# 1. Definition, Scope and Division of Econometrics

## 1.1. DEFINITION AND SCOPE OF ECONOMETRICS

Econometrics deals with the measurement of economic relationships. The term 'econometrics' is formed from two words of Greek origin, οἰκονομια (economy), and μέτρον (measure).

Econometrics is a combination of economic theory, mathematical economics and statistics, but it is completely distinct from each one of these three branches of science.

The following quotation from the opening editorial of *Econometrica* written by R. Frish in 1933 may give a clear idea of the scope and method of econometrics:

> But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definite quantitative character. Nor should econometrics be taken as synonymous with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.

Thus econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships (for example, elasticities, propensities, marginal values) and verifying economic theories. It is a special type of economic analysis and research in which the general economic theory, formulated in mathematical terms, is combined with empirical measurement of economic phenomena. Starting from the relationships of economic theory, we express them in mathematical terms (i.e. we build a model) so that they can be measured. We then use specific methods, called *econometric methods*, in order to obtain numerical estimates of the coefficients of the economic relationships. Econometric methods are statistical methods specifically adapted to the peculiarities of economic phenomena. The most important characteristic of economic relationships is that they contain a random element, which, however, is ignored by economic theory and mathematical economics which postulate exact relationships between the various economic magnitudes. Econometrics has developed methods for dealing with the random component of economic relationships.

An example will make the above clear. Economic theory postulates that the demand for a commodity depends on its price, on the prices of other commodities, on consumers' income and on tastes. This is an exact relationship, because it implies that demand is completely determined by the above four factors. No other factor, except those explicitly mentioned, influences the demand. In mathematical economics we express the above abstract economic relationship of demand in mathematical terms. Thus we may write the following demand equation

$$Q = b_0 + b_1 P + b_2 P_0 + b_3 Y + b_4 t$$

where $Q$ = quantity demanded of a particular commodity
$P$ = price of the commodity
$P_0$ = prices of other commodities
$Y$ = consumers' income
$t$ = tastes
$b_0, b_1, b_2, b_3, b_4$ = coefficients of the demand equation.

The above demand equation is exact, because it implies that the only determinants of the quantity demanded are the four factors which appear in the right-hand side of the equation. Quantity will change only if some of these factors change. No other factor may have any effect on demand. Yet it is common knowledge that in economic life many more factors may affect demand. The invention of a new product, a war, professional changes, institutional changes, changes in law, changes in income distribution, massive population movements (migration), etc., are examples of such factors. Furthermore, human behaviour is inherently erratic. We are influenced by rumours, dreams, prejudices, traditions and other psychological and sociological factors, which make us behave differently even though the conditions in the market (prices) and our incomes remain the same. In econometrics the influence of these 'other' factors is taken into account by the introduction into the economic relationships of a random variable, with specific characteristics, which will be discussed in later chapters. In our example the demand function studied with the tools of econometrics would be of the (stochastic) form

$$Q = b_0 + b_1 P + b_2 P_0 + b_3 Y + b_4 t + u$$

where $u$ stands for the random factors which affect the quantity demanded.

It is essential to stress that econometrics presupposes the existence of a body of economic theory. Economic theory should come first, because it sets the hypotheses about economic behaviour which should be tested with the application of econometric techniques. In testing a theory we start from its mathematical formulation, which constitutes *the model* or the *maintained hypothesis*. In our



example of the demand function the maintained hypothesis is

$$Q = b_0 + b_1 P + b_2 P_0 + b_3 Y + b_4 t + u$$

The next step is to confront the model with observational data referring to the actual behaviour of the economic units — consumers or producers. The aim of this stage is to establish whether the theory can explain the actual behaviour of the economic units, i.e. whether the theory is compatible with the facts. If the theory is compatible with the actual data, we accept the theory as valid. If the theory is incompatible with the observed behaviour, we either reject the theory or, in the light of the empirical evidence of the data, we may modify it. In the latter case one needs additional new observations in order to test the revised version of the theory.

The procedure to be followed when testing a theory may be schematically presented as in Figure 1.1.

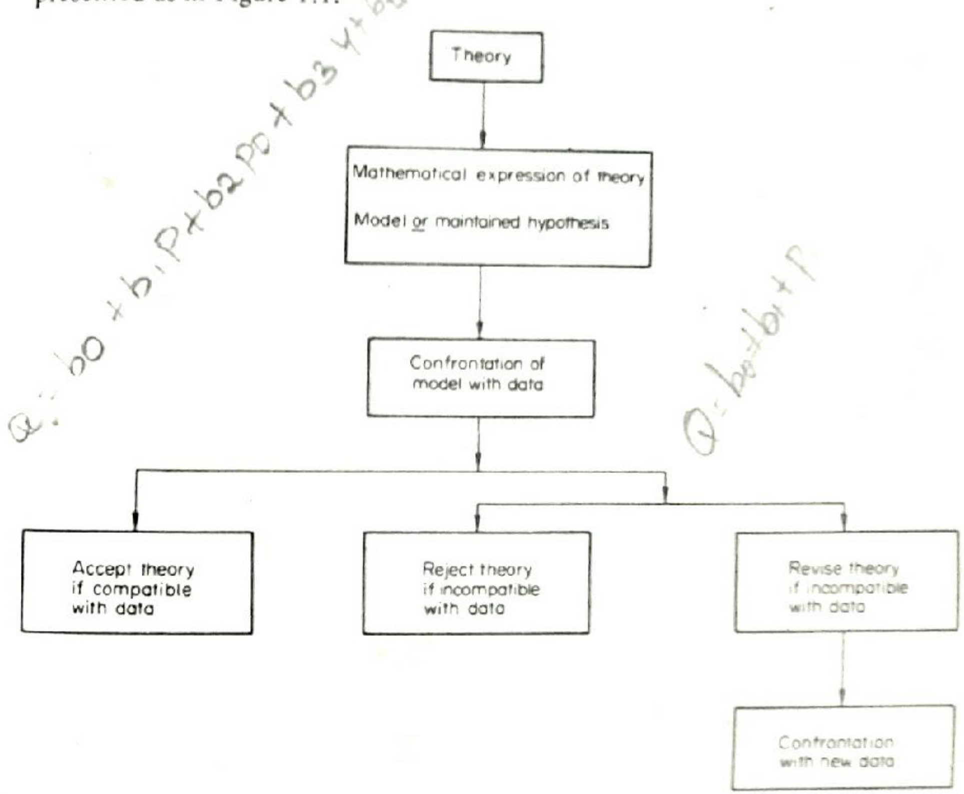



Figure 1.1

The procedure outlined above is not intended to imply that when testing a theory the researcher should restrict himself only to factors suggested by economic theory. If these factors do not provide a satisfactory explanation of economic behaviour, the research worker is certainly entitled to look for other factors. Experimentation with alternative formulations, each including various explanatory factors, has proved a most valuable guide to the revision and restatement of the hypotheses of economic theory. Econometrics, by establishing the usefulness or the insignificance of factors suggested by economic theories, has given new insight into various fields of economics and often provided evidence which has led to a reshaping of theoretical economics. One of the most striking examples in this respect is the investment function. (See M. K. Evans, *Macroeconomic Activity*, Harper & Row, 1969, Chapters 4–8.)

Various writers have argued that there is no need for a pre-existing body of theory: one may start with a set of observed data and from this derive a behavioural theory. This argument is known as 'measurement with no theory'. Such an approach seems absurd given that economics in its present state does provide a large number of hypotheses which may be tested empirically. A pre-existing body of theory saves a lot of time by showing which of the mass of · data available are of interest in any particular case. Furthermore, measurement alone may yield results which are not meaningful; for example it has been found that the number of storks and the number of babies born in New York show a strong statistical correlation, which clearly does not make sense. However, if the researcher chooses to adopt the 'measurement with no theory' approach, the following considerations should be borne in mind. An econometrician with clever experimentation can always arrive at some formulation which he may present as a 'theory'. However, in this case the researcher cannot claim that his 'theory' has been tested from the evidence of his original data. The information of these data has been used for the derivation of the 'theory' and cannot be used again for testing it. In other words, one should distinguish clearly between the test of already existing theory by using observational data, and the use of observations for formulating a new theory. Such new theory cannot be tested against the same data used for its derivation. One needs additional observations for its verification.

We said that econometrics is the integration of economic theory, mathematical economics and statistics. We examine below the relationship between econometrics, mathematical economics and statistics, pointing out the main differences between these branches of science.

### 1.1.1. ECONOMETRICS AND MATHEMATICAL ECONOMICS

Mathematical economics states economic theory in terms of mathematical symbols. There is no essential difference between mathematical economics and economic theory. Both state the same relationships, but while economic theory uses verbal exposition, mathematical economics employs mathematical symbolism. Both express the various economic relationships in an exact form. Neither economic theory nor mathematical economics allows for random elements which might affect the relationship and make it stochastic. Furthermore, they do not provide numerical values for the coefficients of the relationships.

Econometrics differs from mathematical economics. Although econometrics presupposes the expression of economic relationships in mathematical form,



like mathematical economics it does not assume that economic relationships are exact. On the contrary, econometrics assumes that relationships are not exact. Econometric methods are designed to take into account random disturbances which create deviations from the exact behavioural patterns suggested by economic theory and mathematical economics. Furthermore, econometric methods provide numerical values of the coefficients of economic phenomena. For example, economic theory suggests that the demand for a product which covers a basic human need is inelastic, provided the commodity does not have close substitutes. This information is of little assistance to policy-makers, because the coefficient of elasticity may assume any value between 0 and 1. Econometrics can supply precise estimates of elasticities and other parameters of economic theory.

### 1.1.2. ECONOMETRICS AND STATISTICS

Econometrics differs both from mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or charts them, and then attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. Economic statistics is mainly a descriptive aspect of economics. It does not provide explanations of the development of the various variables and it does not provide measurement of the parameters of economic relationships.

Mathematical (or inferential) statistics deals with methods of measurement, which are developed on the basis of controlled experiments in laboratories. Statistical methods of measurement are not appropriate for economic relationships, which cannot be measured on the basis of evidence provided by controlled experiments, because such experiments cannot be designed for economic phenomena. In physics and some other sciences the researcher can hold all other conditions constant and change only one element in performing an experiment. He can then record the results of such a change and apply the classical statistical methods to deduce the laws governing the phenomenon being investigated. In studying the economic behaviour of human beings one cannot change only one factor while keeping all other factors constant. In the real world all variables change continuously and simultaneously, so that controlled experiments are impossible. We cannot change only incomes, keeping prices, tastes and other factors constant, because the latter will change as a result of income changes.

Econometrics uses statistical methods after adapting them to the problems of economic life. These adapted statistical methods are called econometric methods. In particular, econometric methods are adjusted so that they become appropriate for the measurement of economic relationships which are stochastic, that is they include random elements. The adjustment consists primarily in specifying the stochastic (random) elements that are supposed to operate in the real world and enter into the determination of the observed data, so that the latter can be interpreted as a (random) sample to which the methods of statistics can be applied.

## 1.2. GOALS OF ECONOMETRICS

We can distinguish three main goals of econometrics: (1) analysis, i.e. testing of economic theory; (2) policy-making, i.e. supplying numerical estimates of the coefficients of economic relationships, which may be then used for decision-making; (3) forecasting, i.e. using the numerical estimates of the coefficients in order to forecast the future values of the economic magnitudes. Of course, these goals are not mutually exclusive. Successful econometric applications should really include some combination of all three aims.

### 1.2.1. ANALYSIS: TESTING ECONOMIC THEORY

In the earlier stages of the development of economic theory economists formulated the basic principles of the functioning of the economic system using verbal exposition and applying a deductive procedure. The earlier economic theories started from a set of observations concerning the behaviour of individuals as consumers or producers. Some basic assumptions were set regarding the motivation of individual economic units. Thus in demand theory it was assumed that the consumer aims at the maximisation of his satisfaction (utility) from the expenditure of his income, given the prices of the commodities. Similarly, producers were assumed to be motivated by maximisation of their profits. From these assumptions the economists by pure logical reasoning derived some general conclusions (laws) concerning the working processes of the economic system. Economic theories thus developed in an abstract level were not tested against economic reality. In other words no attempt was made to examine whether the theories explained adequately the actual economic behaviour of individuals.

Econometrics aims primarily at the verification of economic theories. In this case we say that the purpose of the research is *analysis*, i.e. obtaining empirical evidence to test the explanatory power of economic theories, to decide how well they *explain* the observed behaviour of the economic units. Today any theory, regardless of its elegance in exposition or its sound logical consistency, cannot be established and generally accepted without some empirical testing.

### 1.2.2. POLICY-MAKING: OBTAINING NUMERICAL ESTIMATES OF THE COEFFICIENTS OF ECONOMIC RELATIONSHIPS FOR POLICY SIMULATIONS

In many cases we apply the various econometric techniques in order to obtain reliable estimates of the individual coefficients of the economic relationships from which we may evaluate elasticities or other parameters of economic theory (multipliers, technical coefficients of production, marginal costs, marginal revenues, etc.). The knowledge of the numerical value of these coefficients is very important for the decisions of firms as well as for the formulation of the economic policy of the government. It helps to compare the effects of alternative policy decisions.

For example, the decision of the government about devaluing the currency will depend to a great extent on the numerical value of the marginal propensity to import, as well as on the numerical values of the price elasticities of exports



and imports. If the sum of price elasticities of exports and imports is less than one in absolute value, the devaluation will not help in eliminating the deficit in the balance of payments.

Similarly, if the price elasticity of demand for a product is less than one (inelastic demand), it does not pay the manufacturer to decrease its price, because his receipts would be reduced.

In a competitive market with linear demand and supply curves of the usual type (downward-sloping demand and upward-sloping supply), the government should not impose a specific excise tax (per unit of output) if its aim is to curb price increases, because such a tax would raise the price, although less than the amount of the tax per unit, *ceteris paribus*.

Such examples show how important is the knowledge of the numerical values of the coefficients of the economic relationships. Econometrics can provide such numerical estimates and has become an essential tool for the formulation of sound economic policies.

### 1.2.3. FORECASTING THE FUTURE VALUES OF ECONOMIC MAGNITUDES

In formulating policy decisions it is essential to be able to forecast the value of the economic magnitudes. Such forecasts will enable the policy-maker to judge whether it is necessary to take any measures in order to influence the relevant economic variables.

For example, suppose that the government wants to decide its employment policy. It is necessary to know what is the current situation of employment as well as what the level of employment will be, say, in five years' time, if no measure whatsoever is taken by the government. With econometric techniques we may obtain such an estimate of the level of employment. If this level is too low, the government will take appropriate measures to avoid its occurrence. If the forecast value of employment is higher than the expected labour force, the government must take different measures in order to avoid inflation.

Forecasting is becoming increasingly important both for the regulation of developed economies as well as for the planning of the economic development of underdeveloped countries.

## 1.3. DIVISION OF ECONOMETRICS

Econometrics may be distinguished into two branches, theoretical econometrics and applied econometrics.

*Theoretical econometrics* includes the development of appropriate methods for the measurement of economic relationships. As mentioned above, econometric techniques are basically statistical techniques which have been adapted to the particular characteristics of economic relationships. Two features of economic reality render the pure methods of mathematical statistics inappropriate for the measurement of economic phenomena. Firstly, the data which are used for the measurement of economic relationships are observations of actual life and are not derived from controlled experiments. In economic life laboratory experiments are not possible, because most of the economic magnitudes change contemporaneously and each influences and is influenced by all the other magnitudes. Accordingly, econometric methods have been developed for the analysis of non-experimental data. Secondly, the economic relationships are not exact, as economic theory and mathematical economics assume them to be. Economic behaviour is to a certain extent erratic, being influenced by unpredictable events. The effects of such factors are taken into account by econometricians through the introduction in the relationship being studied of a special random variable, whose nature will be examined in subsequent chapters.

Econometric methods may be classified into two groups: (1) single-equation techniques, which are methods that are applied to one relationship at a time; and (2) simultaneous-equation techniques, which are methods applied to all the relationships of a model simultaneously. In this book we shall develop various methods of measurement of economic phenomena.

*Applied econometrics* includes the applications of econometric methods to specific branches of economic theory. It examines the problems encountered and the findings of applied research in the fields of demand, supply, production, investment, consumption, and other sectors of economic theory. Applied econometrics involves the application of the tools of theoretical econometrics for the analysis of economic phenomena and forecasting economic behaviour.

# 2. Methodology of Econometric Research

Applied econometric research is concerned with the measurement of the parameters of economic relationships and with the prediction (by means of these parameters) of the values of economic variables.

The relationships of economic theory which can be measured with one or another econometric technique are causal, that is, they are relationships in which some variables are postulated as causes of the variation of other variables. In this sense definitional equations do not require any measurement. For example the equation $Y = C + I + G$ is the mathematical expression of the definition of national income of economic theory. It does not explain the determination of the level of income or the causes of its variations. We stress this point because in many instances researchers tend to 'measure' a relationship which actually is a simple definition and does not express any causal relationship among the variables involved.

In any econometric research we may distinguish four stages.

*Stage A.* The first step in any econometric research is the specification of the model with which one will attempt the measurement of the phenomenon being analysed. This stage is also known as the formulation of the *maintained hypothesis.*

*Stage B.* After the formulation of the model one should obtain estimates of its parameters, that is, the second stage includes the estimation of the model by means of the appropriate econometric method. This stage is known as the testing of the maintained hypothesis.

*Stage C.* Once the model has been estimated, one should proceed with the evaluation of the estimates, that is to say decide on the basis of certain criteria whether the estimates are satisfactory and reliable.

*Stage D.* The final stage of any econometric research is concerned with the evaluation of the forecasting validity of the model. Estimates are useful because they help in decision making. A model, after the estimation of its parameters, can be used in forecasting the values of economic variables. The econometrician must ascertain how good the forecasts are expected to be, in other words he must test the forecasting power of the model.

Stages A and C are the most important for any econometric research. They require the skills of an economist with experience of the functioning of the economic system. Stages B and D are technical and require knowledge of theoretical econometrics.

In this chapter we will discuss in some detail these four stages of econometric

## 2.1. STAGE A. SPECIFICATION OF THE MODEL

The first, and the most important, step the econometrician has to take in attempting the study of any relationship between variables, is to express this relationship in mathematical form, that is to specify the model, with which the economic phenomenon will be explored empirically. This is called the specification of the model or formulation of the *maintained hypothesis*. It involves the determination of: (1) the dependent and explanatory variables which will be included in the model; (2) the *a priori* theoretical expectations about the sign and the size of the parameters of the function. These *a priori* definitions will be the theoretical criteria on the basis of which the results of the estimation of the model will be evaluated; (3) the mathematical form of the model (number of equations, linear or non-linear form of these equations, etc.).

The specification of the econometric model will be based on economic theory and on any available information relating to the phenomenon being studied. Thus the specification of the model presupposes knowledge of economic theory as well as familiarity with the particular phenomenon being studied. The econometrician must know the general laws of economic theory, and furthermore he must gather any other information relevant to the particular characteristics of the relationship as well as all studies already published on the subject by other research workers.

### 2.1.1. VARIABLES OF THE MODEL

From the above sources of information the econometrician will be able to make a list of the variables (regressors) which might influence the dependent variable (regressand). Economic theory indicates the general factors which affect the dependent variable in any particular case. For example, suppose that the econometrician wants to study the demand for a particular product. The first source of his information is the static theory of demand which suggests that the determinants of the demand for any product are its price, the prices of other goods (mainly of substitutes and complements), the level of the income of consumers, and their preferences. On the basis of this information we may write the demand function in the general form

$$Q_z = f(P_z, P_0, Y, T)$$

where $Q_z$ = quantity demanded of commodity $z$
$P_z$ = price of commodity $z$
$P_0$ = price of other commodities
$Y$ = consumers' income
$T$ = a suitable measure of consumers' tastes.

Apart from general economic theory, studies already published in any particular field provide additional knowledge about the factors determining the dependent variable. Thus published results of econometric research on the demand for various products provide evidence that, apart from the above four factors suggested by economic theory, the demand is affected by other factors



such as the level of income earned in previous periods ($Y_{t-1}$, $Y_{t-2}$, etc.), the taxation and credit policy of the government ($G$), and the distribution of income ($Y_d$). Thus the demand function becomes

$$Q_z = f(P_z, P_0, Y, T, Y_{t-1}, Y_{t-2}, G, Y_d)$$

Finally the information about the individual conditions in a particular case, and the actual behaviour of the economic agents (consumers or producers) implements the knowledge of theory and of applied research. If we study the demand for exports of a product, in addition to the above factors we must take into account dumping policies, tariffs of country-buyers, foreign currency restrictions in these countries, etc.

It should be clear that the number of variables to be included in the model depends on the nature of the phenomenon being studied and the purpose of the research. Usually we introduce explicitly in the function only the most important (four or five) explanatory variables. The influence of less important factors is taken into account by the introduction in the model of a random variable, usually denoted by $u$. The values of this random variable cannot be actually observed like the values of the other explanatory variables. We thus have to guess at the pattern of the values of $u$ by making some plausible assumptions about their distribution. The statement of the assumptions about the random variable is part of the specification of the model (see Chapter 4).

### 2.1.2. SIGNS AND MAGNITUDE OF PARAMETERS

The same sources of knowledge – theory, other applied research and information about possible special features of the phenomenon being studied – will contain suggestions about the *sign* of the parameters and possibly of their *size*.

For example assume that we investigate the demand function for a given product

$$Q_z = b_0 + b_1 P_z + b_2 P_j + b_3 Y + u$$

We should expect, according to the general theory of demand, the following findings.

The parameter $b_1$ is expected to have a negative sign, given the 'law of demand' which postulates an inverse relationship between quantity demanded and price.

The parameter $b_3$ related to the variable $Y$ is expected to appear with a positive sign, since income and quantity demanded are positively related, except in the case of inferior goods.

The parameter $b_2$ of the variable $P_j$ is expected to have a positive sign if commodity $j$ is a substitute of commodity $z$, and a negative sign if the two commodities are complementary.

As regards the magnitude of the parameters we note the following. The $b$'s are either elasticities, propensities or other marginal magnitudes of economic theory, or are components of these parameters. In a linear demand function, such as the one in our example, the $b$'s are components of the relevant lasticities.[1] Now the theory of demand suggests that the size of the elasticities lepends mainly on the nature of the commodity and the existence of substitutes. If the product is a 'necessity', price and income elasticities are expected to be small, if it is a 'luxury' these elasticities will be high assuming that the commodity has no close substitutes. The cross elasticity of demand for commodity $z$ with respect to the price of commodity $j$, depends on how close a substitute or a complement commodity $j$ is with respect to commodity $z$. If $j$ is a very close substitute of commodity $z$ the cross elasticity of demand will be very high. Thus, given the units of measurement of the variables, the $b$'s are expected to assume values which would give rise to elasticities of the appropriate theoretical magnitude.

As another example let us examine the simple version of the consumption function which states that consumption ($C$) depends on the level of income ($Y$)

$$C = b_0 + b_1 Y + u$$

In this function the coefficient $b_1$ is the marginal propensity to consume and should be positive with a value less than unity ($0 < MPC < 1$), while the constant intercept ($b_0$) of the function is expected to be positive. The meaning of this positive constant is that even when income is zero, consumption will assume a positive value: people will spend past savings, will borrow or find other means for covering their needs.

To decide in any particular case whether a good is normal or inferior, a 'necessity' or a 'luxury' item, whether it has substitutes and how close these substitutes are, one should know the conditions of the market being studied. For example a television set is a 'necessity' in the United Kingdom, while it is a 'luxury' product in under-developed countries.

Determination of the variables to be included or excluded from a function may be viewed as imposition of zero and non-zero restrictions on the parameters of the variables of the model. That is, once we decide to exclude a variable from a function we actually impose the restriction that its parameter be zero in that function. Similarly if we decide to include a variable in the function this means that we impose the restriction that its parameter assumes a value different from zero. Of course the measurement of the relation may show that some of the included in the function variables are not significant, in which case we may modify our initial hypothesis by excluding these variables. Thus the number of variables to be initially included in the model depends on the nature of the economic phenomenon being studied, while the number of variables which will finally be retained in the model depends on whether the parameter estimates related to the variables pass the economic, statistical and econometric criteria, which we will discuss below.

### 2.1.3. MATHEMATICAL FORM OF THE MODEL

Economic theory may or may not indicate the precise mathematical form of the relationships, or the number of equations to be included in the economic

[1] See pp. 66–7.



model. For example, the theory of demand does not determine whether the demand for a particular commodity should be studied with a single-equation model or with a system of simultaneous equations. Furthermore economic theory does not say whether the demand function will be of a linear or a nonlinear form, demand curves are drawn as straight downward-sloping lines or as curves. However, demand theory contains some information about the mathematical form of a demand function. Static demand theory is based on the assumption that the behaviour of consumers is rational and that they do not suffer from money illusion. This assumption implies that if all prices and incomes change by the same proportion, the rational consumer will not change his consumption patterns, that is he will not change his demand for the various commodities. Thus the demand function should assume a mathematical form which will take into account the rationality assumption of demand theory. In technical jargon we say that the demand function is homogeneous of degree zero. (There are various ways for expressing the rationality assumption of the theory of demand. See L. R. Klein, *An Introduction to Econometrics*, Prentice-Hall International, London 1962, pp. 19–24.)

In most cases economic theory does not explicitly state the mathematical form of economic relationships. It is often helpful to plot the actual data on two-dimensional diagrams, taking two variables at a time (the dependent and each one of the explanatory variables in turn). In most cases the examination of such scatter diagrams throws some light on the form of the function and helps in deciding upon the choice of the mathematical form of the relationship connecting the economic variables. In view of the vagueness of economic theory in this respect it has become a usual practice for the econometrician to experiment with various forms (linear, nonlinear) and then choose from among the various results the ones that are judged as the most satisfactory on the basis of certain criteria which will be discussed below.

Nonlinearities are usually taken into account by a polynomial form, for example

$$Y = b_0 + b_1 X + b_2 X^2 + u$$

or

$$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + u$$

and so on. The number of nonlinear terms which will be retained in the function is decided upon tests of their significance (see Chapter 8).

We should finally note that economic theory does not explicitly state whether a particular phenomenon should be studied with a single equation model or with a multi equation model. It is the econometrician who must decide whether the phenomenon being studied can be adequately described by a single equation or by a system of simultaneous equations. If an economic relationship is complex and we attempt to approximate it by a single equation model, we are almost certainly bound to obtain incorrect estimates of the parameters. Taking into account the complexity of the real world one should hardly expect to study economic phenomena are studied by using single equation models. Yet an important part of applied econometric research is based on single equation models and it measures their coefficients by single equation techniques. This may not be the appropriate procedure, as we shall later see. We note here that the number of equations, that is the size of the model, depends on the complexity of the phenomenon being studied, the purpose for which the model is estimated (forecasting, or obtaining accurate individual values for particular coefficients), the availability of data and the computational facilities available to the research worker. In some cases the model is simplified by dropping some of its equations for lack of data, money or time.

As a final remark we note that the specification is the most important and the most difficult stage of any econometric research. It is often the weakest point of most econometric applications. Some of the reasons for incorrect specification of economic models are easy to see: (1) the imperfection, looseness of statements in economic theories; (2) the limitation on our knowledge of the factors which are operative in any particular case; (3) the formidable obstacles presented by data requirements in the estimation of large models. The most common *errors of specification* are the omission of some variables from the functions, the omission of some equations and the mistaken mathematical form of the functions. It should be noted that almost all the econometric methods are sensitive to errors of specification, that is the estimates of the coefficients obtained from most econometric methods will be incorrect or unreliable if the model is not correctly specified. (See Theil, *Economic Forecasts and Policy*, North Holland, Amsterdam 1965, pp. 204–40. See also Chapter 11.)

## 2.2. STAGE B. ESTIMATION OF THE MODEL

After the model has been specified (formulated) the econometrician must proceed with its estimation, in other words he must obtain numerical estimates of the coefficients of the model.

The estimation of the model is a purely technical stage which requires knowledge of the various econometric methods, their assumptions and the economic implications for the estimates of the parameters.

The stage of estimation includes the following steps.

(1) Gathering of statistical observations (data) on the variables included in the model.

(2) Examination of the identification conditions of the function in which we are interested.

(3) Examination of the aggregation problems involved in the variables of the function.

(4) Examination of the degree of correlation between the explanatory variables, that is, examination of the degree of multicollinearity.

(5) Choice of the appropriate econometric technique for the estimation of the function and critical examination of the assumptions of the chosen technique and of their economic implications for the estimates of the coefficients.



We will attempt to give some idea of the problems involved in each of the above steps, but their full understanding will be possible only after reading the whole book.

### 2.2.1. GATHERING DATA FOR THE ESTIMATION OF THE MODEL

The data used in the estimation of a model may be of various types.

*Time series*

Time series data give information about the numerical values of variables from period to period. For example the data on gross national income in the period 1950–65 form a time series on the variable 'income'.

*Cross-section data*

These data give information on the variables concerning individual agents (consumers or producers) at a given point of time. For example a cross-section sample of consumers is a sample of family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic, social or financial characteristics. Cross-section data may also refer to aggregate variables of different countries (or other regional entities) referring to the same time. Such data are usually called cross-nation (or cross-country) samples and are used for international comparative studies.

*Panel data*

These are repeated surveys of a single (cross-section) sample in different periods of time. They record the behaviour of the same set of individual microeconomic units over time.

*Engineering data*

These data give information about the technical requirements of the method of production (productive processes) employed (by a firm or an industry, or the economy as a whole) for producing a certain commodity. These are collected from the producers of the commodity and are used in studies of production (production functions, input–output relationships, etc.). For example, we can obtain information from the steel firms about the engineering characteristics of their method of steel production and the volume of their output. This information will enable us to find the proportions in which the several methods are employed, and thus we can make a close approximation to the relationship between steel output and input requirements.

*Legislation and other institutional regulations*

Some models can be estimated from direct information about the nature of the relationship involved. This is particularly true for institutional functions, like tax functions. For example, in most countries the taxation of cigarette consumption is determined by law. Taking into account the various tax coefficients for the various brands of tobacco products as well as the volume of consumption of each tobacco brand, it is possible to estimate the tax burden on tobacco. Suppose that this information shows that tobacco is taxed, on the average, at 65 per cent of its retail value. The tax revenue function from tobacco would be related to expenditure on tobacco by the function

$$T = 0{\cdot}65\ C$$

where $T$ = government revenue from tobacco consumption
$C$ = expenditure on tobacco manufactures.

This is a function 'estimated' by reference to the information of the tax legislation: it is an 'institutional' function.

*Data constructed by the econometrician: Dummy variables*

In many cases some factors affecting the dependent variable cannot be measured in any of the above conventional data, because they are qualitative factors. For example profession, religion, sex, are factors affecting the consumption of particular items like bread, meat, cosmetics. Such qualitative attributes can be approximated by the introduction in the function of 'dummy variables', that is indexes which we construct with considerable arbitrariness, but in a way relevant to the influence of the factor concerned. For example if we study the demand for bread with cross-section data, the factor 'sex' could be represented by a dummy variable, which might be assigned the value of one when the individual is a male and the value of zero when the consumer is a female. In this case the coefficient of the dummy variable will be positive if in the real world females consume less bread. As another example suppose we want to estimate the demand for petrol from a cross-section sample. The main determinant of the demand in this case will be the ownership of a car. We may approximate the factor 'car ownership' with a dummy variable which would take the value of zero, if an individual consumer does not own a car and the value of unity if the consumer does own a car.

It should be noted that various problems arise from the use of the one or the other type of data for the estimation of a given econometric model. For example, the meaning of the estimates of the coefficients is different according to whether we use time series or cross-section data. Furthermore, in some cases there is need for pooling together various types of data for the estimation of a model. Such problems will be discussed in detail in subsequent chapters.

2.2.2. EXAMINATION OF THE IDENTIFICATION CONDITION OF THE FUNCTION

Identification is the procedure by which we attempt to establish that the coefficients which we shall estimate by the application of some appropriate econometric technique are actually the true coefficients of the function in which we are interested. This may sound strange at this stage, but there are cases in which we may obtain estimates for which we cannot be certain as to which function they belong: they may either belong to the function on which our interest is focused, or they may belong to some other function which



happens to have the same statistical form (that is it has the same variables as the one which we are studying), or they may be some mixture of coefficients belonging to various functions. For example, suppose we want to estimate the demand function for a product for a period over which incomes and other factors except price have remained constant. Thus both the demand and the supply will depend on the price of the commodity

$$Q_d = f(P) \quad \text{and} \quad Q_s = f(P)$$

Assume that we wish to estimate the demand function by using time series of market data. Such data record the quantity demanded at a certain price, but the quantity bought is at the same time the quantity sold ($D = S$) at the market price $P$. Thus when using the recorded market data on $Q$ and $P$ we do not know whether we are estimating the parameters of the demand function or of the supply function. There are some rules by means of which we may establish identification of the coefficients of a function. These rules are analysed in Chapter 15. We note here that the job of identification is most important since it determines whether a relationship, although theoretically plausible, can be statistically estimated or not.

2.2.3. EXAMINATION OF THE AGGREGATION PROBLEMS OF THE FUNCTION

Aggregation problems arise from the fact that we use aggregative variables in our functions. Such aggregative variables may involve:

(a) *Aggregation over individuals.* For example, total income is the sum of individual incomes; total output is the sum of the output of individual firms, and so on.

(b) *Aggregation over commodities.* We may aggregate over the quantities of various commodities (using appropriate quantity indexes), or over the prices of a group of commodities (using some appropriate price index). For example, if we want to estimate the demand function for 'food', with explanatory variables 'total income', 'the price of food', and 'the price of other commodities', all variables will include a certain level of aggregation.

(c) *Aggregation over time periods.* In many cases statistical sources publish data which refer to a time period different (longer or shorter) than the unit time period required in theory for the functional relationship among the economic variables. For example, the production of most manufacturing commodities is completed in a period shorter than a year. If we use annual figures there may be some error in the coefficients of the production function.

(d) *Spatial aggregation.* For example the population of towns, counties, regions, or, product of regions, of the whole country, of the world as a whole, and so on.

The above sources of aggregation create various complications which may impart some 'aggregation bias' in the estimates of the coefficients. It is important to examine the possibility of such sources of error before estimating the function, and to adjust the aggregative variables or the model accordingly whenever possible. (See R. G. D. Allen, *Mathematical Economics*, Macmillan, London, 1956, chapter 20. Also L. R. Klein, *An Introduction to Econometrics*, Prentice-Hall International, London 1962, pp. 64–6, 86–7, 104–5.)

### 2.2.4. EXAMINATION OF THE DEGREE OF CORRELATION AMONG THE EXPLANATORY VARIABLES

Most economic variables are correlated, in the sense that they tend to change
simultaneously during the various phases of economic activity. Income, employ-
ment, consumption, investment, exports, imports, taxes, tend to grow in booms
and decline in periods of depression. Thus a certain degree of multicollinearity is
inherent in the economic variables due to the growth and technological progress.
If, however, the degree of collinearity is high, the results (measurements)
obtained from econometric applications may be seriously impaired and their use
may be greatly misleading, because in these conditions it may not be computa-
tionally possible to separate the influence of each one explanatory variable. For
example, prices and wages tend to increase together. If we include both these
variables in the set of explanatory variables in a demand function, it is most
probable that the estimated values of the coefficients will be inaccurate and will
show a distorted influence of each individual explanatory variable on demand.
The problem of multicollinearity is discussed in Chapter 11.

### 2.2.5. CHOICE OF THE APPROPRIATE ECONOMETRIC TECHNIQUE

The coefficients of economic relationships may be estimated by various
methods which may be classified in two main groups:

(i) *Single-equation techniques*. These are techniques which are applied to one
equation at a time. The most important are: the Classical Least Squares or
Ordinary Least Squares method, the Indirect Least Squares or Reduced-form
technique, the Two-stage Least Squares method, the Limited Information
Maximum Likelihood method and various methods of Mixed Estimation.

(ii) *Simultaneous-equation techniques*. These are techniques which are
applied to all the equations of a system at once, and give estimates of the
coefficients of all the functions simultaneously. The most important are the
Three-stage Least Squares method and the Full Information Maximum
Likelihood technique.

Which technique will be chosen in any particular case depends on many
factors, such as: (a) The nature of the relationship and its identification
condition. If we study a simple phenomenon which can be satisfactorily
approximated with a single-equation model the method of ordinary least
squares will usually be chosen for its considerable advantages (see Chapter 4).
If, however, the particular function in which we are interested belongs to a
system of simultaneous equations we may use any one of the above techniques,
depending primarily on the identification condition of the function. If the
function is identified, as we shall see in Ch[illegible] choice among
various of the abo[illegible]



sideration some of the following factors. (b) The properties of the estimates of
the coefficients obtained from each technique. In Chapter 6 we shall see that a
'good' estimate should possess the properties of unbiasedness, consistency,
efficiency, sufficiency or a combination of such properties. If one method gives
an estimate which possesses more of these desirable characteristics than any other
estimate from other methods, then the former technique is preferred to the
others. We shall return to this point again. (c) However, which of these desirable
characteristics is the most important, depends on the purpose of the econometric
research. It is usually argued that if the purpose of the model is forecasting, the
property of minimum variance is very important, bias being less important in pre-
dicting the values of economic variables, but if the purpose of the research worker
is analysis or policy-making, in which case he is interested in obtaining good
estimates of individual coefficients, the degree of bias becomes crucial. (d) In
some cases the simplicity of the method is used as a criterion of choice: a
method may be preferred to another because the first involves simpler computa-
tions and has less data requirements than the other. (e) Finally, the time and
cost requirements of the various methods are often important criteria for the
choice of the technique for the estimation of the parameters of a model.

From the above discussion we conclude that the estimation of a model can
be managed with several econometric methods, but in most cases only one
would be, theoretically, the most appropriate for the problem being studied.
However, the theoretically most appropriate econometric technique may not
be applicable due to non-availability or to defects (e.g. multicollinearity) of the
relevant statistical data and other information. Thus it becomes necessary to
choose another less suitable technique, given the data limitations. In most
empirical research data-limitations restrict seriously the possibilities of employ-
ing the theoretically most suitable econometric technique and render inevitable
the use of a less appropriate method. In this case one should interpret the results
of the estimation taking into account the effects and possible errors introduced
into the estimates by the use of the less appropriate technique.

For example the demand function for most goods should be estimated with a
complete model which would take into account the whole working mechanism
of the market of this product. There should be included in this model a
demand equation, a supply equation, a price equation as well as other relevant
equations (tax functions) because it is common knowledge that in all markets
the quantities demanded, the quantities supplied, the price and the taxation
policy are interdependent, each one of these factors influencing and at the
same time being influenced by the others. However, for simplicity, econometri-
cians tend to use single-equation demand models, sacrificing to a certain extent
the accuracy of the estimates in order to facilitate the estimation. Taking into
account the interdependence of quantity and price, however, it is obvious that
the estimates will include some error, which should be taken into account when
interpreting the results of the calaculations.

After choosing the econometric technique for the estimation of his model, the
econometrician should state explicitly the assumptions of this technique and

examine their implications for the estimates of the parameters. Strictly speaking the assumptions relate (a) to the form of the distribution of the random variable $u$ and (b) to the relationships among the explanatory variables. They are assumptions concerning the variables of the model and not the particular method which is applied for the estimation of the model. However, they are usually stated as assumptions of the particular technique. In any case the explicit statement of these assumptions is a very important task; if these assumptions are violated, either the estimates of the parameters will be biased, or it will not be possible to assess their reliability, or both. On the basis of the assumptions of each method the econometrician determines the econometric criteria, which will be used for the evaluation of the results of the computations (see section 2.3 below).

### 2.2.5. 'EXPERIMENTAL APPROACH' VERSUS 'ORTHODOX APPROACH'

In applying econometric methods for the estimation of economic models two approaches have been developed, the 'orthodox approach' and the 'experimental approach'.

The 'orthodox' econometric approach consists in formulating a mathematical model on *a priori* theoretical grounds, and attempting to measure the parameters of that model on the basis of the best available data. Data deficiencies might lead to minor modifications of the model before it could be tested statistically, but broadly speaking, having established his model the 'orthodox' econometrician would tend to stick to it, despite unfavourable statistical results. In other words, following the orthodox approach of econometric research one would proceed as follows:

(1) Collect all information, from theory or from practice, relevant to the phenomenon being studied.

(2) Decide on *a priori* reasoning on the particular mathematical expression of the model.

(3) Estimate the model with the available statistical data.

The model constructed on *a priori* assumptions is considered by the orthodox econometrician as the only true model, irrespective of the results obtained. If these results are 'unfavourable', that is the signs and size of the parameters do not conform to *a priori* knowledge, the econometrician will not reject the model but would try to explain the results by attributing them to data deficiencies mainly. The initial model is considered as 'correct' and would not be revised.

It is obvious that such a rigid approach to applied econometric research is not commendable. First of all in order to stick to an initial formulation of the model, one should be certain that he commands perfect knowledge of all the aspects of the phenomenon being analysed. Such a pretention would be outrageous, given the complexity of economic phenomena and the loose exposition of economic theory. Furthermore, one may pretend to have followed the orthodox approach, while in reality one has experimented to a considerable extent, before settling for the model, which one may present afterwards as being compiled by the most orthodox econometric



Today most econometric research is attempted by the experimental approach. Experimentation with various models has been facilitated by the expansion of the use of electronic computers. In following the experimental approach one starts with simple models containing a small number of equations and variables. These models are formulated on *a priori* considerations, like the models of the orthodox approach, but they are not considered as being rigid. On the contrary, they are modified gradually, on the basis of the statistical evidence accruing from the computations. The econometrician starts from a simple model, which on *a priori* grounds is believed to contain the most important factors of the relationship being analysed. Then additional variables are added, and perhaps the formulation is given a more complex appearance (non-linear forms, etc.). In other words the econometrician experiments with various theoretically plausible models including various variables and/or various mathematical formulations.

The experimental approach combines the theoretical considerations (*a priori* criteria) with the empirical observations available and is designed to extract the maximum of information from the available data. As calculations are carried out by adding other explanatory variables in various combinations, or by adding other equations, or by changing the mathematical form of the functions, or by using alternative econometric methods for the estimation of the models, the econometrician is able to observe the effects of such changes in an attempt to achieve the best model, the best explanation of the phenomenon being analysed. Each time a new variable (or any other change) is introduced because it is thought to improve the explanation of the phenomenon, three statistical effects on the model will normally result.

(1) The new variable (or change) will have some effect, minor or major, on the systematic part of the relation. In other words, the new variable will or will not be shown to explain a significant part of the variation in the dependent variable.

(2) It will affect the non-systematic (residual) part of the relationship, for example because of errors of measurement in this new variable.

(3) It will have some minor or major effect upon the coefficients of the variables already included in the equation (model). We should notice that if an important variable is omitted, not only will the overall fit of the relation be worse, but the coefficients of the included variables may well be distorted from the values which would be obtained from a complete analysis. In this case the introduction of the new variable will 'correct' the value of the coefficients of the other explanatory variables.[1]

It is obvious from the above discussion that the experimental approach to econometric analysis has more advantages in comparison to the orthodox approach. In particular it renders possible a better use of the available data and information. The experimentation may involve models with (a) various variables,

[1] See R. Stone, 'The Analysis of Market Demand', *Journal of the Royal Statistical Society*, Great Britain 1945, vol. CVIII. See also Chapter 11.8.

(b) various mathematical forms, (c) various numbers of equations, (d) various econometric methods. The process of choosing between the various models involves both the *a priori* and economic-theoretical considerations of the 'orthodox' econometrician, and also a sifting of the statistical evidence given by the experimental approach.

We should note that both the alternative lines of approach have a certain degree of arbitrariness: the orthodox approach makes *a priori* assumptions, while the second makes *a posteriori* choice. What matters is that the investigator should give a full description of his method of research, so that one can judge how much reliability can be attached to the results obtained.[1]

Some authors have criticised the experimental approach on the grounds that (a) the degree of subjective judgement it involves is higher than in the orthodox approach, and (b) the use of the same sample of data for the estimation of various models implies a loss of degrees of freedom which is overlooked in most cases. The meaning of 'degrees of freedom' is discussed briefly in Appendix I.

We agree that the experimental approach is not *the* perfect approach. There is a considerable realism in the argument that if an econometrician is clever and persistent he can always find an equation that fits the data satisfactorily. What is worse, he may argue that his equation is theoretically plausible, i.e. he may attempt to revise economic theory on the basis of his results, a procedure which may not always be justifiable.

The argument of loss of degrees of freedom is often referred to as '*the problem of data mining*'. (See M. Friedman, in 'Conference on Business Cycles', Universities NBER (New York, 1951), pp. 107–14. Also C. F. Christ, *Econometric Models and Methods*, Wiley 1966, New York, pp. 8–9.) This argument is based on purely statistical considerations and runs as follows. The reliability of the estimates is judged on the basis of statistical tests of significance (discussed in Chapter 5), which assume that the maintained hypothesis (the model which we test against the data) is known with certainty. In the experimental approach the maintained hypothesis is not known with certainty, but is chosen because it gives the best fit to the available sample data. This decision implies that in the hypothetical repeating sampling procedure on which the classical tests of significance are based, we use not all possible samples, but only those samples that fit the data well: in this way we introduce a non-random factor in the process for selecting samples, which restricts our freedom of choice. This loss of degrees of freedom should be taken into account in order to adjust the test procedure, otherwise the tests will not be valid. In most cases, however, the appropriately adjusted statistical test is not known. Thus researchers tend mostly to ignore the problem completely. Some writers have suggested a new method of research which incorporates actual numerical *a priori* knowledge in the model, a fact that reduces the need for experimentation to a great extent.

[1] See A. Koutsoyiannis, *An Econometric Study of the Leaf Tobacco Market of Greece*, 1962, Papadimitropoulos Press, pp. 8–9.



This method is known as 'mixed estimation' and will be discussed in Chapter 17. It is the author's belief that the 'data mining' problem is not important for econometrics. Statistical considerations may become highly restrictive for the purpose of econometrics. Some 'loose' interpretation of statistical rules is at times essential if econometrics is to be helpful in testing economic theory and in measuring economic relationships.

## 2.3. STAGE C. EVALUATION OF ESTIMATES

After the estimation of the model the econometrician must proceed with the evaluation of the results of the calculations, that is with the determination of the reliability of these results. The evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically satisfactory. For this purpose we use various criteria which may be classified into three groups. Firstly, economic *a priori* criteria, which are determined by economic theory. Secondly, statistical criteria, determined by statistical theory. Thirdly, econometric criteria, determined by econometric theory.

### 2.3.1. ECONOMIC '*A PRIORI*' CRITERIA

These are determined by the principles of economic theory and refer to the sign and the size of the parameters of economic relationships.

As we have already mentioned, the coefficients of economic models are the 'constants' of economic theory: elasticities, marginal values, multipliers, propensities, etc. Economic theory defines the signs of these coefficients and in broad lines their magnitude. In econometric jargon we say that economic theory imposes restrictions on the signs and values of the parameters of economic relationships.

For example, let us examine the liquidity preference function of an economy. The Keynesian theory of liquidity preference postulates that the main determinants of the demand for money are the level of income ($Y$) and the rate of interest ($i$). This theory suggests that there is a positive relationship between the demand for money ($M$) and the level of income: the larger the income, the larger the amount of money held in the form of cash balances, because the larger the income, the larger the amount required to carry out the transactions. On the contrary, there is a negative relationship between the demand for money and the rate of interest: the higher the rate of interest, the lower the amount of money demanded (to hold in idle balances), because (a) the loss from not lending the money is high, and (b) because a high $i$ implies a low price of bonds and other securities, a fact that makes the purchase of such securities attractive in the expectation of reselling them at a higher price later and thus having capital gains. The liquidity preference function may be expressed in the mathematical form

$$M = b_0 + b_1 Y + b_2 i + u$$

On the basis of the above theory the *a priori* criteria to be used for the evaluation of the estimates of the liquidity preference function may be stated as follows

The sign of $b_1$ is expected to be positive while the sign of $b_2$ is expected to be negative. As regards the magnitude of these parameters not much information is provided by the theory of liquidity preference. However, knowledge of the habits of firms and individuals of an economy may help in setting *a priori* limits to the sizes of $b_1$ and $b_2$.

If the estimates of the parameters turn up with signs or size not conforming to economic theory, they should be rejected, unless there is good reason to believe that in the particular instance the principles of economic theory do not hold. In such cases the reasons for accepting the estimates with the 'wrong' sign or magnitude must be stated clearly. However, in most cases the wrong sign or size of the parameters may be attributed to deficiencies of the empirical data employed for the estimation of the model.[1] In other words either the observations are not representative of the relationship, or their number is inadequate, or some assumptions of the method employed are violated. In general, if the *a priori* theoretical criteria are not satisfied, the estimate should be considered unsatisfactory

### 2.3.2. STATISTICAL CRITERIA: FIRST-ORDER TESTS

These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. The most widely used statistical criteria are the *correlation coefficient* and *the standard deviation* (or standard error) *of the estimates.* These criteria will be explained in subsequent chapters, but a few comments are appropriate here.

The estimates of the parameters of the model are obtained from a sample of observations of the variables included in the relationship. The sampling theory of statistics prescribes some tests for finding out how accurate these estimates are.

The square of the correlation coefficient is a statistical number, computed from the data of the sample, which shows the percentage of the total variation of the dependent variable being explained by the changes of the explanatory variables. It is a measure of the extent to which the explanatory variables are responsible for the changes in the dependent variable of the relationship (see Chapter 5).

The standard deviation or standard error of the estimates is a measure of the dispersion of the estimates around the true parameter. The larger the standard error of a parameter, the less reliable it is, and vice versa (see Chapter 5 and Appendix I).

It should be noted that the statistical criteria are secondary only to the *a priori* theoretical criteria. The estimates of the parameters should be rejected in general if they happen to have the 'wrong' sign (or size) even though the correlation coefficient is high, or the standard errors suggest that the estimates are statistically significant. In such cases the parameters, though statistically satisfactory, are theoretically implausible, that is to say they make no sense on the basis of the *a priori* theoretical-economic criteria.

[1] See, for example, J. Johnston, *Statistical C*[illegible] Hill, 1962, for a discussion of the data [illegible]



The importance of the statistical criteria in evaluating the results of the estimates of the coefficients is further discussed in Chapter 5.

### 2.3.3. ECONOMETRIC CRITERIA: SECOND-ORDER TESTS

These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as second-order tests (as tests of the statistical tests); in other words they determine the reliability of the statistical criteria, and in particular of the standard errors of the parameter estimates. They help us establish whether the estimates have the desirable properties of unbiasedness, consistency, etc. (see Chapter 6).

If the assumptions of the econometric method applied by the investigator are not satisfied, either the estimates of the parameters cease to possess some of their desirable properties (for example become biased) or the statistical criteria lose their validity and become unreliable for the determination of the significance of these estimates.

We said that the econometric criteria aim at the detection of the violation or validity of the assumptions of the econometric method employed in any particular application. The assumptions of the various econometric techniques differ and hence there are various econometric criteria for each method. These will be discussed in connection with the various techniques. Some examples may illustrate the meaning of the econometric criteria.

All econometric techniques listed in page 20 have the common assumption that the values of the random variable included in the model are not connected one to the other. This is known as the assumption of non-autocorrelated random disturbances (see Chapters 4 and 10). If this assumption is violated the standard errors of the parameters are not a reliable criterion for the evaluation of the statistical significance of the coefficients. To test the validity of the assumption of non-autocorrelated disturbances, we may compute a statistic, known as the 'Durbin–Watson *d* statistic', from the names of the inventors (see Chapter 10). The '*d*' statistic is an econometric criterion used in the evaluation of the results of the estimates.

Another example is the 'test' aiming at establishing the identification conditions of a relationship. All econometric methods assume that the function to which they are applied is identified, since otherwise the estimation of the coefficients is meaningless. The application of the formal rules of identification, which will be developed in Chapter 15, consists of an econometric test, aiming at the detection of the fulfilment of one of the basic assumptions of all econometric techniques.

From the above discussion it should be clear that the evaluation of the results obtained from the estimation of the model, is a very complex procedure. The econometrician must use all the above criteria, economic, statistical and econometric, before he can accept or reject the estimates.

When the assumptions of an econometric technique are not satisfied it is customary to respecify the model (e.g. introduce new variables or omit some meets the assumptions of the econometric theory. We then proceed ... form which estimation of the new model and with re-application of all the tests. This process of re-specification of the model and re-estimation will continue until the results pass all the economic, statistical and econometric tests. (See E. Kane, *Economic Statistics and Econometrics*, Harper & Row, International edition, 1969, pp. 352–3.)

## 2.4. STAGE D. EVALUATION OF THE FORECASTING POWER OF THE ESTIMATED MODEL

We have said that the objective of any econometric research is to obtain good numerical estimates of the coefficients of economic relationships and to use them for the prediction of the values of economic variables. Forecasting is one of the prime aims of econometric research.

Before using an estimated model for forecasting the value of the dependent variable we must assess by some way or another the predictive power of the model. It is conceivably possible that the model is economically meaningful and statistically and econometrically correct for the sample period for which the model has been estimated, yet it may very well not be suitable for forecasting due, for example, to rapid change in the structural parameters of the relationship in the real world.

The final stage of any applied econometric research is the investigation of the stability of the estimates, their sensitivity to changes in the size of the sample. We must establish whether the estimated function performs adequately outside the sample of data, whose 'average' variation it represents. Extra-sample performance is an important and independent test of the results obtained by applying an econometric technique. It is a test independent of the statistical and econometric tests applied in the previous stage.

One way of establishing the forecasting power of a model is to use the estimates of the model for a period not included in the sample. The estimated value (forecast value) is compared with the actual (realised) magnitude of the relevant dependent variable. Usually there will be a difference between the actual and the forecast value of the variable, which is tested with the aim of establishing whether it is (statistically) significant. If after conducting the relevant test of significance, we find that the difference between the realised value of the dependent variable and that estimated from the model is statistically significant, we conclude that the forecasting power of the model, its extra-sample performance, is poor.

Another way of establishing the stability of the estimates and the performance of the model outside the sample of data from which it has been estimated, is to re-estimate the function with an expanded sample, that is a sample including additional observations. The original estimates ... new estimates. The diff...





There may be various reasons for a model's poor forecasting performance. (a) The values of the explanatory variables used in the forecast may not be accurate. (b) The estimates of the coefficients ($\hat{b}$'s) may be poor, due to deficiencies of the sample data. (c) The estimates are 'good' for the period of the sample, but the structural background conditions of the model may have changed from the period that was used as the basis for the estimation of the model, and therefore the old estimates are not 'good' for forecasting. In this event the whole model needs re-estimation before it can be used for prediction.

We shall discuss the problems of the forecasting performance of estimated models in Chapter 20, but for the moment we give a simplified example of the forecasting procedure. Suppose that we estimate the demand function for a given commodity with a single equation model using time series data for the period 1950–68, as follows

$$\hat{Q}_t = 100 + 5\,Y_t - 30\,P_t$$

This equation is then used for 'forecasting' the demand of the commodity in the year 1970, a period outside the sample data.

Given $\quad Y_{1970} = £1000 \quad$ and $\quad P_{1970} = 5$ shillings

$$\hat{Q}_t = 100 + 5\,(1000) - 30\,(5) = 4{,}950 \text{ tons}$$

If the actual demand for this commodity in 1970 is 4,500 tons, there is a difference of 450 tons between the estimated from the model and the actual market demand for the product. This difference can be tested for significance by various methods (see Chapters 8 and 20). If it is found significant, we try to find out what are the sources of the error in the forecast, in order to improve the forecasting power of our model.

## 2.5. DESIRABLE PROPERTIES OF AN ECONOMETRIC MODEL

An econometric model is a model whose parameters have been estimated with some appropriate econometric technique.

The 'goodness' of an econometric model is judged customarily according to the following desirable properties.

(1) *Theoretical plausibility.* The model should be compatible with the postulates of economic theory. It must describe adequately the economic phenomena to which it relates.

(2) *Explanatory ability.* The model should be able to explain the observations of the actual world. It must be consistent with the observed behaviour of the economic variables whose relationship it determines.

(3) *Accuracy of the estimates of the parameters.* The estimates of the coefficients should be accurate in the sense that they should approximate as best as possible the true parameters of the structural model. The estimates should if possible possess the desirable properties of unbiasedness, ...

(6) Forecasting ability. The model should produce satisfactory predictions of future values of the dependent (endogenous) variables.

(7) Simplicity. The model should represent the economic relationships with maximum simplicity. The fewer the equations and the simpler their mathematical form, the better the model is considered, ceteris paribus (that is to say provided that the other desirable properties are not affected by the simplifications of the model).

The more of the above properties a model possesses the better it is considered for any practical purpose. (See L. [illegible] Econometric Models and [illegible] [illegible] [illegible] [illegible] [illegible] Econometric Forecasts and Policy. [illegible] [illegible] [illegible] [illegible] [illegible].)

EXERCISES

Exercises relating to the material of this chapter are included in Appendix [illegible].

## 3.1. GENERAL NOTES

There are various methods for measuring the relationships existing between economic variables. The simplest are correlation analysis and regression analysis. We shall start from correlation analysis, because, although it has serious limitations and throws little light on the nature of the relationship existing between variables, it will make the student familiar with the correlation coefficient, which is an essential statistic of regression analysis.

Correlation may be defined as the degree of relationship existing between two or more variables. The degree of relationship existing between two variables is called simple correlation. The degree of relationship connecting three or more variables is called multiple correlation. In this chapter we shall examine only simple correlation, postponing the discussion on multiple correlation until a later chapter, after the examination of regression analysis. (Actually the multiple correlation coefficient cannot be interpreted without reference to the multiple regression analysis.)

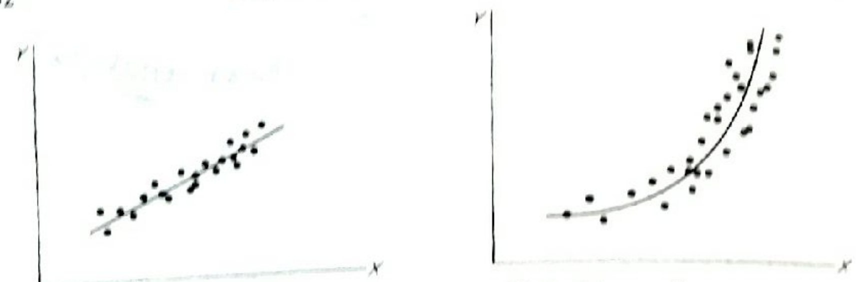
Correlation may be *linear*, when all points $(X, Y)$ on a scatter diagram seem to cluster near a straight line, or *nonlinear*, when all points seem to lie near a curve.

Two variables may have a positive correlation, a negative correlation, or they may be uncorrelated. This holds both for linear and nonlinear correlation.

*Positive correlation.* Two variables are said to be positively correlated if they tend to change together in the same direction, that is, if they tend to increase or decrease together. Such positive correlation is postulated by economic theory for the quantity of a commodity supplied and its price. When the price increases the quantity supplied increases, and conversely, when price falls the quantity supplied decreases. The scatter diagram of two variables positively correlated appears in figure 3.1. All points in the scatter diagram seem to lie near a line or a curve with a positive slope. If all points lie *on* the line (or curve) the correlation is said to be *perfect positive.*

*Negative correlation.* Two variables are said to be negatively correlated if they tend to change in the opposite direction: when $X$ increases $Y$ decreases, and vice versa. For example, the quantity of a commodity demanded and its price are negatively correlated. When price increases, demand for the commodity decreases and when price falls demand increases. The scatter diagram appears in figure 3.2; the points cluster around a line (or curve) with a negative slope. If all points lie on the line (or curve) the correlation is said to be *perfect negative.*

*No correlation, or, zero correlation.* Two variables are uncorrelated when



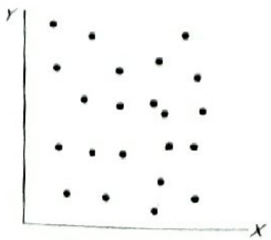
(a) Positive linear correlation (b) Positive nonlinear correlation

Figure 3.1

(a) Negative linear correlation (b) Negative nonlinear correlation

Figure 3.2

they tend to change with no connection to each other. The scatter diagram will appear as in figure 3.3. The points are dispersed all over the surface of the $XY$ plane. For example one should expect zero correlation between the height of the inhabitants of a country and the production of steel, or between the weight of students and the colour of their hair.



Zero correlation

Figure 3.3

## 3.2. MEASURE OF LINEAR CORRELATION: THE POPULATION CORRELATION COEFFICIENT $\rho$, AND ITS SAMPLE ESTIMATE $r$

In the light of the above discussion it appears that we can determine the kind of correlation between two variables by direct observation of the scatter diagram. In addition, the scatter diagram indicates the ***strength*** of the relationship between the two variables. If the points lie close to the line, the correlation



is strong. On the other hand a greater dispersion of points about the line implies weaker correlation. Yet inspection of a scatter diagram gives only a rough idea of the relationship between variables $X$ and $Y$. For a precise quantitive measurement of the degree of correlation between $Y$ and $X$ we use a parameter which is called *the correlation coefficient* and is usually designated by the Greek letter $\rho$, having as subscripts the variables whose correlation it measures. $\rho$ refers to the correlation of all the values of the population of $X$ and $Y$. Its estimate from any particular sample (the sample statistic for correlation) is denoted by $r$ with the relevant subscripts. For example if we measure the correlation between $X$ and $Y$ the population correlation coefficient is represented by $\rho_{XY}$ and its sample estimate by $r_{XY}$. We will establish that the sample correlation coefficient is defined by the formula

$$r_{XY} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{3.1}$$

where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. (Throughout this book lower case letters will denote deviations from the mean of the variables and capital letters the observed values, unless otherwise stated.)

We will use a simple example from the theory of supply. Economic theory suggests that the quantity of a commodity supplied in the market depends on its price, *ceteris paribus*. When price increases the quantity supplied increases, and *vice versa*, when the market price falls producers offer smaller quantities of their commodity for sale. In other words economic theory postulates that price ($X$) and quantity supplied ($Y$) are positively correlated.

Our problem is to define a measure with which we will determine the correlation between price $X$ and quantity supplied $Y$. Our first task is to gather observations of prices and quantities supplied during a given time period. A set of hypothetical observations appears in table 3.1.

Table 3.1

| *Time period (in days)* | *Quantity supplied $Y_i$ (in tons)* | *Price $X_i$ (in shillings)* |
|---|---|---|
| 1 | $Y_1 = 10$ | $X_1 = 2$ |
| 2 | $Y_2 = 20$ | $X_2 = 4$ |
| 3 | $Y_3 = 50$ | $X_3 = 6$ |
| 4 | $Y_4 = 40$ | $X_4 = 8$ |
| 5 | $Y_5 = 50$ | $X_5 = 10$ |
| 6 | $Y_6 = 60$ | $X_6 = 12$ |
| 7 | $Y_7 = 80$ | $X_7 = 14$ |
| 8 | $Y_8 = 90$ | $X_8 = 16$ |
| 9 | $Y_9 = 90$ | $X_9 = 18$ |
| 10 | $Y_{10} = 120$ | $X_{10} = 20$ |
| $n = 10$ | $\sum_1^n Y_i = 610$ | $\sum_1^n X_i = 110$ |

By plotting the above observations on a rectangular co-ordinate system, we get the scatter diagram of figure 3.4.[1]

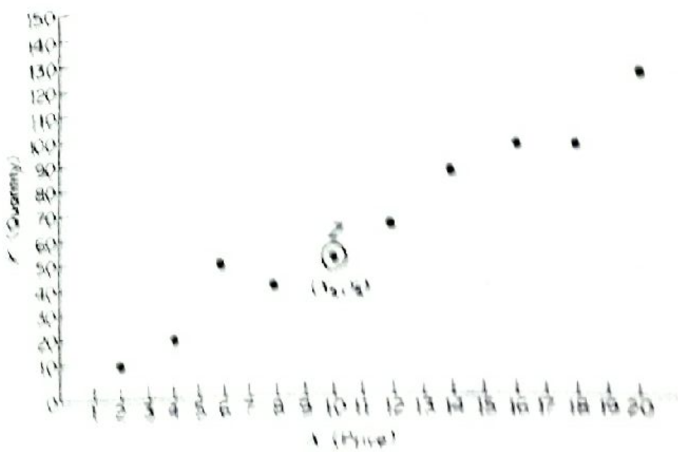



Figure 3.4

Each point of the scatter diagram represents a pair of price – quantity in a given period. For example, point $\mathscr{E}$ represents the pair $(X_8, Y_8)$, that is the price, which is 10 shillings and the quantity supplied, which is 50 tons, during the 8th period. Looking at the diagram we see that the points tend to cluster around a line with a positive slope. This suggests that there exists a positive linear correlation between price and quantity. In order to find the exact measure of correlation we work as follows.

(1) We compute the mean value of the variables

$$\bar{X} = \frac{\sum X_i}{n} = \frac{110}{10} = 11 \quad \text{and} \quad \bar{Y} = \frac{\sum Y_i}{n} = \frac{610}{10} = 61$$

(2) We draw the perpendiculars $\bar{X}\bar{X}'$ and $\bar{Y}\bar{Y}'$ from the means, $\bar{X}$ and $\bar{Y}$, thus dividing the area of the rectangular co-ordinate system into four quadrants I, II, III and IV (figure 3.5).

(3) We next take the deviation of each value of $X$ and $Y$ from their mean and denote this difference by lower case letters

$$x_i = (X_i - \bar{X}) \quad \text{and} \quad y_i = (Y_i - \bar{Y})$$

Examining the deviations of the values of the variables $X$ and $Y$ from their means, we observe that their products can provide a measure of the correlation between the variables $X$ and $Y$.

[1] It is necessary to start correlation analysis by plotting the sample observations on a scatter diagram, in order to see whether the relationship is linear or nonlinear. Because if the relationship between $X$ and $Y$ exists, but is nonlinear, the formulae which will be developed in the present chapter break down (see below, page 43).



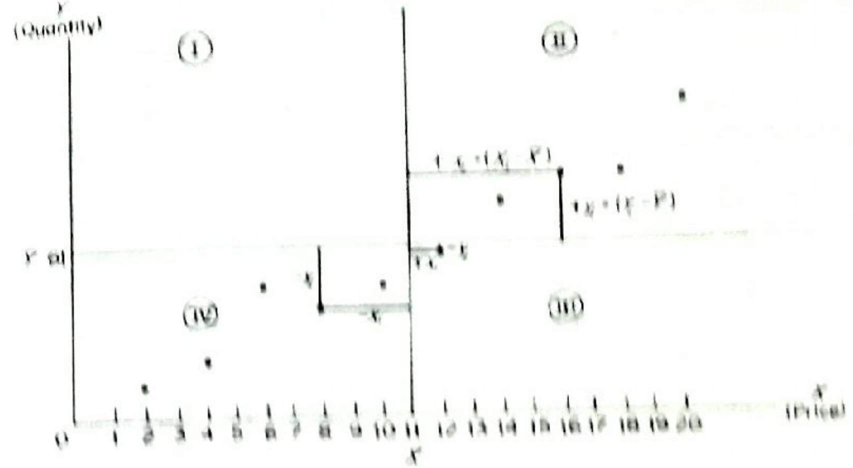



Figure 3.5

**(a) In quadrants II and IV the product $(X_i - \bar{X})(Y_i - \bar{Y}) = x_i y_i$ is positive, because both deviations $x_i$ and $y_i$ have the same sign, both being either positive or negative.**

**(b) In quadrants I and III the product $(X_i - \bar{X})(Y_i - \bar{Y}) = x_i y_i$ is negative, because the deviations of the $x_i$'s have the opposite sign of the deviations of the $y_i$'s lying in the same quadrants.**

Thus if most observations fall in quadrants II and IV, the correlation between $Y$ and $X$ is positive. If on the other hand most of the quantity – price pairs fall in quadrants I and III, the correlation between $Y$ and $X$ will be negative. If the observations are scattered at random all over the four quadrants, the positive and negative products $x_i y_i$ will tend to cancel each other out, and the sum of products will tend to approach zero. If the sum of all products of the deviations of variables $X$ and $Y$ from their means is positive, the correlation between $X$ and $Y$ will be positive, while if the sum of the products of deviations is negative, the correlation between $X$ and $Y$ will be negative. Symbolically

if $$\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_1^n x_i y_i > 0$$

the correlation between $X$ and $Y$ is positive,

if $$\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_1^n x_i y_i < 0$$

the correlation between $X$ and $Y$ is negative.

Thus the sum of the products of the deviations $\sum x_i y_i$ provides a measure of the association between $X$ and $Y$. However, this measure has two basic defects. Firstly, it is affected by the number of observations. The greater the number of observations, the greater the number of products will be, and therefore the value of the sum $\sum x_i y_i$ will be different. Thus if $X$ and $Y$ are positively related, an increase in the number of observations would make the correlation appear stronger, without this being necessarily true. Secondly, the sum $\Sigma x_i y_i$ is affected by the units of measurement of the variables $X$ and $Y$. For example, the correlation in the above case would appear higher if the supply was measured in kilograms and the price in pence although the observations would be exactly the same.

To correct the first defect we divide the sum $\Sigma x_i y_i$ by the number of observations $n$

$$\frac{\Sigma x_i y_i}{n} = S_{XY}$$

This expression is the *covariance* of $X$ and $Y$ and is obviously a better measure of correlation than the simple sum $\Sigma x_i y_i$, because it will not change directly with the number of observations in the sample. However, it still has the defect of being influenced by the units in which the variables $X$ and $Y$ are measured. To correct this defect we divide the covariance by the standard deviations of the variables, which are measured in the same units as the variables themselves, so that the ratio becomes a pure number, independent of any change in the units of measurement of $X$ and $Y$. The resulting ratio is the sample *correlation coefficient* $r$, which is an estimate of the population correlation coefficient $\rho$.

$$r = \frac{\sum_{i}^{n} x_i y_i}{n S_X S_Y} = \frac{S_{YX}}{S_X S_Y} \qquad (3.2)$$

where $S_{XY}$ = covariance of $X$ and $Y = (\Sigma x_i y_i)/n$

$$S_X = \text{standard deviation of } X,\ S_X = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}} = \sqrt{\frac{\Sigma x_i^2}{n}}$$

$$S_Y = \text{standard deviation of } Y,\ S_Y = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{n}} = \sqrt{\frac{\Sigma y_i^2}{n}}$$

Substituting the values of $S_{XY}$, $S_X$ and $S_Y$ in expression 3.2 we find

$$r = \frac{\Sigma x_i y_i}{n\sqrt{(\Sigma x_i^2/n)(\Sigma y_i^2/n)}} = \frac{\Sigma x_i y_i}{\sqrt{(\Sigma x_i^2)(\Sigma y_i^2)}} \qquad (3.3)$$

This formula is expressed in deviations of the variables from their means. If we want to use the actual values of the observations we use the following form:

$$r = \frac{n\Sigma(X_i Y_i) - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2}\ \sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}} \qquad (3.4)$$

This formula is derived from (3.3) through the following transformations:

Given,

$$r = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}}$$



(1) The numerator can be expanded as follows

$$\Sigma xy = \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma(XY - X\bar{Y} - Y\bar{X} + \bar{X}\bar{Y}) = \Sigma XY - \bar{X}\Sigma Y - \bar{Y}\Sigma X + n\bar{X}\bar{Y}$$

$$= \Sigma XY - \frac{\Sigma X}{n}\Sigma Y - \frac{\Sigma Y}{n}\Sigma X + n\frac{\Sigma X}{n}\frac{\Sigma Y}{n}$$

$$= \frac{n\Sigma XY - \Sigma X\Sigma Y}{n} \qquad (3.5)$$

(2) From the denominator of expression (3.3) we get

$$\Sigma x^2 = \Sigma(X - \bar{X})^2 = \Sigma(X^2 - 2X\bar{X} + \bar{X}^2) = \Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}^2$$

$$= \Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}\frac{\Sigma X}{n}$$

$$= \Sigma X^2 - \bar{X}\Sigma X$$

$$= \Sigma X^2 - \frac{(\Sigma X)^2}{n} \qquad (3.6)$$

$$= \frac{n\Sigma X^2 - (\Sigma X)^2}{n} \qquad (3.7)$$

Similarly

$$\Sigma y^2 = \frac{n\Sigma Y^2 - (\Sigma Y)^2}{n} \qquad (3.8)$$

(3) Substituting (3.5), (3.6) and (3.8) in (3.3) we get

$$r = \frac{\{n\Sigma XY - (\Sigma X)(\Sigma Y)\}/n}{\sqrt{\{n\Sigma X^2 - (\Sigma X)^2\}/n}\ \sqrt{\{n\Sigma Y^2 - (\Sigma Y)^2\}/n}} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

The above formulae of the correlation coefficient have two points of interest

(1) The formulae are symmetric with respect to $X$ and $Y$, that is $r_{YX} = r_{XY}$.

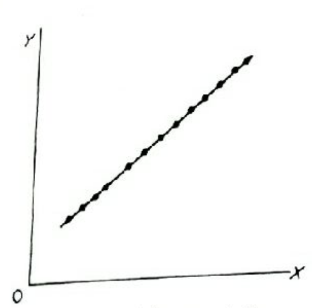
(2) The formulae are applicable only for linear relationships. Expressions for nonlinear relationships will be developed in Chapter 7 in connection with regression analysis.

## 3.3. NUMERICAL VALUES OF THE CORRELATION COEFFICIENT

The correlation coefficient is a measure of the degree of covariability of the variables $X$ and $Y$. The values that the correlation coefficient may assume vary from $-1$ to $+1$. When $r$ is positive, $X$ and $Y$ increase or decrease together. $r = +1$ implies that there is perfect positive correlation between $X$ and $Y$. Diagrammatically, all the observations on $Y$ and $X$ lie on a straight line with a positive slope (figure 3.6).

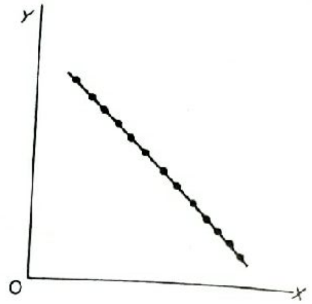
When $r$ is negative, $X$ and $Y$ move in opposite directions. If $r = -1$, there exists a perfect negative correlation between $X$ and $Y$. Diagrammatically, all the observations of $Y$ and $X$ lie on a line with a negative slope (figure 3.7).

When $r$ is zero, then the two variables are uncorrelated.




Perfect positive correlation

Figure 3.6




Perfect negative correlation

Figure 3.7

We shall prove that the $r$ will assume the value of unit when the two variables are perfectly linearly correlated. In this case all the observations will lie on a line with a positive or negative slope according to whether the correlation is positive or negative.

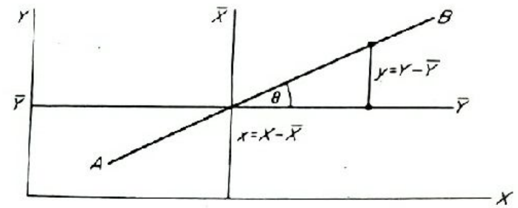



Figure 3.8

In figure 3.8 we picture the case of perfect positive correlation between $X$ and $Y$. The line depicting the relation forms an angle $\theta$ with the parallel $\overline{YY}$ to the horizontal axis. From elementary trigonometry it is known that $\tan\theta = y/x$. Therefore$\{(x)\,.\,(\tan\theta)\} = (y)$. Substituting this result in the formula of the correlation coefficient we find

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2\,\Sigma y^2}} = \frac{\Sigma x(x)\,.\,(\tan\theta)}{\sqrt{\Sigma x^2\,\Sigma\{(x)\,.\,(\tan\theta)\}^2}}$$

$$r = \frac{(\tan\theta)\,.\,(\Sigma x^2)}{\sqrt{(\Sigma x^2)\,.\,(\tan\theta)^2\,.\,(\Sigma x^2)}} = \frac{(\tan\theta)\,(\Sigma x^2)}{\sqrt{(\tan\theta)^2\,(\Sigma x^2)^2}} = 1$$

In practice, we almost never observe either perfect correlation or zero correlation. Usually $r$ assumes some value between zero and one. The closer that value is to one, the greater is the degree of covariability, that is the closer will the scatter of points approach a straight line. On the other hand, the greater the scatter of points in the diagram, the closer $r$ is to zero.

We said that $r$ is the sample estimate of the population correlation coefficient $\rho$. As a statistical estimate $r$ is inevitably subject to some error and should be tested for its reliability. Tests of significance for $r$ are explained in Chapters 5 and 8.

*Example.* Suppose we want to compute the correlation coefficient between the variables $Y$ (quantity supplied) and $X$ (price) with the observations included in table 3.1.

Table 3.2 Data for the estimation of the sample correlation coefficient $r_{YX}$

| $n$ | $Y_i$ | $X_i$ | $x_i = X_i - \bar{X}$ | $y_i = Y_i - \bar{Y}$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ | $X_i Y_i$ | $X_i^2$ | $Y_i^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 2 | −9 | −51 | 81 | 2,601 | 459 | 20 | 4 | 100 |
| 2 | 20 | 4 | −7 | −41 | 49 | 1,681 | 287 | 80 | 16 | 400 |
| 3 | 50 | 6 | −5 | −11 | 25 | 121 | 55 | 300 | 36 | 2,500 |
| 4 | 40 | 8 | −3 | −21 | 9 | 441 | 63 | 320 | 64 | 1,600 |
| 5 | 50 | 10 | −1 | −11 | 1 | 121 | 11 | 500 | 100 | 2,500 |
| 6 | 60 | 12 | +1 | −1 | 1 | 1 | −1 | 720 | 144 | 3,600 |
| 7 | 80 | 14 | +3 | +19 | 9 | 361 | 57 | 1,120 | 196 | 6,400 |
| 8 | 90 | 16 | +5 | +29 | 25 | 841 | 145 | 1,440 | 256 | 8,100 |
| 9 | 90 | 18 | +7 | +29 | 49 | 841 | 203 | 1,620 | 324 | 8,100 |
| 10 | 120 | 20 | +9 | +59 | 81 | 3,481 | 531 | 2,400 | 400 | 14,400 |
| $n = 10$ | $\Sigma Y_i = 610$ | $\Sigma X_i = 110$ | $\Sigma x_i = 0$ | $\Sigma y_i = 0$ | $\Sigma x_i^2 = 330$ | $\Sigma y_i^2 = 10,490$ | $\Sigma x_i y_i = 1,810$ | $\Sigma X_i Y_i = 8,520$ | $\Sigma X_i^2 = 1,540$ | $\Sigma Y_i^2 = 47,700$ |

*Computation of r using deviations from the means*

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

We need compute the terms $\sum x_i y_i$, $\sum x_i^2$, $\sum y_i^2$ which appear in the formula. The computations are given in table 3.2.

We see that $\sum xy = 1{,}810$, $\sum x^2 = 330$, $\sum y^2 = 10{,}490$. Substituting

$$r = \frac{1{,}810}{\sqrt{330}\ \sqrt{10{,}490}} = 0{\cdot}975$$

*Computation of r using actual observations*

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\ \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

From the formula we see that we need compute the terms

$$\sum XY \quad \sum X^2 \quad \sum Y^2 \quad \sum X \quad \sum Y \quad (\sum X)^2 \quad (\sum Y)^2$$

The computations are shown in table 3.2. Substituting we find

$$r = \frac{(10)\,.\,(8{,}520) - (610)\,.\,(110)}{\sqrt{(10)\,.\,(1540) - 12{,}100}\ \sqrt{(10)\,.\,(47{,}700) - 372{,}100}} = 0.975$$

## 3.4. THE RANK CORRELATION COEFFICIENT

The formulae of the linear correlation coefficient developed in the previous section are based on the assumption that the variables involved are quantitative and that we have accurate data for their measurement. However, in many cases the variables may be qualitative (or binary variables) and hence cannot be measured numerically. For example, profession, education, preferences for particular brands, are such categorical variables. Furthermore, in many cases precise values of the variables may not be available, so that it is impossible to calculate the value of the correlation coefficient with the formulae developed in the preceding section. For such cases it is possible to use another statistic, the *rank correlation coefficient* (or Spearman's correlation coefficient). We rank the observations in a specific sequence, for example in order of size, importance, etc., using the numbers 1, 2, . . . , n. In other words we assign *ranks* to the data values. Hence the relationship between their ranks instead of their actual numerical values. Hence the name of the statistic as rank correlation coefficient. If two variables $X$ and $Y$ are ranked in such way the rank correlation coefficient may be computed by the formula

$$r' = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad (3.9)$$

where $D$ = difference between ranks of corresponding pairs of $X$ and $Y$
$n$ = number of observations.



The values that $r'$ may assume range from +1 to −1. (For a test of significance of $r'$ see Chapter 6.)

Two points are of interest when applying the rank correlation coefficient. Firstly, it does not matter whether we rank the observations in ascending or descending order. However, we must use the same rule of ranking for both variables. Second, if two (or more) observations have the same value we assign to them the *mean rank*. Some examples will illustrate the application of the rank correlation coefficient.

*Example 1.* The following table shows how ten students were ranked according to their performance in their class work and their final examinations. We want to find out whether there is a relationship between the accomplishments of the students during the whole year and their performance in their exams.

| Students → | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranking based on class work | 2 | 5 | 6 | 1 | 4 | 10 | 7 | 9 | 3 | 8 |
| Ranking based on exam marks | 1 | 6 | 4 | 2 | 3 | 7 | 8 | 10 | 5 | 9 |

The differences between the two rankings is given in the following table.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $D$ | 1 | −1 | 2 | −1 | 1 | 3 | −1 | −1 | −2 | −1 | |
| $D^2$ | 1 | 1 | 4 | 1 | 1 | 9 | 1 | 1 | 4 | 1 | $\sum D^2 = 24$ |

The rank correlation coefficient is

$$r' = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(24)}{10(10^2-1)} = 0{\cdot}855$$

The high value of the rank correlation coefficient indicates that there is a close relationship between class work and exam performance: students with good record all over the year do well in their examinations and *vice versa*.

*Example 2.* A market researcher asks two smokers to express their preference for twelve different brands of cigarettes. Their replies are shown in the following table.

| Brands of cigarettes | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smoker Z | 9 | 10 | 4 | 1 | 8 | 11 | 3 | 2 | 5 | 7 | 12 | 6 |
| Smoker W | 7 | 8 | 3 | 1 | 10 | 12 | 2 | 6 | 5 | 4 | 11 | 9 |

The differences of preferences of the two smokers are shown below

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D$ | 2 | 2 | 1 | 0 | −2 | −1 | 1 | −4 | 0 | 3 | 1 | −3 | |
| $D^2$ | 4 | 4 | 1 | 0 | 4 | 1 | 1 | 16 | 0 | 9 | 1 | 9 | $\sum D^2 = 50$ |

The rank correlation coefficient

$$r' = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(50)}{12(12^2-1)} = 0{\cdot}827$$