon \\ &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,\end{aligned}$$

where $\textit{Size} = \textit{Width} \times \textit{Height}$ and $\textit{Aspect Ratio} = \textit{Width}/\textit{Height}$. By simple arithmetic, we can see that this model shares the problem found with the consumption model in Example 2.5—in this case, $x_2 - x_4 = x_3 + x_4$. So, this model is, like the previous one, not estimable—it is not identified. It is useful to think of the problem from a different perspective here (so to speak). In the linear model, it must be possible for the variables in the model to vary linearly independently. But, in this instance, while it is possible for any pair of the three covariates to vary independently, the three together cannot. The "model," that is, the theory, is an entirely reasonable model as it stands. Art buyers might very well consider all three of these features in their valuation of a Monet painting. However, it is not possible to learn about that from the observed data, at least not with this linear regression model.

The full rank assumption is occasionally interpreted to mean that the variables in $\mathbf{X}$ must be able to vary independently from each other. This is clearly not the case in Example 2.6, which is a flawed model. But it is also not the case in the linear model

$$E[y|x,z] = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 z + \varepsilon.$$

There is nothing problematic with this model—nor with the model in Example 2.2 or the translog model in Example 2.4. Nonetheless, x and x^2 cannot vary independently. The resolution of this seeming contradiction is to sharpen what we mean by the variables in the model varying independently. First, it remains true that $\mathbf{X}$ must have full column rank to carry out the *linear* regression. But, independent variation of the variables in the model is a different concept. The columns of $\mathbf{X}$ are not necessarily the set of variables in the model. In the equation above, the "variables" are only x and z. The identification problem we consider here would state that it must be possible for z to vary independently

from x. If z is a deterministic function of x, then it is not possible to identify an effect in the model for variable z separately from that for x.

2.3.3 REGRESSION

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$E[\varepsilon_i|\mathbf{X}] = 0. \tag{2-6}$$

For the full set of observations, we write Assumption A3 as

$$\text{ASSUMPTION A3:} \quad E[\boldsymbol{\varepsilon}|\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1|\mathbf{X}] \\ E[\varepsilon_2|\mathbf{X}] \\ \vdots \\ E[\varepsilon_n|\mathbf{X}] \end{bmatrix} = \mathbf{0}. \tag{2-7}$$

There is a subtle point in this discussion that the observant reader might have noted. In (2-7), the left-hand side states, in principle, that the mean of each ε_i *conditioned on all observations* $\mathbf{x}_j$ is zero. This strict exogeneity assumption states, in words, that no observations on $\mathbf{x}$ convey information about the expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $E[\varepsilon_i|\cdot]$, $\mathbf{x}_j$ *at some other observation*, such as in the previous time period, might. Our assumption at this point is that there is no information about $E[\varepsilon_i|\cdot]$ contained in *any* observation $\mathbf{x}_j$. Later, when we extend the model, we will study the implications of dropping this assumption. [See Wooldridge (1995).] We will also assume that the disturbances convey no information about each other. That is, $E[\varepsilon_i|\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.

The zero conditional mean implies that the unconditional mean is also zero, because by the **Law of Iterated Expectations** [Theorem B.1, (B-66)],

$$E[\varepsilon_i] = E_{\mathbf{x}}[E[\varepsilon_i|\mathbf{X}]] = E_{\mathbf{x}}[0] = 0.$$

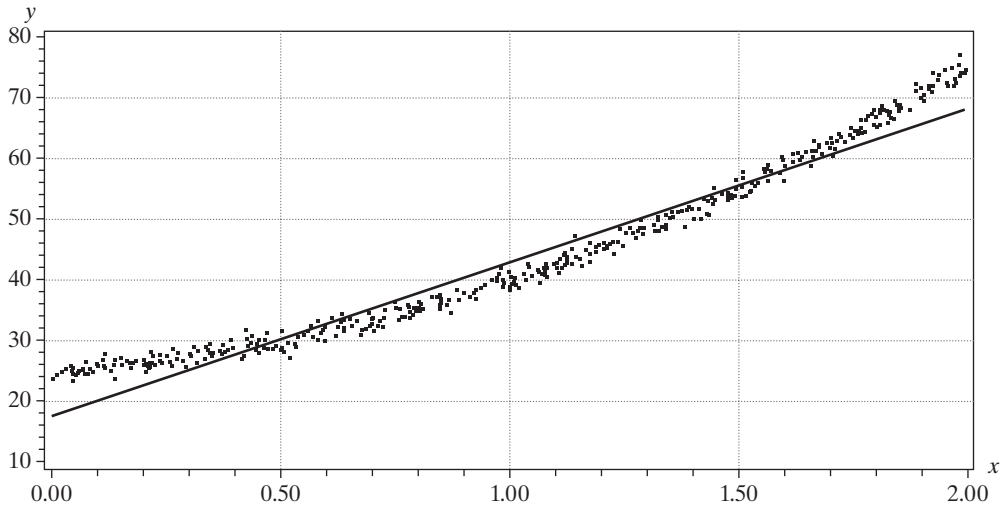
For each ε_i, by Theorem B.2, $\text{Cov}[E[\varepsilon_i|\mathbf{X}], \mathbf{X}] = \text{Cov}[\varepsilon_i, \mathbf{X}]$, Assumption A3 implies that $\text{Cov}[\varepsilon_i, \mathbf{x}] = 0$ for all i. The converse is not true; $E[\varepsilon_i] = 0$ does not imply that $E[\varepsilon_i|\mathbf{x}_i] = 0$. Example 2.7 illustrates the difference.

Example 2.7 Nonzero Conditional Mean of the Disturbances

Figure 2.2 illustrates the important difference between $E[\varepsilon_i] = 0$ and $E[\varepsilon_i|x_i] = 0$. The overall mean of the disturbances in the sample is zero, but the mean for specific ranges of x is distinctly nonzero. A pattern such as this in observed data would serve as a useful indicator that the specification of the linear regression should be questioned. In this particular case, the true conditional mean function (which the researcher would not know in advance) is actually $E[y|x] = 25 + 5x(1 + 2x)$. The sample data are suggesting that a linear specification is not appropriate for these data. A quadratic specification would seem to be a good candidate. This modeling strategy is pursued in an application in Example 6.6.

In most cases, the zero overall mean assumption is not restrictive. Consider a two-variable model and suppose that the mean of ε is $\mu \neq 0$. Then $\alpha + \beta x + \varepsilon$ is the same

FIGURE 2.2 Disturbances with Nonzero Conditional Mean and Zero Unconditional Mean.



as $(\alpha + \mu) + \beta x + (\varepsilon - \mu)$. Letting $\alpha' = \alpha + \mu$ and $\varepsilon' = \varepsilon - \mu$ produces the original model. For an application, see the discussion of frontier production functions in Section 19.2.4. But if the original model does not contain a constant term, then assuming $E[\varepsilon_i] = 0$ could be substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory.[3] Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption A3 also implies that

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}. \tag{2-8}$$

Assumptions A1 and A3 comprise the *linear regression model*. The regression of $\mathbf{y}$ on $\mathbf{X}$ is the conditional mean, $E[\mathbf{y}|\mathbf{X}]$, so that without Assumption A3, $\mathbf{X}\boldsymbol{\beta}$ is *not* the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

2.3.4 HOMOSCEDASTIC AND NONAUTOCORRELATED DISTURBANCES

The fourth assumption concerns the variances and covariances of the disturbances:

$$\text{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2, \qquad \text{for all } i = 1, \ldots, n,$$

[3] Models that describe first differences of variables might well be specified without constants. Consider $y_t - y_{t-1}$. If there is a constant term α on the right-hand side of the equation, then y_t is a function of αt, which is an explosive regressor. Models with linear time trends merit special treatment in the time-series literature. We will return to this issue in Chapter 21.

and

$$\text{Cov}[\varepsilon_i, \varepsilon_j|\mathbf{X}] = 0, \qquad \text{for all } i \neq j.$$

Constant variance is labeled **homoscedasticity**. Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Survey data on household expenditure patterns often display marked **heteroscedasticity**, even after accounting for income and household size.

Uncorrelatedness across observations is labeled generically nonautocorrelation. In Figure 2.1, there is some suggestion that the disturbances might not be truly independent across observations. Although the number of observations is small, it does appear that, on average, each disturbance tends to be followed by one with the same sign. This "inertia" is precisely what is meant by **autocorrelation**, and it is assumed away at this point. Methods of handling autocorrelation in economic data occupy a large proportion of the literature and will be treated at length in Chapter 20. Note that nonautocorrelation does not imply that observations y_i and y_j are uncorrelated. The assumption is that *deviations* of observations from their expected values are uncorrelated.

The two assumptions imply that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1|\mathbf{X}] & E[\varepsilon_1\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n|\mathbf{X}] \\ E[\varepsilon_2\varepsilon_1|\mathbf{X}] & E[\varepsilon_2\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n|\mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1|\mathbf{X}] & E\varepsilon_n\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n|\mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

which we summarize in Assumption A4:

$$\boxed{\text{Assumption A4:} \qquad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}.} \tag{2-9}$$

By using the variance decomposition formula in (B-69), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \sigma^2\mathbf{I}.$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications. Models such as $\text{Var}[\varepsilon_t|\varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}^2$, a "GARCH" model (see Chapter 20), do not violate our conditional variance assumption, but do assume that $\text{Var}[\varepsilon_t|\varepsilon_{t-1}] \neq \text{Var}[\varepsilon_t]$.

2.3.5 DATA GENERATING PROCESS FOR THE REGRESSORS

It is common to assume that $\mathbf{x}_i$ is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $\mathbf{x}_i$ is fertilizer concentration and water applied. The assumption of **nonstochastic regressors** at this point would be a mathematical convenience. With it, we could use the results of elementary statistics to obtain our results by treating the vector $\mathbf{x}_i$ simply as a known constant in the probability distribution of y_i. With this simplification, Assumptions A3 and A4 would be made unconditional and the counterparts would now simply state that the probability distribution of ε_i involves none of the constants in $\mathbf{X}$.

Social scientists are almost never able to analyze experimental data, and relatively few of their models are built around nonrandom regressors. Clearly, for example, in any model of the macroeconomy, it would be difficult to defend such an asymmetric treatment of aggregate data. Realistically, we have to allow the data on $\mathbf{x}_i$ to be random the same as y_i. So an alternative formulation is to assume that $\mathbf{x}_i$ is a random vector and our formal assumption concerns the nature of the random process that produces $\mathbf{x}_i$. If $\mathbf{x}_i$ is taken to be a random vector, then Assumptions A1 through A4 become a statement about the joint distribution of y_i and $\mathbf{x}_i$. The precise nature of the regressor and how we view the sampling process will be a major determinant of our derivation of the statistical properties of our estimators and test statistics. In the end, the crucial assumption is A3, the uncorrelatedness of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$. Now, we do note that this alternative is not completely satisfactory either, because $\mathbf{X}$ may well contain nonstochastic elements, including a constant, a time trend, and dummy variables that mark specific episodes in time. This makes for an ambiguous conclusion, but there is a straightforward and economically useful way out of it. We will allow $\mathbf{X}$ to be any mixture of constants and random variables, and the mean and variance of ε_i are both independent of all elements of $\mathbf{X}$.

ASSUMPTION A5: $\mathbf{X}$ may be fixed or random.	**(2-10)**

2.3.6 NORMALITY

It is convenient to assume that the disturbances are **normally distributed**, with zero mean and constant variance. That is, we add normality of the distribution to Assumptions A3 and A4.

ASSUMPTION A6: $\varepsilon \mid \mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$.	**(2-11)**

In view of our description of the source of ε, the conditions of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. A useful implication of Assumption A6 is that it implies that observations on ε_i are statistically independent as well as uncorrelated. [See the third point in Section B.9, (B-97) and (B-99).]

Normality is usually viewed as an unnecessary and possibly inappropriate addition to the regression model. Except in those cases in which some alternative distribution is explicitly assumed, as in the stochastic frontier model discussed in Chapter 19, the normality assumption may be quite reasonable. But the assumption is not necessary

to obtain most of the results we use in multiple regression analysis. It will prove useful as a starting point in constructing confidence intervals and test statistics, as shown in Section 4.7 and Chapter 5. But it will be possible to discard this assumption and retain for practical purposes the important statistical results we need for the investigation.

2.3.7 INDEPENDENCE AND EXOGENEITY

The term *independent* has been used several ways in this chapter.

In Section 2.2, the right-hand-side variables in the model are denoted the independent variables. Here, the notion of independence refers to the sources of variation. In the context of the model, the variation in the independent variables arises from sources that are outside of the process being described. Thus, in our health services versus income example in the introduction, we have suggested a theory for how variation in demand for services is associated with variation in income and, possibly, variation in insurance coverage. But, we have not suggested an explanation of the sample variation in income; income is assumed to vary for reasons that are outside the scope of the model. Nor have we suggested a behavioral model for insurance take up. This will be a convenient definition to use for **exogeneity** of a variable x.

The assumption in (2-6), $E[\varepsilon_i|\mathbf{X}] = 0$, is mean independence. Its implication is that variation in the disturbances in our data is not explained by variation in the independent variables. Situations in which $E[\varepsilon_i|\mathbf{X}] \neq 0$ arise frequently, as we will explore in Chapter 8 and others. When $E[\varepsilon|x] \neq 0$, x is **endogenous** in the model. The most straightforward instance is a left-out variable. Consider the model in Example 2.2. In a simple model that contains only *Education* but which has inappropriately omitted *Age*, it would follow that *Age* implicitly appears in the disturbance:

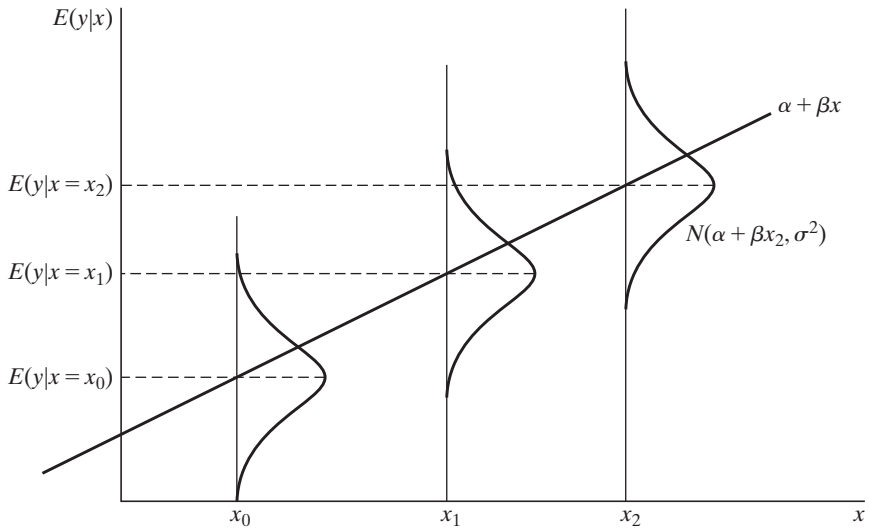
$$Income = \gamma_1 + \gamma_2 Education + (\gamma_3 Age + u) = \gamma_1 + \gamma_2 Education + \varepsilon.$$

If *Education* and (the hidden variable) *Age* are correlated, then *Education* is endogenous in this equation, which is no longer a regression because $E[\varepsilon|Education] = \gamma_3 E[Age|Education] + E[u|Education] \neq 0$.

We have also assumed in Section 2.3.4 that the disturbances are uncorrelated with each other (Assumption A4 in Table 2.1). This implies that $E[\varepsilon_i|\varepsilon_j] = 0$ when $i \neq j$—the disturbances are also mean independent of each other. Conditional normality of the disturbances assumed in Section 2.3.6 (Assumption A6) implies that they are statistically independent of each other, which is a stronger result than mean independence and stronger than we will need in most applications.

Finally, Section 2.3.2 discusses the **linear independence** of the columns of the data matrix, $\mathbf{X}$. The notion of independence here is an algebraic one relating to the column rank of $\mathbf{X}$. In this instance, the underlying interpretation is that it must be possible for the variables in the model to vary linearly independently of each other. Thus, in Example 2.6, we find that it is not possible for the logs of surface area, aspect ratio, and height of a painting all to vary independently of one another. The modeling implication is that, if the variables cannot vary independently of each other, then it is not possible to analyze them in a linear regression model that assumes the variables can each vary while holding the others constant. There is an ambiguity in this discussion of independence of the variables. We have both *age* and *age squared* in a model in Example 2.2. These cannot vary independently, but there is no obstacle to formulating a linear regression

FIGURE 2.3 The Normal Linear Regression Model.



model containing both *age* and *age squared*. The resolution is that *age* and *age squared*, though not *functionally* independent, are *linearly* independent in **X**. That is the crucial assumption in the linear regression model.

2.4 SUMMARY AND CONCLUSIONS

This chapter has framed the linear regression model, the basic platform for model building in econometrics. The assumptions of the classical regression model are summarized in Figure 2.3, which shows the two-variable case.

Key Terms and Concepts

- Autocorrelation
- Central limit theorem
- Conditional mean
- Conditional median
- Conditional variance
- Conditional variation
- Constant elasticity
- Counterfactual
- Covariate
- Dependent variable
- Deterministic relationship
- Disturbance
- Endogeneity
- Exogeneity
- Explained variable
- Explanatory variable
- Flexible functional form
- Full rank
- Heteroscedasticity
- Homoscedasticity
- Identification condition
- Impact of treatment on the treated
- Independent variable
- Law of Iterated Expectations
- Linear independence
- Linear regression model
- Loglinear model
- Mean independence
- Multiple linear regression model
- Nonautocorrelation
- Nonstochastic regressors
- Normality
- Normally distributed
- Path diagram
- Population regression equation
- Random sample
- Regressand
- Regression function
- Regressor
- Semilog
- Translog model