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.

Although it is a matter of personal taste and tradition, here 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 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 regressions 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.

Unless stated otherwise, the letter Y will denote the dependent variable and the X's (X_1, X_2,..., 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). Let us now discuss 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.

1.7 THE NATURE AND SOURCES OF DATA FOR ECONOMIC ANALYSIS

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 introductory data shown are an examples 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. government budgets), **quinquennially**, that is, every 5 years (e.g., the census of manufactures), or **decennially** that is, every 10 years (e.g., the census of population).

Sometimes 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. In fact, **time series econometrics** most based on 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 it's mean and variance do not vary systematically over time.*

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 surveys of consumer expenditures conducted by the University of Michigan, and, of course, the opinion polls by Gallup and other organizations.

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. Generally, we see that there are some states that produce huge amounts of eggs and some that produce very little. When we 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, the data on eggs produced and their prices show how widely it's scattered.

Pooled Data

In pooled, or combined, data are elements of both time series and cross-section data. The above data 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 given data are pooled data in that the Consumer Price Index (CPI) is time series data, whereas the data on the CPI are cross-sectional data.

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. 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 behaviour.

To be continued ...