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.

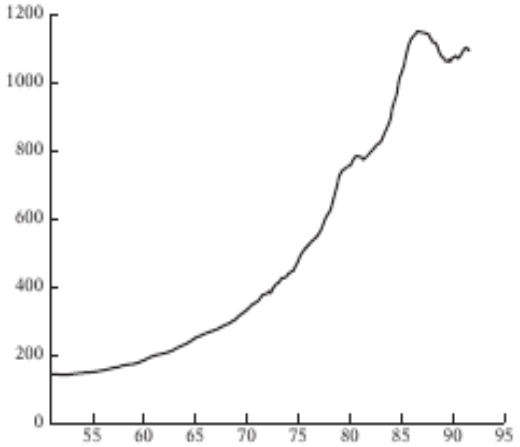

FIGURE 1.5 M1 money supply: United States, 1951:01–1999:09.

[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.

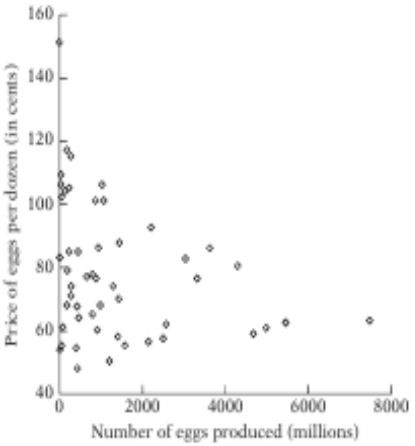

FIGURE 1.6 Relationship between eggs produced and prices, 1990.

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.