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-**

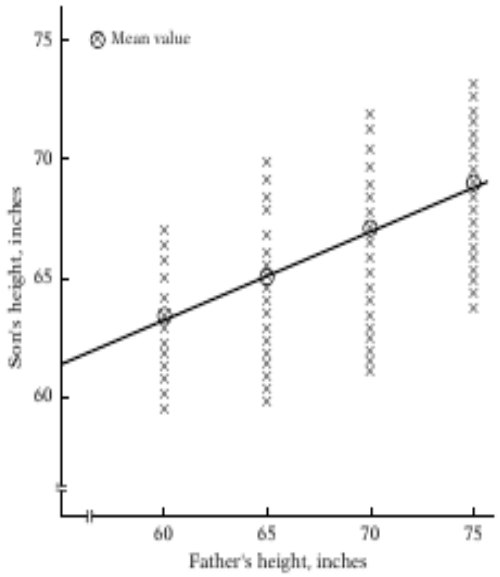

FIGURE 1.1 Hypothetical distribution of sons' heights corresponding to given heights of fathers.

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

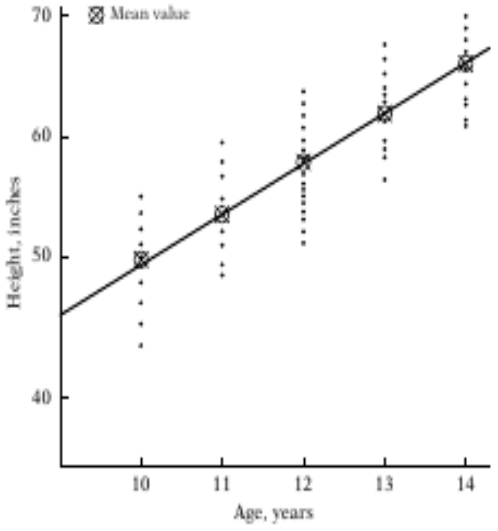

FIGURE 1.2 Hypothetical distribution of heights corresponding to selected ages.

[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.

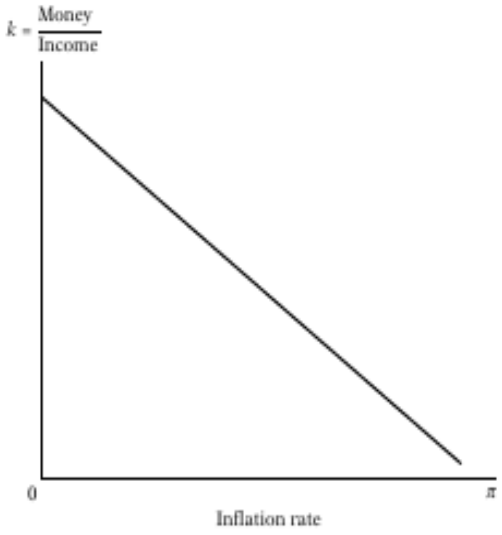

FIGURE 1.4 Money holding in relation to the inflation rate π.

6. From monetary economics it is known that, other things remaining the same, the higher the rate of inflation π, 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 kth explanatory variable. The subscript i or t will denote the ith or the tth observation or value. X_{ki} (or X_{kt}) will denote the ith (or tth) 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.