# 1

# THE NATURE OF REGRESSION ANALYSIS

As mentioned in the Introduction, regression is a main tool of econometrics, and in this chapter we consider very briefly the nature of this tool.

## 1.1 HISTORICAL ORIGIN OF THE TERM *REGRESSION*

The term *regression* was introduced by Francis Galton. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole.[1] In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton's *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.[2] He found that the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a group of short fathers was greater than their fathers' height, thus "regressing" tall and short sons alike toward the average height of all men. In the words of Galton, this was "regression to mediocrity."

[1]Francis Galton, "Family Likeness in Stature," *Proceedings of Royal Society, London*, vol. 40, 1886, pp. 42–72.

[2]K. Pearson and A. Lee, "On the Laws of Inheritance," *Biometrika*, vol. 2, Nov. 1903, pp. 357–462.

## 1.2 THE MODERN INTERPRETATION OF REGRESSION

The modern interpretation of regression is, however, quite different. Broadly speaking, we may say

> Regression analysis is concerned with the study of the dependence of one variable, the *dependent variable*, on one or more other variables, the *explanatory variables*, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

The full import of this view of regression analysis will become clearer as we progress, but a few simple examples will make the basic concept quite clear.

### Examples

**1.** Reconsider Galton's law of universal regression. Galton was interested in finding out why there was a stability in the distribution of heights in a population. But in the modern view our concern is not with this explanation but rather with finding out how the *average* height of sons changes, given the fathers' height. In other words, our concern is with predicting the average height of sons knowing the height of their fathers. To see how this can be done, consider Figure 1.1, which is a **scatter diagram,** or **scatter-**

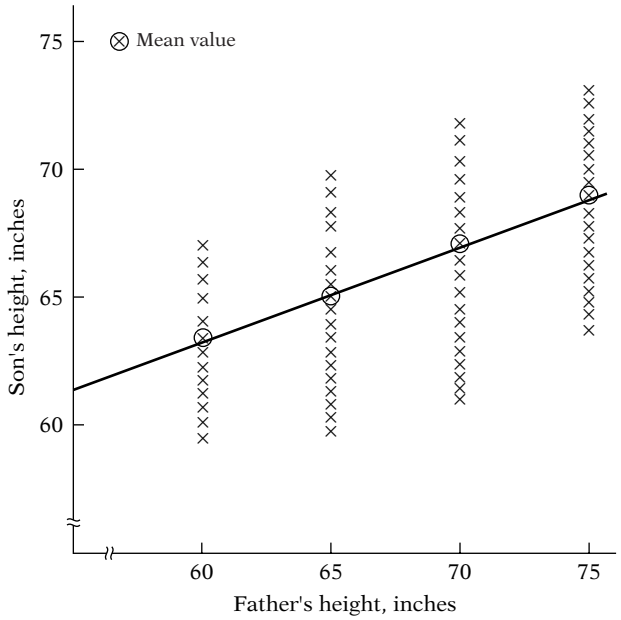


**FIGURE 1.1** Hypothetical distribution of sons' heights corresponding to given heights of fathers.

**gram.** This figure shows the distribution of heights of sons in a hypothetical population corresponding to the given or *fixed* values of the father's height. Notice that corresponding to any given height of a father is a *range* or distribution of the heights of the sons. However, notice that despite the variability of the height of sons for a given value of father's height, the average height of sons generally increases as the height of the father increases. To show this clearly, the circled crosses in the figure indicate the *average* height of sons corresponding to a given height of the father. Connecting these averages, we obtain the line shown in the figure. This line, as we shall see, is known as the **regression line.** It shows how the *average* height of sons increases with the father's height.[3]

**2.** Consider the scattergram in Figure 1.2, which gives the distribution in a hypothetical population of heights of boys measured at *fixed* ages. Corresponding to any given age, we have a range, or distribution, of heights. Obviously, not all boys of a given age are likely to have identical heights. But height *on the average* increases with age (of course, up to a certain age), which can be seen clearly if we draw a line (the regression line) through the

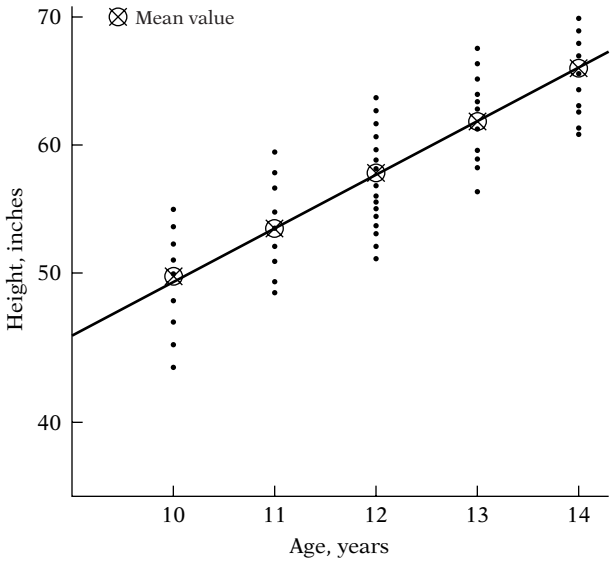


**FIGURE 1.2** Hypothetical distribution of heights corresponding to selected ages.

[3]At this stage of the development of the subject matter, we shall call this regression line simply the *line connecting the mean, or average, value of the dependent variable (son's height) corresponding to the given value of the explanatory variable (father's height)*. Note that this line has a positive slope but the slope is less than 1, which is in conformity with Galton's regression to mediocrity. (Why?)

circled points that represent the average height at the given ages. Thus, knowing the age, we may be able to predict from the regression line the average height corresponding to that age.

**3.** Turning to economic examples, an economist may be interested in studying the dependence of personal consumption expenditure on after-tax or disposable real personal income. Such an analysis may be helpful in estimating the marginal propensity to consume (MPC), that is, average change in consumption expenditure for, say, a dollar's worth of change in real income (see Figure I.3).

**4.** A monopolist who can fix the price or output (but not both) may want to find out the response of the demand for a product to changes in price. Such an experiment may enable the estimation of the **price elasticity** (i.e., price responsiveness) of the demand for the product and may help determine the most profitable price.

**5.** A labor economist may want to study the rate of change of money wages in relation to the unemployment rate. The historical data are shown in the scattergram given in Figure 1.3. The curve in Figure 1.3 is an example of the celebrated *Phillips curve* relating changes in the money wages to the unemployment rate. Such a scattergram may enable the labor economist to predict the average change in money wages given a certain unemployment rate. Such knowledge may be helpful in stating something about the inflationary process in an economy, for increases in money wages are likely to be reflected in increased prices.


**FIGURE 1.3** Hypothetical Phillips curve.

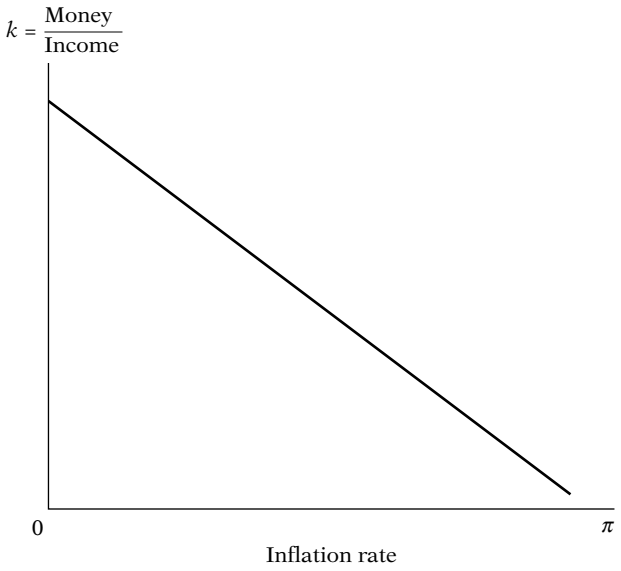


**FIGURE 1.4** Money holding in relation to the inflation rate $\pi$.

**6.** From monetary economics it is known that, other things remaining the same, the higher the rate of inflation $\pi$, the lower the proportion $k$ of their income that people would want to hold in the form of money, as depicted in Figure 1.4. A quantitative analysis of this relationship will enable the monetary economist to predict the amount of money, as a proportion of their income, that people would want to hold at various rates of inflation.

**7.** The marketing director of a company may want to know how the demand for the company's product is related to, say, advertising expenditure. Such a study will be of considerable help in finding out the **elasticity of demand** with respect to advertising expenditure, that is, the percent change in demand in response to, say, a 1 percent change in the advertising budget. This knowledge may be helpful in determining the "optimum" advertising budget.

**8.** Finally, an agronomist may be interested in studying the dependence of crop yield, say, of wheat, on temperature, rainfall, amount of sunshine, and fertilizer. Such a dependence analysis may enable the prediction or forecasting of the average crop yield, given information about the explanatory variables.

The reader can supply scores of such examples of the dependence of one variable on one or more other variables. The techniques of regression analysis discussed in this text are specially designed to study such dependence among variables.

## 1.3 STATISTICAL VERSUS DETERMINISTIC RELATIONSHIPS

From the examples cited in Section 1.2, the reader will notice that in regression analysis we are concerned with what is known as the *statistical*, not *functional* or *deterministic*, dependence among variables, such as those of classical physics. In statistical relationships among variables we essentially deal with **random** or **stochastic**[4] variables, that is, variables that have probability distributions. In functional or deterministic dependency, on the other hand, we also deal with variables, but these variables are not random or stochastic.

The dependence of crop yield on temperature, rainfall, sunshine, and fertilizer, for example, is statistical in nature in the sense that the explanatory variables, although certainly important, will not enable the agronomist to predict crop yield exactly because of errors involved in measuring these variables as well as a host of other factors (variables) that collectively affect the yield but may be difficult to identify individually. Thus, there is bound to be some "intrinsic" or random variability in the dependent-variable crop yield that cannot be fully explained no matter how many explanatory variables we consider.

In deterministic phenomena, on the other hand, we deal with relationships of the type, say, exhibited by Newton's law of gravity, which states: Every particle in the universe attracts every other particle with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Symbolically, $F = k(m_1 m_2 / r^2)$, where $F =$ force, $m_1$ and $m_2$ are the masses of the two particles, $r =$ distance, and $k =$ constant of proportionality. Another example is Ohm's law, which states: For metallic conductors over a limited range of temperature the current $C$ is proportional to the voltage $V$; that is, $C = (\frac{1}{k})V$ where $\frac{1}{k}$ is the constant of proportionality. Other examples of such deterministic relationships are Boyle's gas law, Kirchhoff's law of electricity, and Newton's law of motion.

In this text we are not concerned with such deterministic relationships. Of course, if there are errors of measurement, say, in the $k$ of Newton's law of gravity, the otherwise deterministic relationship becomes a statistical relationship. In this situation, force can be predicted only approximately from the given value of $k$ (and $m_1$, $m_2$, and $r$), which contains errors. The variable $F$ in this case becomes a random variable.

## 1.4 REGRESSION VERSUS CAUSATION

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. In the words of Kendall and Stuart, "A statistical relationship, however strong and however

[4]The word *stochastic* comes from the Greek word *stokhos* meaning "a bull's eye." The outcome of throwing darts on a dart board is a stochastic process, that is, a process fraught with misses.

suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."[5]

In the crop-yield example cited previously, there is no *statistical reason* to assume that rainfall does not depend on crop yield. The fact that we treat crop yield as dependent on rainfall (among other things) is due to nonstatistical considerations: Common sense suggests that the relationship cannot be reversed, for we cannot control rainfall by varying crop yield.

In all the examples cited in Section 1.2 the point to note is that **a statistical relationship in itself cannot logically imply causation.** To ascribe causality, one must appeal to a priori or theoretical considerations. Thus, in the third example cited, one can invoke economic theory in saying that consumption expenditure depends on real income.[6]

## 1.5 REGRESSION VERSUS CORRELATION

Closely related to but conceptually very much different from regression analysis is **correlation analysis,** where the primary objective is to measure the *strength* or *degree* of *linear association* between two variables. The **correlation coefficient,** which we shall study in detail in Chapter 3, measures this strength of (linear) association. For example, we may be interested in finding the correlation (coefficient) between smoking and lung cancer, between scores on statistics and mathematics examinations, between high school grades and college grades, and so on. In regression analysis, as already noted, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables. Thus, we may want to know whether we can predict the average score on a statistics examination by knowing a student's score on a mathematics examination.

Regression and correlation have some fundamental differences that are worth mentioning. In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution. The explanatory variables, on the other hand, are assumed to have fixed values (in repeated sampling),[7] which was made explicit in the definition of regression given in Section 1.2. Thus, in Figure 1.2 we assumed that the variable age was fixed at given levels and height measurements were obtained at these levels. In correlation analysis, on the

[5]M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Charles Griffin Publishers, New York, 1961, vol. 2, chap. 26, p. 279.

[6]But as we shall see in Chap. 3, classical regression analysis is based on the assumption that the model used in the analysis is the correct model. Therefore, the direction of causality may be implicit in the model postulated.

[7]It is crucial to note that the explanatory variables may be intrinsically stochastic, but for the purpose of regression analysis we assume that their values are fixed in repeated sampling (that is, $X$ assumes the same values in various samples), thus rendering them in effect nonrandom or nonstochastic. But more on this in Chap. 3, Sec. 3.2.

other hand, we treat any (two) variables symmetrically; there is no distinction between the dependent and explanatory variables. After all, the correlation between scores on mathematics and statistics examinations is the same as that between scores on statistics and mathematics examinations. Moreover, both variables are assumed to be random. As we shall see, most of the correlation theory is based on the assumption of randomness of variables, whereas most of the regression theory to be expounded in this book is conditional upon the assumption that the dependent variable is stochastic but the explanatory variables are fixed or nonstochastic.[8]

## 1.6 TERMINOLOGY AND NOTATION

Before we proceed to a formal analysis of regression theory, let us dwell briefly on the matter of terminology and notation. In the literature the terms *dependent variable* and *explanatory variable* are described variously. A representative list is:

| Dependent variable | Explanatory variable |
|---|---|
| ⇕ | ⇕ |
| Explained variable | Independent variable |
| ⇕ | ⇕ |
| Predictand | Predictor |
| ⇕ | ⇕ |
| **Regressand** | **Regressor** |
| ⇕ | ⇕ |
| Response | Stimulus |
| ⇕ | ⇕ |
| Endogenous | Exogenous |
| ⇕ | ⇕ |
| Outcome | Covariate |
| ⇕ | ⇕ |
| Controlled variable | Control variable |

Although it is a matter of personal taste and tradition, in this text we will use the dependent variable/explanatory variable or the more neutral, regressand and regressor terminology.

If we are studying the dependence of a variable on only a single explanatory variable, such as that of consumption expenditure on real income, such a study is known as *simple*, or **two-variable, regression analysis.** However, if we are studying the dependence of one variable on more than

[8]In advanced treatment of econometrics, one can relax the assumption that the explanatory variables are nonstochastic (see introduction to Part II).

one explanatory variable, as in the crop-yield, rainfall, temperature, sunshine, and fertilizer examples, it is known as **multiple regression analysis.** In other words, in two-variable regression there is only one explanatory variable, whereas in multiple regression there is more than one explanatory variable.

The term **random** is a synonym for the term **stochastic.** As noted earlier, a random or stochastic variable is a variable that can take on any set of values, positive or negative, with a given probability.[9]

Unless stated otherwise, the letter $Y$ will denote the dependent variable and the $X$'s ($X_1, X_2, \ldots, X_k$) will denote the explanatory variables, $X_k$ being the $k$th explanatory variable. The subscript $i$ or $t$ will denote the $i$th or the $t$th observation or value. $X_{ki}$ (or $X_{kt}$) will denote the $i$th (or $t$th) observation on variable $X_k$. $N$ (or $T$) will denote the total number of observations or values in the population, and $n$ (or $t$) the total number of observations in a sample. As a matter of convention, the observation subscript $i$ will be used for **cross-sectional data** (i.e., data collected at one point in time) and the subscript $t$ will be used for **time series data** (i.e., data collected over a period of time). The nature of cross-sectional and time series data, as well as the important topic of the nature and sources of data for empirical analysis, is discussed in the following section.

## 1.7 THE NATURE AND SOURCES OF DATA FOR ECONOMIC ANALYSIS[10]

The success of any econometric analysis ultimately depends on the availability of the appropriate data. It is therefore essential that we spend some time discussing the nature, sources, and limitations of the data that one may encounter in empirical analysis.

### Types of Data

Three types of data may be available for empirical analysis: **time series, cross-section,** and **pooled** (i.e., combination of time series and cross-section) data.

**Time Series Data** The data shown in Table I.1 of the Introduction are an example of time series data. A *time series* is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as **daily** (e.g., stock prices, weather reports), **weekly** (e.g., money supply figures), **monthly** [e.g., the unemployment rate, the Consumer Price Index (CPI)], **quarterly** (e.g., GDP), **annually** (e.g.,

[9]See **App. A** for formal definition and further details.

[10]For an informative account, see Michael D. Intriligator, *Econometric Models, Techniques, and Applications,* Prentice Hall, Englewood Cliffs, N.J., 1978, chap. 3.

government budgets), **quinquennially,** that is, every 5 years (e.g., the census of manufactures), or **decennially** (e.g., the census of population). Sometime data are available both quarterly as well as annually, as in the case of the data on GDP and consumer expenditure. With the advent of high-speed computers, data can now be collected over an extremely short interval of time, such as the data on stock prices, which can be obtained literally continuously (the so-called *real-time quote*).

Although time series data are used heavily in econometric studies, they present special problems for econometricians. As we will show in chapters on **time series econometrics** later on, most empirical work based on time series data assumes that the underlying time series is **stationary.** Although it is too early to introduce the precise technical meaning of stationarity at this juncture, *loosely speaking a time series is stationary if its mean and variance do not vary systematically over time.* To see what this means, consider Figure 1.5, which depicts the behavior of the M1 money supply in the United States from January 1, 1959, to July 31, 1999. (The actual data are given in exercise 1.4.) As you can see from this figure, the M1 money supply shows a steady upward **trend** as well as variability over the years, suggesting that the M1 time series is not stationary.[11] We will explore this topic fully in Chapter 21.

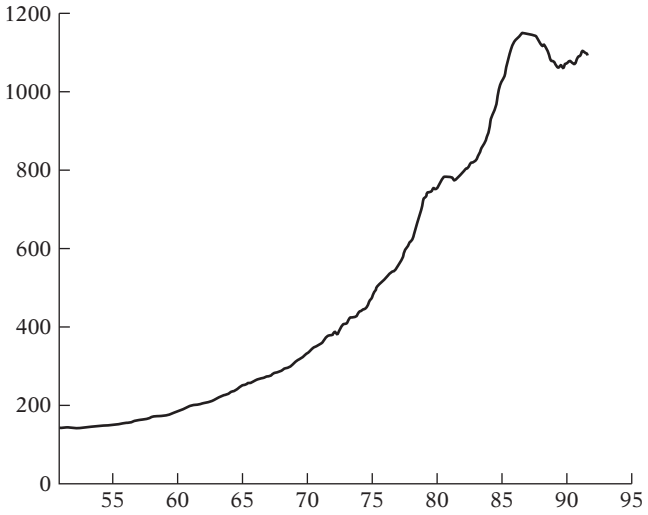


**FIGURE 1.5** M1 money supply: United States, 1951:01–1999:09.

[11]To see this more clearly, we divided the data into four time periods: 1951:01 to 1962:12; 1963:01 to 1974:12; 1975:01 to 1986:12, and 1987:01 to 1999:09: For these subperiods the mean values of the money supply (with corresponding standard deviations in parentheses) were, respectively, 165.88 (23.27), 323.20 (72.66), 788.12 (195.43), and 1099 (27.84), all figures in billions of dollars. This is a rough indication of the fact that the money supply over the entire period was not stationary.

**Cross-Section Data** Cross-section data are data on one or more variables collected *at the same point in time*, such as the census of population conducted by the Census Bureau every 10 years (the latest being in year 2000), the surveys of consumer expenditures conducted by the University of Michigan, and, of course, the opinion polls by Gallup and umpteen other organizations. A concrete example of cross-sectional data is given in Table 1.1 This table gives data on egg production and egg prices for the 50 states in the union for 1990 and 1991. For each year the data on the 50 states are cross-sectional data. Thus, in Table 1.1 we have two cross-sectional samples.

Just as time series data create their own special problems (because of the stationarity issue), cross-sectional data too have their own problems, specifically the problem of *heterogeneity*. From the data given in Table 1.1 we see that we have some states that produce huge amounts of eggs (e.g., Pennsylvania) and some that produce very little (e.g., Alaska). When we

**TABLE 1.1** U.S. EGG PRODUCTION

| State | $Y_1$ | $Y_2$ | $X_1$ | $X_2$ | State | $Y_1$ | $Y_2$ | $X_1$ | $X_2$ |
|---|---|---|---|---|---|---|---|---|---|
| AL | 2,206 | 2,186 | 92.7 | 91.4 | MT | 172 | 164 | 68.0 | 66.0 |
| AK | 0.7 | 0.7 | 151.0 | 149.0 | NE | 1,202 | 1,400 | 50.3 | 48.9 |
| AZ | 73 | 74 | 61.0 | 56.0 | NV | 2.2 | 1.8 | 53.9 | 52.7 |
| AR | 3,620 | 3,737 | 86.3 | 91.8 | NH | 43 | 49 | 109.0 | 104.0 |
| CA | 7,472 | 7,444 | 63.4 | 58.4 | NJ | 442 | 491 | 85.0 | 83.0 |
| CO | 788 | 873 | 77.8 | 73.0 | NM | 283 | 302 | 74.0 | 70.0 |
| CT | 1,029 | 948 | 106.0 | 104.0 | NY | 975 | 987 | 68.1 | 64.0 |
| DE | 168 | 164 | 117.0 | 113.0 | NC | 3,033 | 3,045 | 82.8 | 78.7 |
| FL | 2,586 | 2,537 | 62.0 | 57.2 | ND | 51 | 45 | 55.2 | 48.0 |
| GA | 4,302 | 4,301 | 80.6 | 80.8 | OH | 4,667 | 4,637 | 59.1 | 54.7 |
| HI | 227.5 | 224.5 | 85.0 | 85.5 | OK | 869 | 830 | 101.0 | 100.0 |
| ID | 187 | 203 | 79.1 | 72.9 | OR | 652 | 686 | 77.0 | 74.6 |
| IL | 793 | 809 | 65.0 | 70.5 | PA | 4,976 | 5,130 | 61.0 | 52.0 |
| IN | 5,445 | 5,290 | 62.7 | 60.1 | RI | 53 | 50 | 102.0 | 99.0 |
| IA | 2,151 | 2,247 | 56.5 | 53.0 | SC | 1,422 | 1,420 | 70.1 | 65.9 |
| KS | 404 | 389 | 54.5 | 47.8 | SD | 435 | 602 | 48.0 | 45.8 |
| KY | 412 | 483 | 67.7 | 73.5 | TN | 277 | 279 | 71.0 | 80.7 |
| LA | 273 | 254 | 115.0 | 115.0 | TX | 3,317 | 3,356 | 76.7 | 72.6 |
| ME | 1,069 | 1,070 | 101.0 | 97.0 | UT | 456 | 486 | 64.0 | 59.0 |
| MD | 885 | 898 | 76.6 | 75.4 | VT | 31 | 30 | 106.0 | 102.0 |
| MA | 235 | 237 | 105.0 | 102.0 | VA | 943 | 988 | 86.3 | 81.2 |
| MI | 1,406 | 1,396 | 58.0 | 53.8 | WA | 1,287 | 1,313 | 74.1 | 71.5 |
| MN | 2,499 | 2,697 | 57.7 | 54.0 | WV | 136 | 174 | 104.0 | 109.0 |
| MS | 1,434 | 1,468 | 87.8 | 86.7 | WI | 910 | 873 | 60.1 | 54.0 |
| MO | 1,580 | 1,622 | 55.4 | 51.5 | WY | 1.7 | 1.7 | 83.0 | 83.0 |

*Note:* $Y_1$ = eggs produced in 1990 (millions)
$Y_2$ = eggs produced in 1991 (millions)
$X_1$ = price per dozen (cents) in 1990
$X_2$ = price per dozen (cents) in 1991
*Source: World Almanac,* 1993, p. 119. The data are from the Economic Research Service, U.S. Department of Agriculture.

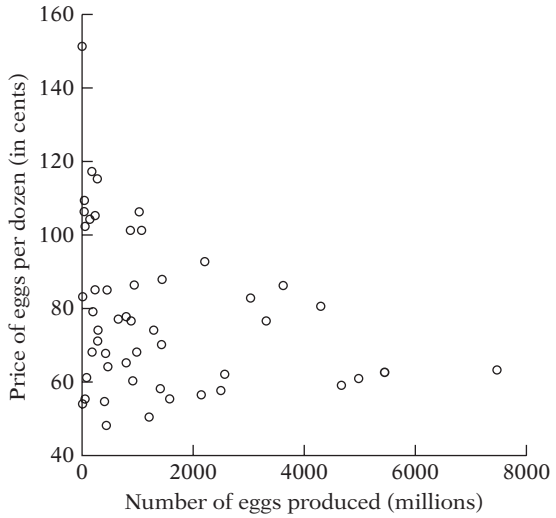


**FIGURE 1.6** Relationship between eggs produced and prices, 1990.

include such heterogeneous units in a statistical analysis, the **size** or **scale effect** must be taken into account so as not to mix apples with oranges. To see this clearly, we plot in Figure 1.6 the data on eggs produced and their prices in 50 states for the year 1990. This figure shows how widely scattered the observations are. In Chapter 11 we will see how the scale effect can be an important factor in assessing relationships among economic variables.

**Pooled Data** In pooled, or combined, data are elements of both time series and cross-section data. The data in Table 1.1 are an example of pooled data. For each year we have 50 cross-sectional observations and for each state we have two time series observations on prices and output of eggs, a total of 100 pooled (or combined) observations. Likewise, the data given in exercise 1.1 are pooled data in that the Consumer Price Index (CPI) for each country for 1973–1997 is time series data, whereas the data on the CPI for the seven countries for a single year are cross-sectional data. In the pooled data we have 175 observations—25 annual observations for each of the seven countries.

**Panel, Longitudinal, or Micropanel Data** This is a special type of pooled data in which the *same* cross-sectional unit (say, a family or a firm) is surveyed over time. For example, the U.S. Department of Commerce carries out a census of housing at periodic intervals. At each periodic survey the same household (or the people living at the same address) is interviewed to find out if there has been any change in the housing and financial conditions of that household since the last survey. By interviewing the same household periodically, the panel data provides very useful information on the dynamics of household behavior, as we shall see in Chapter 16.

### The Sources of Data[12]

The data used in empirical analysis may be collected by a governmental agency (e.g., the Department of Commerce), an international agency (e.g., the International Monetary Fund (IMF) or the World Bank), a private organization (e.g., the Standard & Poor's Corporation), or an individual. Literally, there are thousands of such agencies collecting data for one purpose or another.

**The Internet** The Internet has literally revolutionized data gathering. If you just "surf the net" with a keyword (e.g., exchange rates), you will be swamped with all kinds of data sources. In **Appendix E** we provide some of the frequently visited web sites that provide economic and financial data of all sorts. Most of the data can be downloaded without much cost. You may want to bookmark the various web sites that might provide you with useful economic data.

The data collected by various agencies may be **experimental** or **nonexperimental.** In experimental data, often collected in the natural sciences, the investigator may want to collect data while holding certain factors constant in order to assess the impact of some factors on a given phenomenon. For instance, in assessing the impact of obesity on blood pressure, the researcher would want to collect data while holding constant the eating, smoking, and drinking habits of the people in order to minimize the influence of these variables on blood pressure.

In the social sciences, the data that one generally encounters are nonexperimental in nature, that is, not subject to the control of the researcher.[13] For example, the data on GNP, unemployment, stock prices, etc., are not directly under the control of the investigator. As we shall see, this lack of control often creates special problems for the researcher in pinning down the exact cause or causes affecting a particular situation. For example, is it the money supply that determines the (nominal) GDP or is it the other way round?

### The Accuracy of Data[14]

Although plenty of data are available for economic research, the quality of the data is often not that good. There are several reasons for that. First, as noted, most social science data are nonexperimental in nature. Therefore, there is the possibility of observational errors, either of omission or commission. Second, even in experimentally collected data errors of measurement arise from approximations and roundoffs. Third, in questionnaire-type surveys, the problem of nonresponse can be serious; a researcher is lucky to

[12]For an illuminating account, see Albert T. Somers, *The U.S. Economy Demystified: What the Major Economic Statistics Mean and their Significance for Business,* D.C. Heath, Lexington, Mass., 1985.

[13]In the social sciences too sometimes one can have a controlled experiment. An example is given in exercise 1.6.

[14]For a critical review, see O. Morgenstern, *The Accuracy of Economic Observations*, 2d ed., Princeton University Press, Princeton, N.J., 1963.

get a 40 percent response to a questionnaire. Analysis based on such partial response may not truly reflect the behavior of the 60 percent who did not respond, thereby leading to what is known as (sample) **selectivity bias.** Then there is the further problem that those who respond to the questionnaire may not answer all the questions, especially questions of financially sensitive nature, thus leading to additional selectivity bias. Fourth, the sampling methods used in obtaining the data may vary so widely that it is often difficult to compare the results obtained from the various samples. Fifth, economic data are generally available at a highly aggregate level. For example, most macrodata (e.g., GNP, employment, inflation, unemployment) are available for the economy as a whole or at the most for some broad geographical regions. Such highly aggregated data may not tell us much about the individual or microunits that may be the ultimate object of study. Sixth, because of confidentiality, certain data can be published only in highly aggregate form. The IRS, for example, is not allowed by law to disclose data on individual tax returns; it can only release some broad summary data. Therefore, if one wants to find out how much individuals with a certain level of income spent on health care, one cannot do that analysis except at a very highly aggregate level. But such macroanalysis often fails to reveal the dynamics of the behavior of the microunits. Similarly, the Department of Commerce, which conducts the census of business every 5 years, is not allowed to disclose information on production, employment, energy consumption, research and development expenditure, etc., at the firm level. It is therefore difficult to study the interfirm differences on these items.

Because of all these and many other problems, **the researcher should always keep in mind that the results of research are only as good as the quality of the data.** Therefore, if in given situations researchers find that the results of the research are "unsatisfactory," the cause may be not that they used the wrong model but that the quality of the data was poor. Unfortunately, because of the nonexperimental nature of the data used in most social science studies, researchers very often have no choice but to depend on the available data. But they should always keep in mind that the data used may not be the best and should try not to be too dogmatic about the results obtained from a given study, especially when the quality of the data is suspect.

### A Note on the Measurement Scales of Variables[15]

The variables that we will generally encounter fall into four broad categories: *ratio scale, interval scale, ordinal scale, and nominal scale.* It is important that we understand each.

**Ratio Scale** For a variable $X$, taking two values, $X_1$ and $X_2$, the ratio $X_1/X_2$ and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a

[15]The following discussion relies heavily on Aris Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data,* Cambridge University Press, New York, 1999, p. 24.

natural ordering (ascending or descending) of the values along the scale. Therefore, comparisons such as $X_2 \leq X_1$ or $X_2 \geq X_1$ are meaningful. Most economic variables belong to this category. Thus, it is meaningful to ask how big is this year's GDP compared with the previous year's GDP.

**Interval Scale** An interval scale variable satisfies the last two properties of the ratio scale variable but not the first. Thus, the distance between two time periods, say (2000–1995) is meaningful, but not the ratio of two time periods (2000/1995).

**Ordinal Scale** A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering). Examples are grading systems (A, B, C grades) or income class (upper, middle, lower). For these variables the ordering exists but the distances between the categories cannot be quantified. Students of economics will recall the *indifference curves* between two goods, each higher indifference curve indicating higher level of utility, but one cannot quantify by how much one indifference curve is higher than the others.

**Nominal Scale** Variables in this category have none of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories. *Question:* What is the reason why such variables cannot be expressed on the ratio, interval, or ordinal scales?

As we shall see, econometric techniques that may be suitable for ratio scale variables may not be suitable for nominal scale variables. Therefore, it is important to bear in mind the distinctions among the four types of measurement scales discussed above.

## 1.8 SUMMARY AND CONCLUSIONS

**1.** The key idea behind regression analysis is the statistical dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables.

**2.** The objective of such analysis is to estimate and/or predict the mean or average value of the dependent variable on the basis of the known or fixed values of the explanatory variables.

**3.** In practice the success of regression analysis depends on the availability of the appropriate data. This chapter discussed the nature, sources, and limitations of the data that are generally available for research, especially in the social sciences.

**4.** In any research, the researcher should clearly state the sources of the data used in the analysis, their definitions, their methods of collection, and any gaps or omissions in the data as well as any revisions in the data. Keep in mind that the macroeconomic data published by the government are often revised.

**5.** Since the reader may not have the time, energy, or resources to track down the data, the reader has the right to presume that the data used by the researcher are properly gathered and that the computations and analysis are correct.

## EXERCISES

**1.1.** Table 1.2 gives data on the Consumer Price Index (CPI) for seven industrialized countries with 1982–1984 = 100 as the base of the index.
- **a.** From the given data, compute the inflation rate for each country.[16]
- **b.** Plot the inflation rate for each country against time (i.e., use the horizontal axis for time and the vertical axis for the inflation rate.)
- **c.** What broad conclusions can you draw about the inflation experience in the seven countries?
- **d.** Which country's inflation rate seems to be most variable? Can you offer any explanation?

**TABLE 1.2** CPI IN SEVEN INDUSTRIAL COUNTRIES, 1973–1997 (1982−1984 = 100)

| Year | Canada | France | Germany | Italy | Japan | U.K. | U.S. |
|---|---|---|---|---|---|---|---|
| 1973 | 40.80000 | 34.60000 | 62.80000 | 20.60000 | 47.90000 | 27.90000 | 44.40000 |
| 1974 | 45.20000 | 39.30000 | 67.10000 | 24.60000 | 59.00000 | 32.30000 | 49.30000 |
| 1975 | 50.10000 | 43.90000 | 71.10000 | 28.80000 | 65.90000 | 40.20000 | 53.80000 |
| 1976 | 53.90000 | 48.10000 | 74.20000 | 33.60000 | 72.20000 | 46.80000 | 56.90000 |
| 1977 | 58.10000 | 52.70000 | 76.90000 | 40.10000 | 78.10000 | 54.20000 | 60.60000 |
| 1978 | 63.30000 | 57.50000 | 79.00000 | 45.10000 | 81.40000 | 58.70000 | 65.20000 |
| 1979 | 69.20000 | 63.60000 | 82.20000 | 52.10000 | 84.40000 | 66.60000 | 72.60000 |
| 1980 | 76.10000 | 72.30000 | 86.70000 | 63.20000 | 90.90000 | 78.50000 | 82.40000 |
| 1981 | 85.60000 | 81.90000 | 92.20000 | 75.40000 | 95.30000 | 87.90000 | 90.90000 |
| 1982 | 94.90000 | 91.70000 | 97.10000 | 87.70000 | 98.10000 | 95.40000 | 96.50000 |
| 1983 | 100.4000 | 100.4000 | 100.3000 | 100.8000 | 99.80000 | 99.80000 | 99.60000 |
| 1984 | 104.7000 | 108.1000 | 102.7000 | 111.5000 | 102.1000 | 104.8000 | 103.9000 |
| 1985 | 109.0000 | 114.4000 | 104.8000 | 121.1000 | 104.1000 | 111.1000 | 107.6000 |
| 1986 | 113.5000 | 117.3000 | 104.7000 | 128.5000 | 104.8000 | 114.9000 | 109.6000 |
| 1987 | 118.4000 | 121.1000 | 104.9000 | 134.4000 | 104.8000 | 119.7000 | 113.6000 |
| 1988 | 123.2000 | 124.4000 | 106.3000 | 141.1000 | 105.6000 | 125.6000 | 118.3000 |
| 1989 | 129.3000 | 128.7000 | 109.2000 | 150.4000 | 108.1000 | 135.3000 | 124.0000 |
| 1990 | 135.5000 | 133.0000 | 112.2000 | 159.6000 | 111.4000 | 148.2000 | 130.7000 |
| 1991 | 143.1000 | 137.2000 | 116.3000 | 169.8000 | 115.0000 | 156.9000 | 136.2000 |
| 1992 | 145.3000 | 140.5000 | 122.1000 | 178.8000 | 116.9000 | 162.7000 | 140.3000 |
| 1993 | 147.9000 | 143.5000 | 127.6000 | 186.4000 | 118.4000 | 165.3000 | 144.5000 |
| 1994 | 148.2000 | 145.8000 | 131.1000 | 193.7000 | 119.3000 | 169.4000 | 148.2000 |
| 1995 | 151.4000 | 148.4000 | 133.5000 | 204.1000 | 119.1000 | 175.1000 | 152.4000 |
| 1996 | 153.8000 | 151.4000 | 135.5000 | 212.0000 | 119.3000 | 179.4000 | 156.9000 |
| 1997 | 156.3000 | 153.2000 | 137.8000 | 215.7000 | 121.3000 | 185.0000 | 160.5000 |

[16]Subtract from the current year's CPI the CPI from the previous year, divide the difference by the previous year's CPI, and multiply the result by 100. Thus, the inflation rate for Canada for 1974 is $[(45.2 - 40.8)/40.8] \times 100 = 10.78\%$ (approx.).

**1.2.** **a.** Plot the inflation rate of Canada, France, Germany, Italy, Japan, and the United Kingdom against the United States inflation rate.
**b.** Comment generally about the behavior of the inflation rate in the six countries vis-à-vis the U.S. inflation rate.
**c.** If you find that the six countries' inflation rates move in the same direction as the U.S. inflation rate, would that suggest that U.S. inflation "causes" inflation in the other countries? Why or why not?

**1.3.** Table 1.3 gives the foreign exchange rates for seven industrialized countries for years 1977–1998. Except for the United Kingdom, the exchange rate is defined as the units of foreign currency for one U.S. dollar; for the United Kingdom, it is defined as the number of U.S. dollars for one U.K. pound.
**a.** Plot these exchange rates against time and comment on the general behavior of the exchange rates over the given time period.
**b.** The dollar is said to *appreciate* if it can buy more units of a foreign currency. Contrarily, it is said to depreciate if it buys fewer units of a foreign currency. Over the time period 1977–1998, what has been the general behavior of the U.S. dollar? Incidentally, look up any textbook on macroeconomics or international economics to find out what factors determine the appreciation or depreciation of a currency.

**1.4.** The data behind the M1 money supply in Figure 1.5 are given in Table 1.4. Can you give reasons why the money supply has been increasing over the time period shown in the table?

**TABLE 1.3** EXCHANGE RATES FOR SEVEN COUNTRIES: 1977–1998

| Year | Canada | France | Germany | Japan | Sweden | Switzerland | U.K. |
|---|---|---|---|---|---|---|---|
| 1977 | 1.063300 | 4.916100 | 2.323600 | 268.6200 | 4.480200 | 2.406500 | 1.744900 |
| 1978 | 1.140500 | 4.509100 | 2.009700 | 210.3900 | 4.520700 | 1.790700 | 1.918400 |
| 1979 | 1.171300 | 4.256700 | 1.834300 | 219.0200 | 4.289300 | 1.664400 | 2.122400 |
| 1980 | 1.169300 | 4.225100 | 1.817500 | 226.6300 | 4.231000 | 1.677200 | 2.324600 |
| 1981 | 1.199000 | 5.439700 | 2.263200 | 220.6300 | 5.066000 | 1.967500 | 2.024300 |
| 1982 | 1.234400 | 6.579400 | 2.428100 | 249.0600 | 6.283900 | 2.032700 | 1.748000 |
| 1983 | 1.232500 | 7.620400 | 2.553900 | 237.5500 | 7.671800 | 2.100700 | 1.515900 |
| 1984 | 1.295200 | 8.735600 | 2.845500 | 237.4600 | 8.270800 | 2.350000 | 1.336800 |
| 1985 | 1.365900 | 8.980000 | 2.942000 | 238.4700 | 8.603200 | 2.455200 | 1.297400 |
| 1986 | 1.389600 | 6.925700 | 2.170500 | 168.3500 | 7.127300 | 1.797900 | 1.467700 |
| 1987 | 1.325900 | 6.012200 | 1.798100 | 144.6000 | 6.346900 | 1.491800 | 1.639800 |
| 1988 | 1.230600 | 5.959500 | 1.757000 | 128.1700 | 6.137000 | 1.464300 | 1.781300 |
| 1989 | 1.184200 | 6.380200 | 1.880800 | 138.0700 | 6.455900 | 1.636900 | 1.638200 |
| 1990 | 1.166800 | 5.446700 | 1.616600 | 145.0000 | 5.923100 | 1.390100 | 1.784100 |
| 1991 | 1.146000 | 5.646800 | 1.661000 | 134.5900 | 6.052100 | 1.435600 | 1.767400 |
| 1992 | 1.208500 | 5.293500 | 1.561800 | 126.7800 | 5.825800 | 1.406400 | 1.766300 |
| 1993 | 1.290200 | 5.666900 | 1.654500 | 111.0800 | 7.795600 | 1.478100 | 1.501600 |
| 1994 | 1.366400 | 5.545900 | 1.621600 | 102.1800 | 7.716100 | 1.366700 | 1.531900 |
| 1995 | 1.372500 | 4.986400 | 1.432100 | 93.96000 | 7.140600 | 1.181200 | 1.578500 |
| 1996 | 1.363800 | 5.115800 | 1.504900 | 108.7800 | 6.708200 | 1.236100 | 1.560700 |
| 1997 | 1.384900 | 5.839300 | 1.734800 | 121.0600 | 7.644600 | 1.451400 | 1.637600 |
| 1998 | 1.483600 | 5.899500 | 1.759700 | 130.9900 | 7.952200 | 1.450600 | 1.657300 |

*Source: Economic Report of the President,* January 2000 and January 2001.

**TABLE 1.4** SEASONALLY ADJUSTED M1 SUPPLY: 1959:01–1999:09 (BILLIONS OF DOLLARS)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1959:01 | 138.8900 | 139.3900 | 139.7400 | 139.6900 | 140.6800 | 141.1700 |
| 1959:07 | 141.7000 | 141.9000 | 141.0100 | 140.4700 | 140.3800 | 139.9500 |
| 1960:01 | 139.9800 | 139.8700 | 139.7500 | 139.5600 | 139.6100 | 139.5800 |
| 1960:07 | 140.1800 | 141.3100 | 141.1800 | 140.9200 | 140.8600 | 140.6900 |
| 1961:01 | 141.0600 | 141.6000 | 141.8700 | 142.1300 | 142.6600 | 142.8800 |
| 1961:07 | 142.9200 | 143.4900 | 143.7800 | 144.1400 | 144.7600 | 145.2000 |
| 1962:01 | 145.2400 | 145.6600 | 145.9600 | 146.4000 | 146.8400 | 146.5800 |
| 1962:07 | 146.4600 | 146.5700 | 146.3000 | 146.7100 | 147.2900 | 147.8200 |
| 1963:01 | 148.2600 | 148.9000 | 149.1700 | 149.7000 | 150.3900 | 150.4300 |
| 1963:07 | 151.3400 | 151.7800 | 151.9800 | 152.5500 | 153.6500 | 153.2900 |
| 1964:01 | 153.7400 | 154.3100 | 154.4800 | 154.7700 | 155.3300 | 155.6200 |
| 1964:07 | 156.8000 | 157.8200 | 158.7500 | 159.2400 | 159.9600 | 160.3000 |
| 1965:01 | 160.7100 | 160.9400 | 161.4700 | 162.0300 | 161.7000 | 162.1900 |
| 1965:07 | 163.0500 | 163.6800 | 164.8500 | 165.9700 | 166.7100 | 167.8500 |
| 1966:01 | 169.0800 | 169.6200 | 170.5100 | 171.8100 | 171.3300 | 171.5700 |
| 1966:07 | 170.3100 | 170.8100 | 171.9700 | 171.1600 | 171.3800 | 172.0300 |
| 1967:01 | 171.8600 | 172.9900 | 174.8100 | 174.1700 | 175.6800 | 177.0200 |
| 1967:07 | 178.1300 | 179.7100 | 180.6800 | 181.6400 | 182.3800 | 183.2600 |
| 1968:01 | 184.3300 | 184.7100 | 185.4700 | 186.6000 | 187.9900 | 189.4200 |
| 1968:07 | 190.4900 | 191.8400 | 192.7400 | 194.0200 | 196.0200 | 197.4100 |
| 1969:01 | 198.6900 | 199.3500 | 200.0200 | 200.7100 | 200.8100 | 201.2700 |
| 1969:07 | 201.6600 | 201.7300 | 202.1000 | 202.9000 | 203.5700 | 203.8800 |
| 1970:01 | 206.2200 | 205.0000 | 205.7500 | 206.7200 | 207.2200 | 207.5400 |
| 1970:07 | 207.9800 | 209.9300 | 211.8000 | 212.8800 | 213.6600 | 214.4100 |
| 1971:01 | 215.5400 | 217.4200 | 218.7700 | 220.0000 | 222.0200 | 223.4500 |
| 1971:07 | 224.8500 | 225.5800 | 226.4700 | 227.1600 | 227.7600 | 228.3200 |
| 1972:01 | 230.0900 | 232.3200 | 234.3000 | 235.5800 | 235.8900 | 236.6200 |
| 1972:07 | 238.7900 | 240.9300 | 243.1800 | 245.0200 | 246.4100 | 249.2500 |
| 1973:01 | 251.4700 | 252.1500 | 251.6700 | 252.7400 | 254.8900 | 256.6900 |
| 1973:07 | 257.5400 | 257.7600 | 257.8600 | 259.0400 | 260.9800 | 262.8800 |
| 1974:01 | 263.7600 | 265.3100 | 266.6800 | 267.2000 | 267.5600 | 268.4400 |
| 1974:07 | 269.2700 | 270.1200 | 271.0500 | 272.3500 | 273.7100 | 274.2000 |
| 1975:01 | 273.9000 | 275.0000 | 276.4200 | 276.1700 | 279.2000 | 282.4300 |
| 1975:07 | 283.6800 | 284.1500 | 285.6900 | 285.3900 | 286.8300 | 287.0700 |
| 1976:01 | 288.4200 | 290.7600 | 292.7000 | 294.6600 | 295.9300 | 296.1600 |
| 1976:07 | 297.2000 | 299.0500 | 299.6700 | 302.0400 | 303.5900 | 306.2500 |
| 1977:01 | 308.2600 | 311.5400 | 313.9400 | 316.0200 | 317.1900 | 318.7100 |
| 1977:07 | 320.1900 | 322.2700 | 324.4800 | 326.4000 | 328.6400 | 330.8700 |
| 1978:01 | 334.4000 | 335.3000 | 336.9600 | 339.9200 | 344.8600 | 346.8000 |
| 1978:07 | 347.6300 | 349.6600 | 352.2600 | 353.3500 | 355.4100 | 357.2800 |
| 1979:01 | 358.6000 | 359.9100 | 362.4500 | 368.0500 | 369.5900 | 373.3400 |
| 1979:07 | 377.2100 | 378.8200 | 379.2800 | 380.8700 | 380.8100 | 381.7700 |
| 1980:01 | 385.8500 | 389.7000 | 388.1300 | 383.4400 | 384.6000 | 389.4600 |
| 1980:07 | 394.9100 | 400.0600 | 405.3600 | 409.0600 | 410.3700 | 408.0600 |
| 1981:01 | 410.8300 | 414.3800 | 418.6900 | 427.0600 | 424.4300 | 425.5000 |
| 1981:07 | 427.9000 | 427.8500 | 427.4600 | 428.4500 | 430.8800 | 436.1700 |
| 1982:01 | 442.1300 | 441.4900 | 442.3700 | 446.7800 | 446.5300 | 447.8900 |
| 1982:07 | 449.0900 | 452.4900 | 457.5000 | 464.5700 | 471.1200 | 474.3000 |
| 1983:01 | 476.6800 | 483.8500 | 490.1800 | 492.7700 | 499.7800 | 504.3500 |
| 1983:07 | 508.9600 | 511.6000 | 513.4100 | 517.2100 | 518.5300 | 520.7900 |

(*Continued*)

**TABLE 1.4** (*Continued*)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1984:01 | 524.4000 | 526.9900 | 530.7800 | 534.0300 | 536.5900 | 540.5400 |
| 1984:07 | 542.1300 | 542.3900 | 543.8600 | 543.8700 | 547.3200 | 551.1900 |
| 1985:01 | 555.6600 | 562.4800 | 565.7400 | 569.5500 | 575.0700 | 583.1700 |
| 1985:07 | 590.8200 | 598.0600 | 604.4700 | 607.9100 | 611.8300 | 619.3600 |
| 1986:01 | 620.4000 | 624.1400 | 632.8100 | 640.3500 | 652.0100 | 661.5200 |
| 1986:07 | 672.2000 | 680.7700 | 688.5100 | 695.2600 | 705.2400 | 724.2800 |
| 1987:01 | 729.3400 | 729.8400 | 733.0100 | 743.3900 | 746.0000 | 743.7200 |
| 1987:07 | 744.9600 | 746.9600 | 748.6600 | 756.5000 | 752.8300 | 749.6800 |
| 1988:01 | 755.5500 | 757.0700 | 761.1800 | 767.5700 | 771.6800 | 779.1000 |
| 1988:07 | 783.4000 | 785.0800 | 784.8200 | 783.6300 | 784.4600 | 786.2600 |
| 1989:01 | 784.9200 | 783.4000 | 782.7400 | 778.8200 | 774.7900 | 774.2200 |
| 1989:07 | 779.7100 | 781.1400 | 782.2000 | 787.0500 | 787.9500 | 792.5700 |
| 1990:01 | 794.9300 | 797.6500 | 801.2500 | 806.2400 | 804.3600 | 810.3300 |
| 1990:07 | 811.8000 | 817.8500 | 821.8300 | 820.3000 | 822.0600 | 824.5600 |
| 1991:01 | 826.7300 | 832.4000 | 838.6200 | 842.7300 | 848.9600 | 858.3300 |
| 1991:07 | 862.9500 | 868.6500 | 871.5600 | 878.4000 | 887.9500 | 896.7000 |
| 1992:01 | 910.4900 | 925.1300 | 936.0000 | 943.8900 | 950.7800 | 954.7100 |
| 1992:07 | 964.6000 | 975.7100 | 988.8400 | 1004.340 | 1016.040 | 1024.450 |
| 1993:01 | 1030.900 | 1033.150 | 1037.990 | 1047.470 | 1066.220 | 1075.610 |
| 1993:07 | 1085.880 | 1095.560 | 1105.430 | 1113.800 | 1123.900 | 1129.310 |
| 1994:01 | 1132.200 | 1136.130 | 1139.910 | 1141.420 | 1142.850 | 1145.650 |
| 1994:07 | 1151.490 | 1151.390 | 1152.440 | 1150.410 | 1150.440 | 1149.750 |
| 1995:01 | 1150.640 | 1146.740 | 1146.520 | 1149.480 | 1144.650 | 1144.240 |
| 1995:07 | 1146.500 | 1146.100 | 1142.270 | 1136.430 | 1133.550 | 1126.730 |
| 1996:01 | 1122.580 | 1117.530 | 1122.590 | 1124.520 | 1116.300 | 1115.470 |
| 1996:07 | 1112.340 | 1102.180 | 1095.610 | 1082.560 | 1080.490 | 1081.340 |
| 1997:01 | 1080.520 | 1076.200 | 1072.420 | 1067.450 | 1063.370 | 1065.990 |
| 1997:07 | 1067.570 | 1072.080 | 1064.820 | 1062.060 | 1067.530 | 1074.870 |
| 1998:01 | 1073.810 | 1076.020 | 1080.650 | 1082.090 | 1078.170 | 1077.780 |
| 1998:07 | 1075.370 | 1072.210 | 1074.650 | 1080.400 | 1088.960 | 1093.350 |
| 1999:01 | 1091.000 | 1092.650 | 1102.010 | 1108.400 | 1104.750 | 1101.110 |
| 1999:07 | 1099.530 | 1102.400 | 1093.460 | | | |

*Source:* Board of Governors, Federal Reserve Bank, USA.

**1.5.** Suppose you were to develop an economic model of criminal activities, say, the hours spent in criminal activities (e.g., selling illegal drugs). What variables would you consider in developing such a model? See if your model matches the one developed by the Nobel laureate economist Gary Becker.[17]

**1.6.** *Controlled experiments in economics:* On April 7, 2000, President Clinton signed into law a bill passed by both Houses of the U.S. Congress that lifted earnings limitations on Social Security recipients. Until then, recipients between the ages of 65 and 69 who earned more than $17,000 a year would lose 1 dollar's worth of Social Security benefit for every 3 dollars of income earned in excess of $17,000. How would you devise a study to assess the impact of this change in the law? *Note:* There was no income limitation for recipients over the age of 70 under the old law.

[17]G. S. Becker, "Crime and Punishment: An Economic Approach," *Journal of Political Economy,* vol. 76, 1968, pp. 169–217.

**TABLE 1.5** IMPACT OF ADVERTISING EXPENDITURE

| Firm | Impressions, millions | Expenditure, millions of 1983 dollars |
|---|---|---|
| 1. Miller Lite | 32.1 | 50.1 |
| 2. Pepsi | 99.6 | 74.1 |
| 3. Stroh's | 11.7 | 19.3 |
| 4. Fed'l Express | 21.9 | 22.9 |
| 5. Burger King | 60.8 | 82.4 |
| 6. Coca Cola | 78.6 | 40.1 |
| 7. McDonald's | 92.4 | 185.9 |
| 8. MCl | 50.7 | 26.9 |
| 9. Diet Cola | 21.4 | 20.4 |
| 10. Ford | 40.1 | 166.2 |
| 11. Levi's | 40.8 | 27.0 |
| 12. Bud Lite | 10.4 | 45.6 |
| 13. ATT/Bell | 88.9 | 154.9 |
| 14. Calvin Klein | 12.0 | 5.0 |
| 15. Wendy's | 29.2 | 49.7 |
| 16. Polaroid | 38.0 | 26.9 |
| 17. Shasta | 10.0 | 5.7 |
| 18. Meow Mix | 12.3 | 7.6 |
| 19. Oscar Meyer | 23.4 | 9.2 |
| 20. Crest | 71.1 | 32.4 |
| 21. Kibbles 'N Bits | 4.4 | 6.1 |

*Source:* http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html

**1.7.** The data presented in Table 1.5 was published in the March 1, 1984 issue of the *Wall Street Journal*. It relates to the advertising budget (in millions of dollars) of 21 firms for 1983 and millions of impressions retained per week by the viewers of the products of these firms. The data are based on a survey of 4000 adults in which users of the products were asked to cite a commercial they had seen for the product category in the past week.

**a.** Plot impressions on the vertical axis and advertising expenditure on the horizontal axis.

**b.** What can you say about the nature of the relationship between the two variables?

**c.** Looking at your graph, do you think it pays to advertise? Think about all those commercials shown on Super Bowl Sunday or during the World Series.

*Note:* We will explore further the data given in Table 1.5 in subsequent chapters.