The Nature and Sources of Data for Economic Analysis10

Types of Data:

1. Time Series Data:

A time series is a **data set that tracks a sample over time**. In particular, a time series allows one to see what factors influence certain variables from period to period. Time series analysis can be useful to see how a given asset, security, or economic variable changes over time

Example:

**Weather records, economic indicators and patient health evolution metrics** — all are time series data. ... In investing, a time series tracks the movement of data points, such as a security's price over a specified period of time with data points recorded at regular intervals.

1. *Cross-Section Data*

Cross-sectional data are **the result of a data collection, carried out at a single point in time on a statistical unit**. With cross-sectional data, we are not interested in the change of data over time, but in the current, valid opinion of the respondents about a question in a survey.

Example:

if we want to measure current obesity levels in a population, we could draw a sample of 1,000 people randomly from that population (also known as a cross section of that population), measure their weight and height, and calculate what percentage of that sample is categorized as obese. This cross-sectional sample provides us with a snapshot of that population, at that one point in time. Note that we do not know based on one cross-sectional sample if obesity is increasing or decreasing; we can only describe the current proportion.

1. *Pooled Data*

Data pooling is **a process where data sets coming from different sources are combined**. ... First, that multiple datasets containing information on many patients from different countries or from different institutions is merged into one data file.

Example:

Pooled data is a mixture of time series data and cross-section data. One example is **GNP per capita of all European countries over ten years**. Panel, longitudinal or micropanel data is a type that is pooled data of nature.

1. *Panel, Longitudinal, or Micropanel Data*

Panel data, sometimes referred to as longitudinal data, is **data that contains observations about different cross sections across time**.

**Or**

**Longitudinal data**, sometimes called panel data, is data that is collected through a series of repeated observations of the same subjects over some extended time frame—and is useful for measuring change.

Example:

Examples of groups that may make up panel data series include countries, firms, individuals, or demographic groups.

Source of Data:

**Internet**

The Accuracy of Data

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.

1. As noted, most social science data are nonexperimental in nature. Therefore, there is the  
possibility of observational errors, either of omission or commission.

2. Even in experimentally collected data, errors of measurement arise from approximations and roundoffs.

3. In questionnaire-type surveys, the problem of nonresponse can be serious; a researcher is lucky to get a 40 percent response rate to a questionnaire. Analysis based on such a partial response rate 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 do respond to the questionnaire may not answer all the questions, especially questions of a financially sensitive nature, thus leading to additional selectivity bias.

4. 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.

5. Economic data are generally available at a highly aggregate level.

6. Because of confidentiality, certain data can be published only in highly aggregate form.

Measurement Scales

Ratio Scale

Interval Scale

Ordinal Scale

Nominal Scale

CHAPTER 2:

Two-Variable Regression Analysis: Some Basic Ideas

The Significance of the Stochastic Disturbance Term

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

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

2. *Unavailability of data:* Even if we know what some of the excluded variables are and  
therefore consider a multiple regression rather than a simple regression, we may not have  
quantitative information about these variables. It is a common experience in empirical  
analysis that the data we would ideally like to have often are not available. For example, in  
principle we could introduce family wealth as an explanatory variable in addition to the income variable to explain family consumption expenditure. But unfortunately, information  
on family wealth generally is not available. Therefore, we may be forced to omit the wealth  
variable from our model despite its great theoretical relevance in explaining consumption  
expenditure.

3. *Core variables versus peripheral variables:* Assume in our consumption-income example that besides income *X*1, the number of children per family *X*2, sex *X*3, religion *X*4,  
education *X*5, and geographical region *X*6 also affect consumption expenditure. But it is quite  
possible that the joint influence of all or some of these variables may be so small and at best  
nonsystematic or random that as a practical matter and for cost considerations it does not pay  
to introduce them into the model explicitly. One hopes that their combined effect can be  
treated as a random variable *u i*.10

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

5. *Poor proxy variables:* Although the classical regression model (to be developed in  
Chapter 3) assumes that the variables *Y* and *X* are measured accurately, in practice the data  
9As a matter of fact, in the method of least squares to be developed in Chapter 3, it is assumed  
explicitly that *E*(*ui*|*Xi*) = 0. See Sec. 3.2.  
10A further difficulty is that variables such as sex, education, and religion are difficult to quantify.  
**42** Part One *Single-Equation Regression Models*may be plagued by errors of measurement. Consider, for example, Milton Friedman’s wellknown theory of the consumption function.11 He regards *permanent consumption* (*Y p*) as  
a function of *permanent income* (*X p*)*.* But since data on these variables are not directly observable, in practice we use proxy variables, such as current consumption (*Y*) and current  
income (*X*), which can be observable. Since the observed *Y* and *X* may not equal *Y p* and  
*X p*, there is the problem of errors of measurement. The disturbance term *u* may in this case  
then also represent the errors of measurement. As we will see in a later chapter, if there are  
such errors of measurement, they can have serious implications for estimating the regression coefficients, the *β*’s.

6. *Principle of parsimony:* Following Occam’s razor,12 we would like to keep our regression model as simple as possible. If we can explain the behavior of *Y* “substantially”  
with two or three explanatory variables and if our theory is not strong enough to suggest  
what other variables might be included, why introduce more variables? Let *u i* represent all  
other variables. Of course, we should not exclude relevant and important variables just to  
keep the regression model simple.

7. *Wrong functional form:* Even if we have theoretically correct variables explaining a  
phenomenon and even if we can obtain data on these variables, very often we do not know  
the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (invariable) function of income or a nonlinear (invariable)  
function? If it is the former, *Yi* = *β*1 + *β*2 *Xi* + *u i* is the proper functional relationship  
between *Y* and *X*, but if it is the latter, *Yi* = *β*1 + *β*2 *Xi* + *β*3 *Xi*2 + *u i* may be the correct  
functional form. In two-variable models the functional form of the relationship can often  
be judged from the scattergram. But in a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize scattergrams in  
multiple dimensions