4. The Simple Linear Regression Model
The Ordinary Least Squares Method

There are various econometric methods that can be used to derive estimates of the parameters of economic relationships from statistical observations. In this chapter we shall examine the method of ordinary least squares (OLS) or classical least squares (CLS). The reasons for starting with this method are many. Firstly, the parameter estimates obtained by ordinary least squares have some optimal properties which will be discussed in Chapter 6. Secondly, the computational procedure of OLS is fairly simple as compared with other econometric techniques and the data requirements are not excessive. Thirdly, the least squares method has been used in a wide range of economic relationships with fairly satisfactory results (see Chapter 21), and, despite the improvement of computational equipment and of statistical information which facilitated the use of other more elaborate econometric techniques, OLS is still one of the most commonly employed methods in estimating relationships in econometric models. Fourthly, the mechanics of least squares are simple to understand. Fifthly, OLS is an essential component of most other econometric techniques. In fact, as we will see later, with the exception of the Full Information Maximum Likelihood method, all other techniques involve the application of the least squares method, modified in some respects.

We shall start by the simple linear regression model, that is, by a relationship between two variables, one dependent and one explanatory, related with a linear function. Subsequently we will examine the multiple regression analysis, which refers to the relationship between more than two variables..

4.1. THE SIMPLE LINEAR REGRESSION MODEL

An example.

We will illustrate the meaning of the method of least squares by referring to our earlier example from the theory of supply. The theory of supply in its simplest form postulates that there exists a positive relationship between the quantity supplied of a commodity and its price, *ceteris paribus*. When the price rises the quantity of the commodity supplied increases and *vice versa*. Following the econometric procedure outlined in Chapter 2, our first task is the specification of the supply model, that is, the determination of the dependent (regressand) and the explanatory variables (regressors), the number of equations of the model and their precise mathematical form, and finally the *a priori* expectations concerning the sign and the magnitude of the coefficients. Economic theory provides the following information with respect to the supply function.

(1) The dependent variable is the quantity supplied and the explanatory variable is the price

$$Y = f(X)$$

where Y = quantity supplied
X = price of the commodity.

(2) Economic theory does not specify whether the supply should be studied with a single-equation model or with a more elaborate system of simultaneous equations. In view of this indeterminacy we choose to start our investigation with a single-equation model. In later stages we may study more elaborate models.

(3) Economic theory is not clear about the mathematical form (linear or nonlinear) of the supply function. In textbooks the supply is sometimes depicted by a straight upward-sloping line, or by an upward-sloping curve. The latter implies a nonlinear relationship between quantity and price. Again the econometrician has to decide the form of the supply function. We start by assuming that the variables are related with the simplest possible mathematical form, that is, the relationship between quantity and price is linear of the form

$$Y_i = b_0 + b_1 X_i$$

This form implies that there is a one-way causation between the variables Y and X: price is the cause of changes in the quantity supplied, but not the other way around.

The parameters of the supply function are b_0 and b_1, and our aim is to obtain estimates of their numerical values, $\hat{b}_0$ and $\hat{b}_1$.

As regards the sign and size of the constant intercept $\hat{b}_0$, we note that it should be either zero (in which case its meaning is that the quantity is zero when price is zero) or positive (in which case its meaning is that some quantity is supplied even when the price drops to zero). Normally $\hat{b}_0$ should not be negative in the case of a supply function. If $\hat{b}_0$ turns up with a negative sign we should ignore the negative part of the supply function, since a negative quantity does not make sense in economics. However, the sign of $\hat{b}_0$ is crucial in determining the price elasticity of supply, as we will presently see.

Regarding the value of $\hat{b}_1$, we note that in the particular case of a supply function we expect the sign of $\hat{b}_1$ to be positive ($\hat{b}_1 > 0$), since a supply curve is normally upward-sloping.

It is important to examine the relationship between the price elasticity of supply and the coefficients $\hat{b}_0$ and $\hat{b}_1$. Recall that the elasticity is defined by the expression

$$\eta_P = \frac{dQ}{dP} \cdot \frac{P}{Q} = \frac{dY}{dX} \cdot \frac{X}{Y}$$

From the supply function it is obvious that

$$\frac{dY}{dX} = b_1$$

In computing the elasticity from a regression line, we use the estimate $\hat{b}_1$ and the mean values of price ($\bar{X}$) and quantity ($\bar{Y}$) in the sample. Thus

$$\hat{\eta}_P = \hat{b}_1 \cdot \frac{\bar{X}}{\bar{Y}}$$

But, as we will show on page 63,

$$\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}$$

Thus, substituting for $\bar{Y}$ in the expression of the elasticity, we obtain

$$\hat{\eta}_P = \frac{\hat{b}_1 \bar{X}}{\hat{b}_1 \bar{X} + \hat{b}_0}$$

Given that $\hat{b}_1 > 0$, it follows that

(i) the supply will be elastic ($\eta_P > 1$) if $\hat{b}_0$ is negative ($\hat{b}_0 < 0$)
(ii) the supply will be inelastic ($\eta_P < 1$) if $\hat{b}_0$ is positive ($\hat{b}_0 > 0$)
(iii) the supply with have unitary elasticity if $\hat{b} = 0$.

Thus the elasticity of a supply curve (with positive slope) depends on the sign of the constant intercept, $\hat{b}_0$.

(4) The above form of the supply function implies that the relationship between quantity and price is *exact*, that is that all the variation in Y is due solely to changes in X, and that there are no other factors affecting the dependent variable. If this were true all the points of price–quantity pairs, if plotted on a two-dimensional plane, would fall on a straight line. However, if we gather observations on the quantity actually supplied in the market at various prices and we plot them on a diagram we see that they do not fall on a straight line (or any other smooth curve for that matter). Suppose that we have the ten pairs of observations on X and Y shown in table 4.1. The scatter

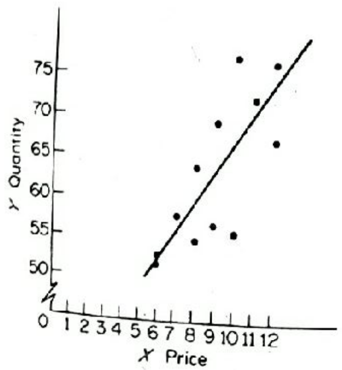


Figure 4.1

diagram of these observations shows that the relationship between price and quantity supplied has a form roughly similar to a straight line (figure 4.1).

The deviations of the observations from the line may be attributed to several factors.

(1) Omission of variables from the function

In economic reality each variable is influenced by a very large number of factors. For instance, the consumption pattern of a family is determined by family income, prices, the composition by age and sex of the family, the past levels of the family income, tastes, religion, social and educational status, wealth, and so on. One could compile an almost non-ending list of such factors. However, not all the factors influencing a certain variable can be included in the function for various reasons. (a) Some of the factors may not be known even to the person most aquainted with the relationship being studied. This lack of knowledge is to a great extent due to incomplete theory about the variation of economic variables in general. (b) Even when known to be relevant, some factors cannot be measured statistically. These are mainly psychological factors, or, in general, qualitative factors (tastes, expectations, religion) which cannot even be approximated satisfactorily with dummy variables. (c) Some factors are random, appearing in an unpredictable way and time, so that their influence cannot be taken satisfactorily into account (e.g. epidemics, earthquakes, wars). (d) Some factors may have, each individually, a very small influence on the dependent variable. Thus their parameter is so small that it cannot be measured in a reliable way (due to rounding errors of the computations). All these factors together, however, may account for a considerable part of the variation of the dependent variable. (e) Even if all factors are known, the available data most often are not adequate for the measurement of all factors influencing a relationship. This is particularly so when we use time series, which are usually short. Thus in most cases only the most important three or four variables are explicitly included in the function. The lack of

Table 4.1

Number of observations	*Y Quantity*	*X Price*
1	69	9
2	76	12
3	52	6
4	56	10
5	57	9
6	77	10
7	58	7
8	55	8
9	67	12
10	53	6
11	72	11
12	64	8

adequate number of observations creates a problem of 'degrees of freedom', which impairs the application of the traditional tests of significance. (See Chapter 5 and Appendix I.)

(2) *Random behaviour of the human beings.* The scatter of points around the line may be attributed to an erratic element which is inherent in human behaviour. Human reactions are to a certain extent unpredictable and may cause deviations from the 'normal' behavioural pattern depicted by the line. For example in a moment's whim a consumer may change his expenditure pattern, although income and prices did not change.

(3) *Imperfect specification of the mathematical form of the model.* We may have linearised a possibly nonlinear relationship. Or we may have left out of the model some equations. The economic phenomena are much more complex than a single equation may reveal, no matter how many explanatory variables it contains. In most cases many variables are simultaneously determined by a system containing many equations. For example price determines and is determined by the quantity supplied. Under such circumstances if we attempt to study the phenomenon with a single-equation model, we are bound to commit an error, which is due to the imperfect specification of the form of the model, that is, of the number of its equations.

(4) *Errors of aggregation.* We often use aggregate data (aggregate consumption, aggregate income), in which we add magnitudes referring to individuals whose behaviour is dissimilar. In this case we say that variables expressing individual peculiarities are missing. For example, in a production function for an industry we add together the factor inputs and outputs of dissimilar entrepreneurs. Changes in the distribution of total output among firms are important in the determination of total output. However, such distributional variables are often missing from the function. There are other types of aggregation which introduce error in the relationship. For example, aggregation over time, spatial aggregation, cross section aggregation, and so on.

(5) *Errors of measurement.* The deviations of the points from the line may be due to errors of measurement of the variables, which are inevitable due to the methods of collecting and processing statistical information.

The first four sources of error render the form of the equation wrong, and they are usually referred to as *error in the equation* or *error of omission*. The fifth source of error is called *error of measurement* or *error of observation*. It is usual of course to have both these types of error simultaneously in the function.

In order to take into account the above sources of error we introduce in econometric functions a random variable which is usually denoted by the letter u and is called *error term* or *random disturbance term* or *stochastic term* of the function, so called because u is supposed to 'disturb' the exact linear relationship which is assumed to exist between X and Y. By introducing this random variable in the function the model is rendered stochastic of the form

$$Y_i = (b_0 + b_1 X_i) + (u_i)$$

The true relationship which connects the variables involved is split into two

parts: a part represented by a line and a part represented by the random term u. The meaning of these two parts may be explained by looking at figure 4.2. The

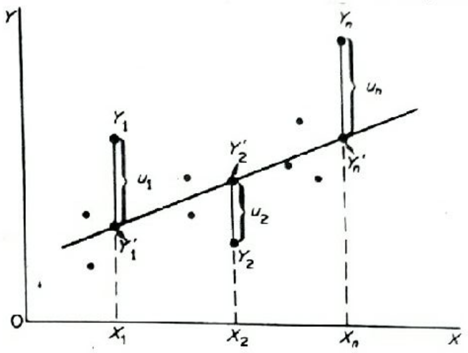


Figure 4.2

scatter of observations represents the true relationship between Y and X. The line represents the exact part of the relationship and the deviations of the observations from the line represent the random component of the relationship. Were it not for the errors in the model, we would observe the points on the line $Y_1', Y_2', \ldots, Y_n'$, corresponding to $X_1, X_2, \ldots, X_n$. However, because of the random disturbances, we observe $Y_1, Y_2, \ldots, Y_n$, corresponding to $X_1, X_2, \ldots, X_n$. These points diverge from the regression line by quantities $u_1, u_2, \ldots, u_n$, where u_i is the random error associated with Y_i. In other words the values of Y corresponding to a value of X will on the average fall on a line, but each individual Y_i will deviate from the line depending on the value of u_i. Hence each $Y_i (i = 1, 2, \ldots, n)$ can be expressed in terms of two components, one component due to X_i and a second component due to the influences included in the random term u_i

$$Y_i \quad = \quad \underbrace{b_0 + b_1 X_i}_{} \quad + \quad \underbrace{u_i}_{}$$

$$\begin{bmatrix}\text{Variation}\\ \text{in } Y_i\end{bmatrix} = \begin{bmatrix}\text{Systematic}\\ \text{variation}\end{bmatrix} + \begin{bmatrix}\text{Random}\\ \text{variation}\end{bmatrix}$$

or

$$\begin{bmatrix}\text{Variation}\\ \text{in } Y_i\end{bmatrix} = \begin{bmatrix}\text{Explained}\\ \text{variation}\end{bmatrix} + \begin{bmatrix}\text{Unexplained}\\ \text{variation}\end{bmatrix}$$

The first component in brackets is the part of the variation in Y explained by the changes in X and the second is the part of the variation not explained by any specific factor, that is to say the variation in Y is due to the random influence of u.

Seen in this light the random term u seems to have a meaning related to the *ceteris paribus* clause of economic theory. Economic theory assumes that the functional relationships between variables are exact under the *ceteris paribus* clause. For example, the demand function $D = b_0 + b_1 P$ postulated by economic theory implies that the quantity of a particular commodity is a linear function of its price alone, 'other things remaining equal'; that is, the price–quantity relationship holds provided that all other factors not appearing explicitly in the function (for example tastes, income, other prices) remain unchanged. However, theories are simplifications of the complex relationships which exist in the real world, so that the *ceteris paribus* clause is very seldom fulfilled. When we collect data on the quantity that would be bought if all 'other things' were constant, but rather the quantities purchased while the prices of various prices we do not observe the quantities purchased while the prices of other goods, incomes, tastes and other factors have all been changing.

In econometrics we may read the true relationship connecting the variables as follows. Y is connected with X by a linear relationship, *ceteris paribus*. If factors other than X remain unchanged then changes in Y would be fully explained by changes in X. However, other factors do not remain equal; hence we introduce u into the function to account for the changes in other variables not included in it explicitly.

We may now look at the final form of our equation $Y_i = b_0 + b_1 X_i + u_i$ in another way. For a given value of X, Y may assume various values depending on the particular (positive or negative) value that u happens to assume. To each value of X corresponds a distribution of various values of u, and therefore Y's. This situation is pictured in figure 4.3. For example if the price of the commodity is equal to X_1, the quantity which will be supplied at this price may assume any value between Y_1' and Y_1'', depending on the value of u in this period. If, for instance, there is a strike of lorry drivers, or a power cut, which delays the delivery of the commodity (these situations being examples of chance events), the quantity will not be Y_1, as the linear equation suggests, but a smaller quantity Y_1^*, due to the above factors which give a value u_1^* to the random term. If, however, there is a rumour of a fall in prices of substitutes or of a new product being developed, the supplier may offer all the stock, which

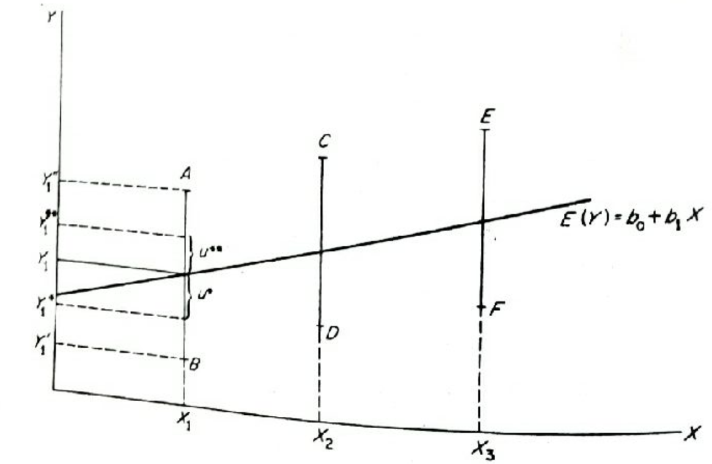


Figure 4.3

he otherwise would offer in future periods, so that at the price X_1 the quantity supplied would be Y_1^{**}, because the change in expectations caused u to assume the value u_1^{**}.

To estimate the coefficients b_0 and b_1 we need observations on X, Y and u. Yet u is never observed like the other explanatory variables,[1] and therefore in order to estimate the function $Y_i = b_0 + b_1 X_i + u_i$, we should 'guess' the values of u, that is we should make some reasonable (plausible) assumptions about the shape of the distribution of each u_i (its mean, variance and covariance[2] with other u's). These assumptions are guesses about the true, but unobservable, values of u_i.

4.2. ASSUMPTIONS OF THE LINEAR STOCHASTIC REGRESSION MODEL

The linear regression model is based on certain assumptions, some of which refer to the distribution of the random variable u, some to the relationship between u and the explanatory variables, and finally some refer to the relationship between the explanatory variables themselves. We will group the assumptions in two categories, (a) stochastic assumptions, (b) other assumptions.

4.2.1. STOCHASTIC ASSUMPTIONS OF ORDINARY LEAST SQUARES

These are assumptions about the distribution of the values of u. They are crucial for the estimates of the parameters and will be explained in detail in subsequent chapters (see Chapters 9–12). It is these assumptions about the random term u that adapt the least squares method, which is a statistical method, to the stochastic nature of economic phenomena. At this stage we will state these assumptions without attempting to explain their implications for the parameter estimates.

Assumption 1 u_i is a random real variable.

The value which u_i may assume in any one period depends on chance; it may be positive, negative or zero. Each value has a certain probability of being assumed by u in any particular instance.

Assumption 2 The mean value of u in any particular period is zero.

This means that for each value of X, u may assume various values, some greater than zero and some smaller than zero, but if we considered all the possible values of u, for any given value of X, they would have an average value equal to zero. With this assumption we may say that $Y_i = b_0 + b_1 X_i$ gives the relationship between X and Y *on the average*, that is, when X

[1] As we shall see readily, we can get an estimate of the u's after the estimation of the regression line and the computation of the residual deviations of the observations from this line.

[2] The covariance of the u's measures the way in which the u's of different periods tend to covary. The covariance of u's and X's measures the way in which the values of u's of different periods tend to vary with the values of X in these periods. (See Appendix I.)

assumes the value X_i the dependent variable will on the average assume the value Y_i (on the line), although the actual value of Y observed in any particular occasion may display some variation: sometimes the value of the dependent variable (corresponding to the given value of X) will be bigger than Y_i, and at other times it will be smaller than the Y_i (on the line). Yet on the average the value of Y will be equal to Y_i when X assumes the value X_i. That is, on the average u is equal to zero.

Assumption 3 The variance of u_i is constant in each period.

The variance of u_i about its mean is constant at all values of X. In other words for all values of X, the u's will show the same dispersion round their mean. In figure 4.3 this assumption is denoted by the fact that the values that u may assume lie within the same limits, irrespective of the value of X: for X_1, u can assume any value within the range AB; for X_2, u can assume any value within the range CD which is equal to AB and so on.

Assumption 4 The variable u_i has a normal distribution.

The values of u (for each X_i) have a bell-shaped symmetrical distribution about their zero mean.

The above four assumptions about the behaviour (distribution) of the values of u may be summarised by the expression

$$u \sim N(0, \sigma_u^2)$$

and are pictured in figure 4.4.

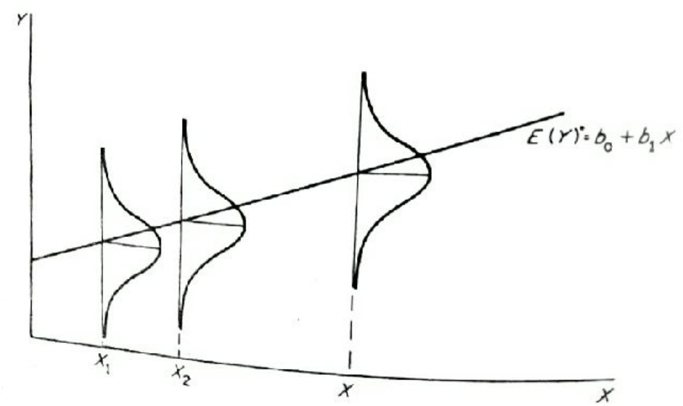


Figure 4.4

Assumption 5 The random terms of different observations (u_i, u_j) are independent.

This means that all the covariances of any u_i with any other u_j are equal to zero. The value which the random term assumed in one period does not depend on the value which it assumed in any other period.

Assumption 6 u is independent of the explanatory variable(s).

The disturbance term is not correlated with the explanatory variable(s). The u's and the X's do not tend to vary together; their covariance is zero. Symbolically

$$\text{cov}(Xu) = E\{[X_i - E(X_i)][u_i - E(u_i)]\} = 0$$

It is, however, conceptually easier and computationally more convenient to make an alternative assumption which ensures zero covariance of the u's and X's.

Assumption 6A The X_i's are a set of fixed values in the hypothetical process of repeated sampling which underlies the linear regression model.

This means that, in taking a large number of samples on Y and X, the X_i values are the same in all samples, but the u_i values do differ from sample to sample, and so of course do the values of Y_i. For example, assume that every day in a market we choose the same prices $X_1, X_2, \ldots, X_n$, and we record the quantities Y_i's sold each day at these prices. The X's do not vary, they are a set of fixed values; while the Y_i's vary for each day due to different random influences. Clearly, under these conditions the covariance of the (fixed) X's and the u's is zero. Because

$$\begin{aligned}
\text{cov}(Xu) &= E\{[X_i - E(X_i)][u_i - E(u_i)]\} \\
&= E\{[X_i - E(X_i)]u_i\} && \text{given } E(u_i) = 0 \\
&= E(X_i u_i) - E(X_i)E(u_i) \\
&= E(X_i u_i) \\
&= X_i E(u_i) && \text{given that the } X_i\text{'s are fixed} \\
&= 0
\end{aligned}$$

In the remainder of this book we will mostly use Assumption 6A, that the explanatory variables are fixed.

Assumption 7 The explanatory variable(s) are measured without error.

u absorbs the influence of omitted variables and possibly errors of measurement in the Y's. That is, we will assume that the regressors are error-free, while the Y values may or may not include errors of measurement.

4.2.2. OTHER ASSUMPTIONS OF ORDINARY LEAST SQUARES

Assumption 8 The explanatory variables are not perfectly linearly correlated.

If there is more than one explanatory variable in the relationship it is assumed that they are not perfectly correlated with each other. Indeed the regressors should not even be strongly correlated, they should not be highly multicollinear.

Assumption 9 The macrovariables should be correctly aggregated.

Usually the variables X and Y are aggregative variables, representing the sum of individual items. For example, in a consumption function $C = b_0 + b_1 Y + u$, C is the sum of the expenditures of all consumers and Y is the sum of all individual incomes. It is assumed that the appropriate aggregation procedure has been adopted in compiling the aggregate variables.

Assumption 10 The relationship being estimated is identified.

It is assumed that the relationship whose coefficients we want to estimate has a unique mathematical form, that is it does not contain the same variables as any other equation related to the one being investigated. Only if this assumption is fulfilled can we be certain that the coefficients which result from our computations are the true parameters of the relationship which we study.

Assumption 11 The relationship is correctly specified.

It is assumed that we have not committed any specification error in determining the explanatory variables, that we have included all the important regressors explicitly in the model, and that its mathematical form (number of equations and their linear or nonlinear nature) is correct.

4.3. THE DISTRIBUTION OF THE DEPENDENT VARIABLE Y

In this section we will establish that the dependent variable Y has a normal distribution with mean

$$E(Y_i) = b_0 + b_1 X_i \tag{4.1}$$

and variance

$$\text{var}(Y_i) = E[Y_i - E(Y_i)] = E(u_i^2) = \sigma_u^2 \tag{4.2}$$

Proof 1. The mean of $Y_i = E(Y_i) = b_0 + b_1 X_i$.
By definition the mean of Y_i is its expected value.

Given $$Y_i = b_0 + b_1 X_i + u_i$$

Taking expected values we find

$$E(Y_i) = E[b_0 + b_1 X_i + u_i]$$
$$= E(b_0 + b_1 X_i) + E(u_i)$$

Given that b_0 and b_1 are parameters and by Assumption 6A the values of X_i's are a set of fixed numbers (in the process of hypothetical repeated sampling)

$$E(b_0 + b_1 X_i) = b_0 + b_1 X_i$$

Furthermore, by Assumption 2

$$E(u_i) = 0$$

Therefore,

$$E(Y_i) = b_0 + b_1 X_i$$

Proof 2. The variance of $Y_i = E[Y_i - E(Y_i)]^2 = \sigma_u^2$.
Substitute $Y_i = b_0 + b_1 X_i + u_i$ and $E(Y_i) = b_0 + b_1 X_i$ in the definition of the variance

$$E[Y_i - E(Y_i)]^2 = E[b_0 + b_1 X_i + u_i - b_0 - b_1 X_i]^2 = E(u_i)^2$$

But, by Assumption 3, the u_i's are homoscedastic, that is, they have the constant variance σ_u^2

$$E(u_i^2) = \sigma_u^2 \text{ constant}$$

Therefore,

$$\text{var}(Y_i) = E[Y_i - E(Y_i)]^2 = \sigma_u^2$$

Proof 3. The distribution of Y_i is normal.
The shape of the distribution of Y_i is determined by the shape of the distribution of u_i, which is normal by Assumption 4. Clearly b_0 and b_1, being constants, do not affect the distribution of Y_i. Furthermore the values of the explanatory variable, X_i, are a set of constant values by Assumption 6A and therefore do not affect the shape of the distribution of Y_i.

4.4. THE LEAST SQUARES CRITERION AND THE 'NORMAL' EQUATIONS OF OLS

Thus far we have completed the work involved in the first stage of any econometric application, namely we have specified the model and stated explicitly its assumptions. The next step is the estimation of the model, that is, the computation of the numerical values of its parameters.

The linear relationship $Y_i = b_0 + b_1 X_i + u_i$ holds for the population of the values of X and Y, so that we could obtain *the* numerical values of b_0 and b_1 only if we could have all the conceivably possible values of X, Y, and u which form the population of these variables. Since this is impossible in practice, we get a sample of observed values of Y and X, we specify the distribution of the u's and we try to get satisfactory estimates of the true parameters of the relationship. This is done by fitting a regression line through the observations of the sample, which we consider as an approximation to the true line. The true relationship between X and Y is

$$Y_i = b_0 + b_1 X_i + u_i$$

the true regression line is

$$E(Y_i) = b_0 + b_1 X_i$$

the estimated relationship is

$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + e_i$$

and the estimated regression line is

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

where $\hat{Y}$ = estimated value of Y, given a specified value of X
$\hat{b}_0$ = estimate of the true intercept b_0
$\hat{b}_1$ = estimate of the true parameter b_1
e = estimate of the true value of the random term u.

The true and the estimated regression lines are shown in figure 4.5. In our example of the supply function, in order to compute the numerical values of the

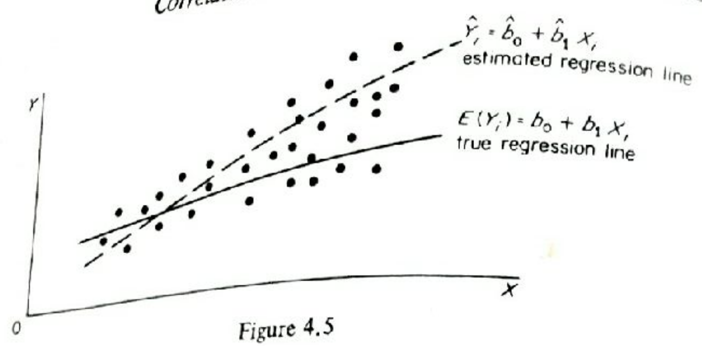
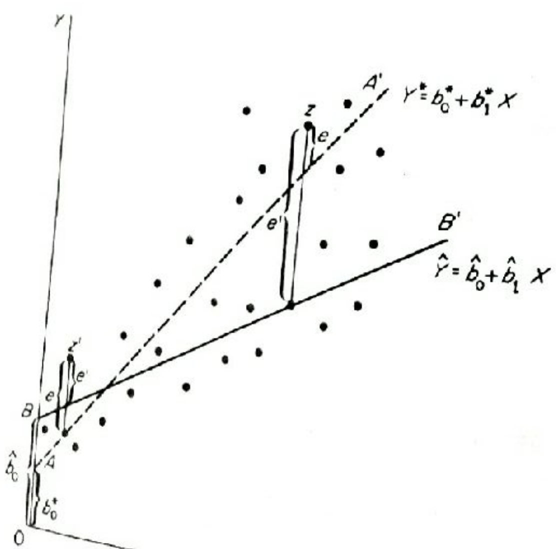


Figure 4.5

true parameters b_0 and b_1 we should have all the conceivable values of quantities supplied at all conceivable prices, which of course is impossible. Consequently we take a sample of observed prices and quantities sold over some period of time and we attempt to obtain the best possible estimate of the supply function.

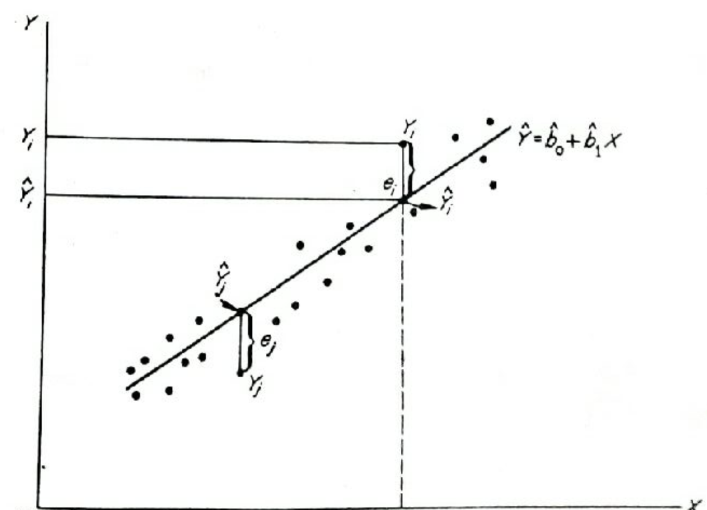
The snag in this procedure is that from a given sample we may obtain an infinite number of estimated regression lines, by assigning different values to the parameters b_0 and b_1. In figure 4.6 we have drawn two such lines, AA' and BB'. When we assign to the parameters the values b_0^* and b_1^* we get the line $AA' = b_0^* + b_1^* X$, while if the parameters are given the values $\hat{b}_0$ and $\hat{b}_1$, the line will be $BB' = \hat{b}_0 + \hat{b}_1 X$, and so on. It is clear, however, that the deviations of the actual sample observations from each line are different. For example point z is closer to line AA', while point z' is nearer to line BB'. In other words if we choose the upper line $Y^* = b_0^* + b_1^* X$, point z will deviate by e, while if we take the line BB' $(\hat{Y} = \hat{b}_0 + \hat{b}_1 X)$, the same point z will deviate from it by a greater distance equal to e' $(e' > e)$.



Clearly the deviations of the observations from the lines depend on their constant intercept (b_0) and their slope (b_1). The choice among all possible lines is done on the basis of what is called *the least squares criterion*. The rationale of this criterion is easy to understand. It is intuitively obvious that the smaller the deviations from the line, the better the fit of the line to the scatter of observations. Consequently from all possible lines we choose the one for which the deviations of the points is the smallest possible. The *least squares criterion* requires that the regression line be drawn (i.e. its parameters be chosen) in such a way as to minimise the sum of the squares of the deviations of the observations from it.

The first step is to draw the line so that the sum of the simple deviations of the observations is zero – some observations will lie above the line and will have a positive deviation, some will lie below the line, in which case they will have a negative deviation, and finally the points lying on the line will have a zero deviation. In summing these deviations the positive values will offset the negative values, so that the final algebraic sum of these residuals will equal zero by definition ($\Sigma e \equiv 0$). This of course does not mean that the deviations disappear when we fit the least squares line, but that their algebraic sum is by construction equal to zero. How then, can one minimise a quantity which is by definition zero? The best solution is to square the deviations and minimise the sum of the squares, (Σe_i^2). The reason for calling this method *the least squares method* should now be clear: the method seeks the minimisation of the *sum of the squares* of the deviations of the actual observations from the line.

Our next task is to express the residual deviations (e's) in terms of the observed values of Y and X in our sample. In figure 4.7 the estimated line is $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$. As already mentioned the sign (^) on top of the dependent variable indicates the estimated (predicted) value of the dependent variable, as distinguished from the observed value of this variable, which is represented by



the simple letter Y_i. If $\hat{b}_0$ and $\hat{b}_1$ are numerically known, from the estimated line we can obtain a prediction of Y, that is, an 'estimated' value of the dependent variable (Y_i) which corresponds to a given value of the explanatory variable (X_i). That is, for each given X, the corresponding $\hat{Y}$ lies on the line. For example when X assumes the value X_i, the equation predicts that the dependent variable will assume the (estimated) value $\hat{Y}_i$. However, the actually observed value of the dependent variable which corresponds to X_i, is Y_i, and not $\hat{Y}_i$ as the line predicts. In other words, the actual observations of Y may not lie on the estimated line. It is apparent that the equation does not predict the values of the dependent variable with perfect accuracy. We have denoted by e_i the difference between the observed value Y_i and its estimated value $\hat{Y}_i$, that is

$$e_i = Y_i - \hat{Y}_i$$

Substituting $\hat{Y}_i$ we find

$$e_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$$

Squaring these deviations and taking their sum we obtain

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$

The sum of squared residual deviations is to be minimised with respect to $\hat{b}_0$ and $\hat{b}_1$. Following the minimisation procedure we get the *normal equations*

$$\Sigma Y = n\hat{b}_0 + \hat{b}_1 \Sigma X \qquad (4.3)$$

$$\Sigma XY = \hat{b}_0 \Sigma X + \hat{b}_1 \Sigma X^2 \qquad (4.4)$$

Formal derivation of the normal equations

We have to minimise the function

$$\Sigma e_i^2 = \Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$

with respect to $\hat{b}_0$ and $\hat{b}_1$. The necessary condition for a minimum is that the first derivatives of the function be equal to zero

$$\frac{\partial \Sigma e^2}{\partial \hat{b}_0} = 0 \quad \text{and} \quad \frac{\partial \Sigma e^2}{\partial \hat{b}_1} = 0$$

To obtain the above derivatives we apply the 'function of a function' rule of differentiation. According to this rule if $y = f(w)$ and $w = f(x)$,

then

$$\frac{dy}{dx} = \frac{dy}{dw} \cdot \frac{dw}{dx}$$

In the case of the above function we let $(Y_i - \hat{b}_0 - \hat{b}_1 X_i) = w$. Thus we have:

Partial derivative with respect to $\hat{b}_0$

$$\frac{\partial \Sigma e^2}{\partial \hat{b}_0} = \frac{\partial \Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2}{\partial \hat{b}_0} = 0$$

$$2\Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i) \, . \, (-1) = 0$$

$$\Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i) = 0 \qquad (4.5)$$

Partial derivative with respect to b_1

$$\frac{\partial \Sigma e^2}{\partial \hat{b}_1} = \frac{\partial \Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2}{\partial \hat{b}_1} = 0$$

$$2\Sigma (Y_i - \hat{b}_0 - \hat{b}_1 X_i) \, . \, (-X_i) = 0$$

$$\Sigma (Y_i X_i - \hat{b}_0 X_i - \hat{b}_1 X_i^2) = 0 \qquad (4.6)$$

Combining equations (4.5) and (4.6) and performing the summations we get

$$\Sigma Y_i - \Sigma \hat{b}_0 - \Sigma \hat{b}_1 X_i = 0$$

$$\Sigma Y_i X_i - \Sigma \hat{b}_0 X_i - \Sigma \hat{b}_1 X_i^2 = 0$$

Applying the usual summation rules (see Appendix I) we obtain the 'normal' equations of OLS

$$\Sigma Y_i = \hat{b}_0 n + \hat{b}_1 \Sigma X_i$$

$$\Sigma Y_i X_i = \hat{b}_0 \Sigma X_i + \hat{b}_1 \Sigma X_i^2$$

Solving the normal equations for $\hat{b}_0$ and $\hat{b}_1$ we obtain the least squares estimates[1]

$$\hat{b}_0 = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{n\Sigma X^2 - (\Sigma X)^2} \qquad (4.7)$$

$$\hat{b}_1 = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad (4.8)$$

It is clear that $\hat{b}_0$ and $\hat{b}_1$ can be estimated by substituting the terms n, ΣX, ΣY, ΣXY and ΣX^2, whose values can be obtained from the sample observations.

The above formulae are expressed in terms of the original sample observations on X and Y. It can be shown that the estimates $\hat{b}_0$ and $\hat{b}_1$ may be obtained by the following formulae which are expressed in deviations of the variables from their means:

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \qquad (4.9)$$

$$\hat{b}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \qquad (4.10)$$

Proof

(1) In Chapter 3 we established that $\Sigma x_i y_i = (n\Sigma XY - \Sigma X \Sigma Y)/n$. (This is the expression 3.5 on p. 37.)

(2) Similarly we have proved (expression 3.6 of Chapter 3) that

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

(3) Substituting in the expression for $\hat{b}_1$ we find

$$\hat{b}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{(n\Sigma XY - \Sigma X \Sigma Y)/n}{(n\Sigma X^2 - (\Sigma X)^2)/n} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

[1] The solution of a system of equations may be obtained by the use of various methods. In Appendix II we explain the method of determinants which is conceptually the simplest of all.

Table 4.2. Worksheet for the estimation of the supply function of commodity Z

n	Y_i Quantity (in tons)	X_i Price (in £ per ton)	X_i^2	X_iY_i	y_i $(Y_i - \bar{Y})$	x_i $(X_i - \bar{X})$	x_iy_i	x_i^2 $(X_i - \bar{X})^2$	$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$	e_i $(Y_i - \hat{Y}_i)$	e_i^2
1	69	9	81	621	+ 6	0	0	0	63.00	6.00	36.00
2	76	12	144	912	+13	+3	39	9	72.75	3.25	10.56
3	52	6	36	312	−11	−3	33	9	53.25	− 1.25	1.56
4	56	10	100	560	− 7	+1	−7	1	66.25	−10.25	105.06
5	57	9	81	513	− 6	0	0	0	63.00	− 6.00	36.00
6	77	10	100	770	+14	+1	14	1	66.25	10.75	115.56
7	58	7	49	406	− 5	−2	10	4	56.50	1.50	2.25
8	55	8	64	440	− 8	−1	8	1	59.75	− 4.75	22.56
9	67	12	144	804	+ 4	+3	12	9	72.75	− 5.75	30.06
10	53	6	36	318	−10	−3	30	9	53.25	− 0.25	0.06
11	72	11	121	792	+ 9	+2	18	4	69.50	2.50	6.25
12	64	8	64	512	+ 1	−1	−1	1	59.75	4.25	18.06
$n = 12$	$\Sigma Y_i = 756$	$\Sigma X_i = 108$	$\Sigma X_i^2 = 1{,}020$	$\Sigma X_i Y_i = 6{,}960$	$\Sigma y_i = 0$	$\Sigma x_i = 0$	$\Sigma x_i y_i = 156$	$\Sigma x_i^2 = 48$	$\Sigma \hat{Y}_i = 756.0$	$\Sigma e_i = 0$	$\Sigma e_i^2 = 383.98$
	$\bar{Y} = 63$	$\bar{X} = 9$							$\bar{\hat{Y}} = 63$		

Dividing the first normal equation through by n we obtain

$$\frac{\Sigma Y}{n} = \frac{\hat{b}_0 n}{n} + \frac{\hat{b}_1 \Sigma X}{n}$$

or

$$\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}$$

that is the regression line passes through the point defined by the means of the variables. This is a very useful result which we will use often in subsequent chapters.

Example. To illustrate the use of the above formulae we will estimate the supply function of commodity z using the data in table 4.2.

We substitute the computed values from table 4.2 into the formulae for $\hat{b}_0$ and $\hat{b}_1$.

(1) *Using the original sample observations*

$$\hat{b}_0 = \frac{\Sigma X^2 \Sigma Y - (\Sigma X)(\Sigma XY)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(1{,}020)(756) - (108)(6{,}960)}{(12)(1{,}020) - (108)^2} = \frac{19{,}440}{576} = 33{\cdot}75$$

$$\hat{b}_1 = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(12)(6{,}960) - (756)(108)}{(12)(1{,}020) - (108)^2} = \frac{1{,}872}{576} \approx 3{\cdot}25$$

(2) *Using the deviations of the variables from their means*

$$\hat{b}_1 = \frac{\Sigma xy}{\Sigma x^2} = \frac{156}{48} = 3{\cdot}25$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 63 - (3{\cdot}25)(9) = 33{\cdot}75$$

Thus the estimated supply function is

$$\hat{Y}_i = 33{\cdot}75 + 3{\cdot}25 X_i$$

4.5 ESTIMATION OF A FUNCTION WHOSE INTERCEPT IS ZERO

In some cases economic theory postulates relationships which have a zero constant intercept, that is, they pass through the origin of the XY plane. For example linear production functions of manufactured products should normally have zero intercept, since output is zero when the factor inputs are zero. In this event we should estimate the function

$$Y = b_0 + b_1 X + u$$

imposing the restriction $b_0 = 0$. The formula for the estimation of $\hat{b}_1$ then becomes

$$\hat{b}_1 = \frac{\Sigma XY}{\Sigma X^2}$$

which involves the actual values of the variables, and not their deviations, as in the case of unrestricted value of b_0.

Proof. We want to fit the line $Y = b_0 + b_1 X_1 + u$, subject to the restriction $b_0 = 0$. This is a restricted minimisation problem: we minimise

$$\Sigma e^2 = \Sigma(Y - \hat{b}_0 - \hat{b}_1 X)^2$$

4. The following results have been obtained from a sample of 11 observations on the value of sales (Y) of a firm and the corresponding prices (X).

$$\bar{X} = 519{\cdot}18 \quad \bar{Y} = 217{\cdot}82$$
$$\Sigma X_i^2 = 3{,}134{,}543 \qquad \Sigma X_i Y_i = 1{,}296{,}836 \qquad \Sigma Y_i^2 = 539{,}512$$

(i) Estimate the regression line of sales on price and interpret the results.
(ii) What is the part of the variation in sales which is not explained by the regression line?
(iii) Estimate the price elasticity of sales.

5. The following table gives the quantities of commodity z bought in each year from 1961–1970 and the corresponding prices.

Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Quantity (in tons)	770	785	790	795	800	805	810	820	840	850
Price (in £)	18	16	15	15	12	10	10	·7	9	6

(i) Estimate the linear demand function for commodity z.
(ii) Calculate the price elasticity of demand.
(iii) Forecast the demand at the mean price of the sample.
(iv) Forecast the demand at $P = 20$.

Note. Additional exercises are included in Appendix III.

5. Statistical Tests of Significance of the Least Squares Estimates: First-Order Tests

In Chapter 4 we developed the formulae for the estimation of the parameters of economic relationships by using the method of least squares. The next stage is to establish criteria for judging the 'goodness' of the parameter estimates. We divide the available criteria into three groups: theoretical *a priori* criteria, statistical criteria and econometric criteria. The theoretical criteria are set by economic theory and refer to the sign and size of the coefficients. They are defined in the first stage of econometric research, that is in the stage of the specification of the model (see Chapter 2). In this chapter we shall develop the *statistical criteria* or *first-order tests* for the evaluation of the parameter estimates. The *econometric criteria* or *second-order tests* will be examined in subsequent chapters.

The two most commonly used tests in econometrics are the following:

The first is the square of the correlation coefficient, r^2, which is used for judging the explanatory power of the linear regression of Y on X. We will prove that r^2 is a measure of the goodness of fit of the regression line to the observed sample values of Y and X.

The second test is based on the standard errors of the parameter estimates and is applied for judging the statistical reliability of the estimates of the regression coefficients $\hat{b}_0$ and $\hat{b}_1$. It provides a measure of the degree of confidence that we may attribute to the estimates $\hat{b}_0$ and $\hat{b}_1$. It enables the researcher to decide how 'good' estimates of the true parameters of the (population) relationship $\hat{b}_0$ and $\hat{b}_1$ are. In Chapter 8 we shall develop an alternative statistical technique for judging the significance of the OLS results, the *Analysis of Variance* technique.

5.1. THE TEST OF THE GOODNESS OF FIT WITH r^2

5.1.1. DEFINITION OF r^2

After the estimation of the parameters and the determination of the least squares regression line, we need to know how 'good' is the fit of this line to the sample observations of Y and X, that is to say we need to measure the dispersion of observations around the regression line. This knowledge is essential, because the closer the observations to the line, the better the goodness of fit, that is the better is the explanation of the variations of Y by the changes in the explanatory variables.

We will prove that a measure of the goodness of fit is the square of the correlation coefficient, r^2, which shows the percentage of the total variation of

variable X.
We plot the observations on a rectangular co-ordinate system. Next we compute the means

$$\bar{X} = \sum_i X_i/n \quad \text{and} \quad \bar{Y} = \sum_i Y_i/n$$

and we draw perpendiculars through the points of these means (figure 5.1).

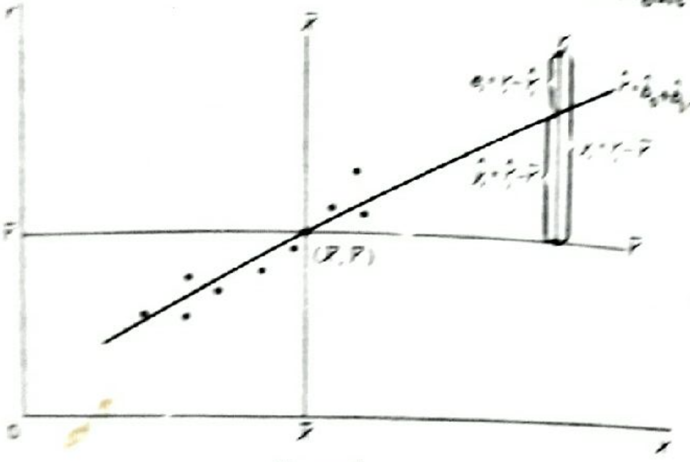


Figure 5.1

By fitting the line $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ we try to obtain the explanation of the variations of the dependent variable Y produced by the changes of the explanatory variable X. However, the fact that the observations deviate from the estimated line shows that the regression line explains only a part of the total variation of the dependent variable. A part of the variation, defined as $e_i = Y_i - \hat{Y}$ remains unexplained.

(1) We may compute the total variation of the dependent variable by comparing each value of Y to the mean value $\bar{Y}$ and adding all the resulting deviations.[1] Denoting the deviations of the values Y_i around their mean $\bar{Y}$ by lower case letters we have

$$[\text{Total variation in } Y] = \sum_i^n y_i^2 = \sum_i^n (Y_i - \bar{Y})^2 \qquad (5.1)$$

Note that in order to find the total variation of the Y's we square the simple deviations, since by definition the sum of the simple deviations of any variable around its mean is identically equal to zero

$$\sum_i^n (Y_i - \bar{Y}) = \sum_i^n y_i = 0$$

[1] When we speak of changes in Y we must define the 'basis of reference', that is, a value of the variable Y, to which we compare any other value that may be assumed by this variable. As such reference value we may take the origin ($Y = 0$) or the mean value ($\bar{Y}$) or any other statistic of Y (the median, etc.). However, it is customary and computationally convenient to take the mean as reference value and express the total variation of Y's as the sum of the deviations of the Y's from their mean.

(2) In the same way we define the deviation of the regressed (that is the estimated from the line) values, $\hat{Y}$'s, from the mean value: $\hat{y}_i = \hat{Y}_i - \bar{Y}$. This is the part of the total variation of Y_i which is explained by the regression line. Thus the sum of the squares of these deviations is the total *explained by the regression line variation* of the dependent variable

$$[\text{Explained Variation}] = \sum_i^n \hat{y}_i^2 = \sum_i^n (\hat{Y}_i - \bar{Y})^2 \qquad (5.2)$$

(3) We have already defined the residual e_i as the difference $e_i = Y_i - \hat{Y}_i$, that is as the part of the variation of the dependent variable which is not explained by the regression line and is attributed to the existence of the disturbance variable u. Thus the sum of the squared residuals gives the total unexplained variation of the dependent variable Y around its mean

$$[\text{Unexplained Variation}] = \sum_i^n e_i^2 = \sum_i^n (Y_i - \hat{Y}_i)^2 \qquad (5.3)$$

In summary

$e_i = Y_i - \hat{Y}_i$ = deviation of the observations Y_i from the regression line
$y_i = Y_i - \bar{Y}$ = deviation of Y_i from its mean
$\hat{y}_i = \hat{Y}_i - \bar{Y}$ = deviation of the regressed value $\hat{Y}_i$ from the mean

Combining these expressions we obtain

$$Y_i = y_i + \bar{Y} \quad \text{and} \quad \hat{Y}_i = \hat{y}_i + \bar{Y}$$

Substituting in the expressions of the residuals we find

$$e_i = (y_i + \bar{Y}) - (\hat{y}_i + \bar{Y}) \qquad (5.4)$$

$$e_i = y_i - \hat{y}_i$$

and

$$y_i = \hat{y}_i + e_i. \qquad (5.5)$$

This equation shows that each deviation of the observed values of Y from its mean consists of two components: the first is the explained by the regression line variation and the second is the unexplained variation. This relationship is shown in figure 5.1.

Substituting 5.5 into 5.1 we obtain

$$\sum_i y_i^2 = \sum_i (\hat{y}_i + e_i)^2$$
$$= \sum_i \hat{y}_i^2 + \sum_i e_i^2 + 2\sum_i \hat{y}_i e_i$$

But $\sum_i \hat{y}_i e_i = 0$.

This can be proved as follows:
(a) We know that $y_i = \hat{Y}_i - \bar{Y}$. But $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$ and $\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}$.
Therefore $\hat{y}_i = (\hat{b}_0 + \hat{b}_1 X_i) - (\hat{b}_0 + \hat{b}_1 \bar{X}) = \hat{b}_1 (X_i - \bar{X}) = \hat{b}_1 x_i$, where $x_i = X_i - \bar{X}$
(b) We also know that $e_i = y_i - \hat{y}_i = y_i - \hat{b}_1 x_i$.
Therefore $\sum \hat{y}_i e_i = \sum (\hat{b}_1 x_i)(y_i - \hat{b}_1 x_i) = \hat{b}_1 (\sum x_i y_i - \hat{b}_1 \sum x_i^2)$.

But $\hat{b}_1 = \Sigma x_i y_i / \Sigma x_i^2$.
Therefore we may write:

$$\Sigma \hat{y}_i e_i = \hat{b}_1 \left(\Sigma x_i y_i - \frac{\Sigma x_i y_i}{\Sigma x_i^2} \cdot \Sigma x_i^2 \right) = 0$$

Therefore

$$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \tag{5.6}$$

or

$$\begin{bmatrix} \text{Total} \\ \text{Variation} \end{bmatrix} = \begin{bmatrix} \text{Explained} \\ \text{Variation} \end{bmatrix} + \begin{bmatrix} \text{Unexplained} \\ \text{(residual) Variation} \end{bmatrix}$$

The explained variation expressed as a percentage of total variation is $\Sigma \hat{y}^2 / \Sigma y^2$.

But $\hat{y} = \hat{b}_1 x$. Substituting we find

$$\frac{\Sigma \hat{y}^2}{\Sigma y^2} = \frac{\Sigma (\hat{b}_1 x)^2}{\Sigma y^2} = \hat{b}_1^2 \frac{\Sigma x^2}{\Sigma y^2}$$

Given that $\hat{b}_1 = \Sigma xy / \Sigma x^2$, we get

$$\frac{\Sigma \hat{y}^2}{\Sigma y^2} = \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} \cdot \frac{(\Sigma x^2)}{\Sigma y^2} = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)}$$

Comparing this result with the formula of the correlation coefficient developed in Chapter 3 we see that

$$\frac{\Sigma \hat{y}^2}{\Sigma y^2} = r^2 \tag{5.7}$$

since

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

Thus r^2 determines the proportion of the variation in Y which is explained by variations in X. For this reason r^2 is sometimes called the *coefficient of determination*. For example, if $r^2_{Y.X} = 0{\cdot}90$, this means that the regression line gives a good fit to the observed data, since this line explains 90 per cent of the total variation of the Y values around their mean. The remaining 10 per cent of the total variation in Y is unaccounted for by the regression line and is attributed to the factors included in the disturbance variable u.

5.1.2. LIMITING VALUES OF THE COEFFICIENT OF DETERMINATION, r^2

It can be proved that the coefficient of determination can assume values lying between zero and one, that is to say

$$0 \leqslant r^2 \leqslant 1$$

Proof. We have proved $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$. Dividing through by Σy^2 we get

$$1 = \frac{\Sigma \hat{y}^2}{\Sigma y^2} + \frac{\Sigma e^2}{\Sigma y^2} \quad \text{or} \quad 1 = r^2 + \frac{\Sigma e^2}{\Sigma y^2}$$

therefore

$$r^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2}$$

Recall that $\Sigma e^2 / \Sigma y^2$ is the proportion of the unexplained variation of the Y's around their mean $\bar{Y}$. If all the observations lie on the regression line, there will be no scatter of points; in other words the total variation of Y is explained completely by the estimated regression line, and consequently there will be no unexplained variation; that is $\Sigma e^2 / \Sigma y^2 = 0$ and hence $r^2 = 1$. On the other hand, if the regression line explains only part of the variation in Y, there will be some unexplained variation, $(\Sigma e^2 / \Sigma y^2 > 0)$. Therefore r^2 will be smaller than 1. Finally if the regression line does not explain any part of the variation of Y, $\Sigma e^2 / \Sigma y^2 = 1$, because $\Sigma y^2 = \Sigma e^2$. Therefore in this case $r^2 = 0$.

5.1.3. RELATIONSHIP BETWEEN r^2 AND THE SLOPE $\hat{b}_1$

The relationship between the square of the correlation coefficient, r^2, and the slope of the regression line is given by the formula

$$r^2 = \hat{b}_1 \frac{\Sigma xy}{\Sigma y^2} \tag{5.8}$$

Proof. We found that

$$r^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)}$$

Rearranging slightly we obtain

$$r^2 = \frac{(\Sigma xy)}{(\Sigma x^2)} \cdot \frac{(\Sigma xy)}{(\Sigma y^2)}$$

But $\Sigma xy / \Sigma x^2 = \hat{b}_1$. Hence $r^2 = \hat{b}_1 \cdot (\Sigma xy / \Sigma y^2)$.

In summary, r^2 may be computed in various ways

$$r^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)}$$

$$r^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2} \quad \text{or} \quad r^2 = \hat{b}_1 \cdot \frac{\Sigma xy}{\Sigma y^2} \quad \text{or} \quad r^2 = \hat{b}_1^2 \cdot \frac{\Sigma x^2}{\Sigma y^2}$$

Example. The coefficient of determination of the supply function estimated in Chapter 4 is found as follows.

$$\hat{Y}_i = 33{\cdot}75 + 3{\cdot}25\, X_i$$

$$\Sigma e_i^2 = 383{\cdot}98$$

$$\Sigma y_i^2 = 894$$

Thus

$$r^2_{YX} = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

$$= 0{\cdot}570$$

5.2. TESTS OF SIGNIFICANCE OF THE PARAMETER ESTIMATES

Since $\hat{b}_0$ and $\hat{b}_1$ are sample estimates of the parameters b_0 and b_1 we must test their statistical reliability. In order to apply the standard tests of significance we must, among other things, know the mean and variance.[1] We will first develop formulae for the computation of the mean and variance of the least squares estimates. We will next explain the procedure of the standard error test and the 't' test for judging the statistical significance of the estimates. Finally, we explain the construction of confidence intervals for the estimates $\hat{b}_0$ and $\hat{b}_1$.

5.2.1. MEAN AND VARIANCE OF THE LEAST-SQUARES PARAMETER ESTIMATES

In this section we will establish the following results

(1) Mean of $\hat{b}_0$:

$$E(\hat{b}_0) = b_0. \tag{5.9}$$

(2) Variance of $\hat{b}_0$:

$$\text{var}(\hat{b}_0) = E[\hat{b}_0 - b_0]^2 = \sigma_u^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2} \tag{5.10}$$

(3) Mean of $\hat{b}_1$:

$$E(\hat{b}_1) = b_1. \tag{5.11}$$

(4) Variance of $\hat{b}_1$:

$$\text{var}(\hat{b}_1) = E[\hat{b}_1 - b_1]^2 = \sigma_u^2 \frac{1}{\Sigma x_i^2} \tag{5.12}$$

(5) Estimate of the variance of u:

$$\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-K} \tag{5.13}$$

where K = total number of parameters estimated from the regression.

5.2.2. THE MEAN OF $\hat{b}_1$

We *assume* that we draw repeated samples of size n from the population of Y and X, and for each sample we estimate the parameters $\hat{b}_0$ and $\hat{b}_1$. This is known as *hypothetical repeated sampling procedure*. If all the possible samples are taken, then the mean value of $\hat{b}_1$ will be its expected value: (mean $\hat{b}_1$) = $E(\hat{b}_1)$. To find the value of the mean in terms of the observations of our sample of Y and X we work as follows.

We established that

$$\hat{b}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

Substituting $y_i = (Y_i - \bar{Y})$ we obtain

$$\hat{b}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{\Sigma x_i (Y_i - \bar{Y})}{\Sigma x_i^2} = \frac{\Sigma x_i Y_i}{\Sigma x_i^2} - \frac{\bar{Y}\Sigma x_i}{\Sigma x_i^2}$$

But by definition the sum of the deviations of a variable from its mean is identically equal to zero, $\Sigma x_i = 0$. Therefore

$$\hat{b}_1 = \frac{\Sigma x_i Y_i}{\Sigma x_i^2} = \Sigma \left[\frac{x_i}{\Sigma x_i^2} Y_i \right] \tag{5.14}$$

[1] See Appendix I.

By Assumption 6A of the method of least squares, the values of X are a set of fixed values, which do not change from sample to sample. Consequently the ratio $x_i/\Sigma x_i^2$ will be constant from sample to sample, and if we denote this ratio by k_i we may write the estimate $\hat{b}_1$ in the form

$$\hat{b}_1 = \Sigma k_i Y_i.$$

By substituting the value of $Y_i = b_0 + b_1 X_i + u_i$ and rearranging the factors in the resultant expression we find

$$\hat{b}_1 = \Sigma k_i (b_0 + b_1 X_i + u_i)$$

$$= b_0 \Sigma k_i + b_1 \Sigma k_i X_i + \Sigma k_i u_i$$

But $\Sigma k_i = 0$ and $\Sigma k_i X_i = 1$.

Proof 1

$$\Sigma k_i = \frac{\Sigma x_i}{\Sigma x_i^2} = \frac{\Sigma (X_i - \bar{X})}{\Sigma x_i^2} = \frac{0}{\Sigma x_i^2} = 0$$

Proof 2

$$\Sigma\, k_i X_i = \frac{\Sigma x_i X_i}{\Sigma x_i^2} = \frac{\Sigma (X_i - \bar{X}) X_i}{\Sigma x_i^2} = \frac{\Sigma X_i^2 - \bar{X}\Sigma X_i}{\Sigma x_i^2} = 1$$

given $\Sigma x_i^2 = \Sigma X_i^2 - \bar{X}\Sigma X_i$ (see Chapter 3, expression 3.6).

Therefore

$$\hat{b}_1 = b_1 + \Sigma k_i u_i = b_1 + \frac{\Sigma x_i u_i}{\Sigma x_i^2}$$

Taking expected values, and noting that by Assumption 6A the X_i's are fixed, we obtain

$$E(\hat{b}_1) = E(b_1) + E\ \frac{\Sigma x_i u_i}{\Sigma x_i^2} = E(b_1) + \frac{\Sigma x_i E(u_i)}{\Sigma x_i^2}$$

Since b_1, the true population parameter is constant, $E(b_1) = b_1$. Furthermore by Assumption 2 the mean value of u is zero ($E(u_i) = 0$), so that the second term in the right-hand side vanishes and we have

$$\text{Mean of } \hat{b}_1 = E(\hat{b}_1) = b_1.$$

The mean of the ordinary least squares estimate $\hat{b}_1$ is equal to the true value of the population parameter b_1. This result has been established by making use of Assumption 2 and Assumption 6A.

5.2.3. THE VARIANCE OF $\hat{b}_1$

It can be proved that

$$\text{var}(\hat{b}_1) = E[\hat{b}_1 - E(\hat{b}_1)]^2 = E[\hat{b}_1 - b_1]^2 = \sigma_u^2 \frac{1}{\Sigma x_i^2}$$

Proof. We established in 5.14 that

$$\hat{b}_1 = \frac{\Sigma x_i Y_i}{\Sigma x_i^2} = \Sigma k_i Y_i$$

where $k_i = \dfrac{x_i}{\Sigma x_i^2}$ = constant weights in the process of hypothetical repeated sampling.

Therefore

$$\text{var}(\hat{b}_1) = \text{var}(\Sigma k_i Y_i) = \Sigma k_i^2 \text{ var}(Y_i)$$

given that $k_i = x_i/\Sigma x_i^2$ are constant weights, independent of the values of Y_i by Assumption 6A.

But $\text{var}(Y_i) = \sigma_u^2$ (see Chapter 4, expression 4.2). Therefore

$$\text{var}(\hat{b}_1) = \Sigma k_i^2 \, \sigma_u^2 = \sigma_u^2 \, \Sigma k_i^2$$

$$= \sigma_u^2 \, \Sigma \left(\frac{x_i}{\Sigma x_i^2} \right)^2 = \sigma_u^2 \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2}$$

$$= \sigma_u^2 \frac{1}{\Sigma x_i^2}$$

5.2.4. THE MEAN OF $\hat{b}_0$

It can be proved that

$$E(\hat{b}_0) = b_0$$

Proof. We have established in Chapter 4 (expression 4.9) that

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

Substituting $\hat{b}_1 = \Sigma k_i Y_i$ we obtain

$$\hat{b}_0 = \overline{Y} - \overline{X} \Sigma k_i Y_i = \frac{\Sigma Y}{n} - \overline{X} \Sigma k_i Y_i$$

Taking Y_i as a common factor we may write

$$\hat{b}_0 = \Sigma \left[\frac{1}{n} - \overline{X} k_i \right] Y_i \qquad (5.15)$$

Taking expected values

$$E(\hat{b}_0) = \Sigma \left[\frac{1}{n} - \overline{X} k_i \right] E(Y_i)$$

given that n, $\overline{X}$ and k_i are constant from sample to sample.

But in Chapter 4 (expression 4.1) we established that

Therefore $\qquad E(Y_i) = b_0 + b_1 X_i$

$$E(\hat{b}_0) = \Sigma \left[\frac{1}{n} - \overline{X} k_i \right] (b_0 + b_1 X_i)$$

$$= \Sigma \left[\frac{b_0}{n} - \overline{X} k_i b_0 + \frac{b_1 X_i}{n} - \overline{X} k_i b_1 X_i \right]$$

$$= b_0 + b_1 \overline{X} - b_1 \overline{X}$$

since $\Sigma k_i = 0$ and $\Sigma k_i X_i = 1$ (see page 75). Therefore

$$E(\hat{b}_0) = b_0$$

5.2.5. THE VARIANCE OF $\hat{b}_0$

It can be proved that

$$\text{var}(\hat{b}_0) = E[\hat{b}_0 - E(\hat{b}_0)]^2 = E[\hat{b}_0 - b_0]^2 = \sigma_u^2 \frac{\Sigma X^2}{n \Sigma x^2}$$

Proof. We established in 5.15 that

$$\hat{b}_0 = \Sigma \left[\frac{1}{n} - \overline{X} k_i \right] Y_i$$

Therefore

$$\text{var}(\hat{b}_0) = \text{var} \left[\Sigma \left(\frac{1}{n} - \overline{X} k_i \right) Y_i \right]$$

$$= \Sigma \left[\frac{1}{n} - \overline{X} k_i \right]^2 \text{var}(Y_i)$$

But $\text{var}(Y_i) = \sigma_u^2$ (see Chapter 4 (expression 4.2)).

Therefore

$$\text{var}(\hat{b}_0) = \sigma_u^2 \, \Sigma \left[\frac{1}{n^2} - \frac{2\overline{X} k_i}{n} + \overline{X}^2 k_i^2 \right]$$

Since $\Sigma k_i = 0$ and $\Sigma k_i^2 = \dfrac{1}{\Sigma x^2}$, we obtain

$$\text{var}(\hat{b}_0) = \sigma_u^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\Sigma x_i^2} \right] = \sigma_u^2 \left[\frac{\Sigma x_i^2 + n\overline{X}^2}{n \Sigma x_i^2} \right]$$

Now $\Sigma x_i^2 = \Sigma (X_i - \overline{X})^2 = \Sigma X_i^2 - n\overline{X}^2$. Therefore

$$\text{var}(\hat{b}_0) = \sigma_u^2 \frac{\Sigma X_i^2}{n \Sigma x_i^2}$$

Another convenient expression for the variance of $\hat{b}_0$ is

$$\text{var}(\hat{b}_0) = \sigma_u^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\Sigma x_i^2} \right)$$

5.2.6. THE VARIANCE OF THE RANDOM VARIABLE u

The formulae of the variance of $\hat{b}_0$ and $\hat{b}_1$ involve the variance of the random term u, σ_u^2. However, the true variance of u_i cannot be computed since the values of u_i are not observable. But we may obtain an unbiased estimate of σ_u^2 from the expression

$$\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-2}$$

where $e_i = Y_i - \hat{Y}_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$

Proof. We use the device of repeated (hypothetical) sampling, through which we obtain all possible samples of size n, compute a regression line for each sample and find the values of the residuals e_i from each regression. The variance of the residuals ($e_i = Y_i - \hat{Y}_i$) is defined as the expected value of the squared differences of e_i's from their mean, that is:

$$\text{var}(e) = E[e_i - E(e)]^2 = E(e_i^2)$$

since by definition $E(u) = 0$.
The problem is to express this variance in terms of the sample observations of Y and X. We have established in 4.4 that

$$e_i = y_i - \hat{y}_i$$

But $y_i = b_1 x_i + (u_i - \bar{u})$ (where $\bar{u}$ = mean value of u in the particular sample). Because

$$Y_i = b_0 + b_1 X_i + u_i$$

thus

$$\bar{Y} = b_0 + b_1 \bar{X} + \bar{u}$$

and we get, by subtraction,

$$y_i = (Y_i - \bar{Y}) = b_1(X_i - \bar{X}) + (u_i - \bar{u}) = b_1 x_i + (u_i - \bar{u})$$

Note that while $E(u) = 0$ (that is in taking a very large number of samples we expect u to have a mean value of zero by assumption), in any particular single sample $\bar{u}$ is not necessarily zero.
Similarly $\hat{y}_i = \hat{b}_1 x_i$.
Therefore

$$e_i = b_1 x_i + (u_i - \bar{u}) - \hat{b}_1 x_i$$
$$= (u_i - \bar{u}) - (\hat{b}_1 - b_1) x_i$$

The summation over the n sample values of the squares of the residuals yields

$$\Sigma e_i^2 = \Sigma[(u_i - \bar{u}) - (\hat{b}_1 - b_1) x_i]^2$$
$$= \Sigma[(u_i - \bar{u})^2 + (\hat{b}_1 - b_1)^2 x_i^2 - 2(u_i - \bar{u})(\hat{b}_1 - b_1) x_i]$$
$$= \Sigma(u_i - \bar{u})^2 + (\hat{b}_1 - b_1)^2 \Sigma x_i^2 - 2(\hat{b}_1 - b_1) \Sigma x_i(u_i - \bar{u})$$

Taking expected values we have

$$E(\Sigma e^2) = E[\Sigma(u_i - \bar{u})^2] + E[(\hat{b}_1 - b_1)^2 \Sigma x_i^2] - 2E[(\hat{b}_1 - b_1) \Sigma x_i(u_i - \bar{u})]$$

The right-hand side terms may be rearranged as follows:

(a) $$E[\Sigma(u - \bar{u})^2] = E[\Sigma u^2 - \bar{u} \Sigma u_i]$$
$$= E\left[\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right]$$
$$= \Sigma E(u^2) - \frac{1}{n} E[\Sigma u]^2$$
$$= n\sigma_u^2 - \frac{1}{n} E(u_1 + u_2 + \ldots + u_n)^2 \qquad \text{(since } E(u_i^2) = \sigma_u^2\text{)}$$
$$= n\sigma_u^2 - \frac{1}{n} E[\Sigma u_i^2 + 2\sum_{i \neq j} u_i u_j]$$
$$= n\sigma_u^2 - \frac{1}{n}\left[\Sigma E(u_i^2) + 2\sum_{i \neq j} E(u_i u_j)\right] = n\sigma_u^2 - \frac{1}{n} n\sigma_u^2 - \frac{2}{n} \Sigma E(u_i u_j)$$
$$= n\sigma_u^2 - \sigma_u^2 \qquad \text{(given } E(u_i u_j) = 0\text{)}$$
$$= \sigma_u^2 (n - 1).$$

(b) $$E[(\hat{b}_1 - b_1)^2 \Sigma x_i^2] = \Sigma x_i^2 \,.\, E(\hat{b}_1 - b_1)^2$$

given that the X's are fixed in all samples. But

$$E(\hat{b}_1 - b_1)^2 = \text{var}(\hat{b}_1) = \sigma_u^2 \frac{1}{\Sigma x^2}$$

Therefore

$$E[(\hat{b}_1 - b_1)^2 \Sigma x_i^2] = \Sigma x_i^2 \sigma_u^2 \frac{1}{\Sigma x_i^2} = \sigma_u^2$$

(c) $$E[(\hat{b}_1 - b_1) \Sigma x_i(u_i - \bar{u})] = E[(\hat{b}_1 - b_1)(\Sigma x_i u_i - \Sigma x_i \bar{u})].$$

But $\hat{b}_1 = b_1 + \Sigma k_i u_i$, from which $(\hat{b}_1 - b_1) = \Sigma k_i u_i$.

Therefore $$E[(\hat{b}_1 - b_1) \Sigma x_i(u_i - \bar{u})] = E[(\Sigma k_i u_i)(\Sigma x_i u_i - \bar{u} \Sigma x_i)]$$
$$= E[(\Sigma k_i u_i)(\Sigma x_i u_i)] \qquad \text{(given } \Sigma x_i = 0\text{)}$$
$$= E\left[\left(\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)(\Sigma x_i u_i)\right] \qquad \left(\text{since } k = \frac{x_i}{\Sigma x_i^2}\right)$$
$$= E\left[\frac{(\Sigma x_i u_i)^2}{\Sigma x_i^2}\right]$$
$$= E\left[\frac{\Sigma x_i^2 u_i^2 + 2\sum_{i \neq j} (x_i x_j)(u_i u_j)}{\Sigma x_i^2}\right]$$
$$= \left[\frac{\Sigma x_i^2 E(u_i^2)}{\Sigma x_i^2} + \frac{2\sum_{i \neq j} (x_i x_j) E(u_i u_j)}{\Sigma x_i^2}\right]$$
$$= \frac{\Sigma x_i^2 E(u_i^2)}{\Sigma x_i^2} \qquad \text{(given } E(u_i u_j) = 0\text{)}$$
$$= E(u_i^2) = \sigma_u^2.$$

Consequently the expected value of the sum of squares of the residuals becomes by substitution:

$$E(\Sigma e^2) = (n - 1)\sigma_u^2 + \sigma_u^2 - 2\sigma_u^2 = (n - 2)\sigma_u^2,$$

from which we get

$$E\left(\frac{\Sigma e^2}{n - 2}\right) = \sigma_u^2$$

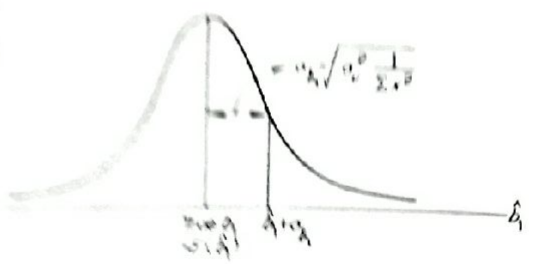
Defining $\hat{\sigma}_u^2 = \Sigma e^2/(n - 2)$ we may write

$$E(\hat{\sigma}_u^2) = \sigma_u^2$$

Thus $\Sigma e^2/(n - 2)$ is an unbiased estimate of the true variance of u.

5.2.7. THE SAMPLING DISTRIBUTION OF THE LEAST SQUARES ESTIMATES

We have found expressions for the mean and variance of the least squares estimates. Given that by Assumption 4 the random variable u is normally distributed, it can be proved that the distribution of the estimates $\hat{b}_0$ and $\hat{b}_1$ is also normal (see R. L. Anderson and T. A. Bancroft, *Statistical Theory in Research*, New York; McGraw-Hill, 1952, pp. 63–4).



TEST OF THE LEAST SQUARES ESTIMATES

... and $\hat{b}_1$ are obtained from a sample of ... Since sampling errors are inevitable in all estimates, ... significance in order to measure the size of the ... confidence in the validity of the estimates. ... purpose. In the present section we will ... namely the *standard error test* which is popular in ... This test helps us to decide whether the estimates ... from zero, i.e. whether the sample from ... might have come from a population whose true ... $b_1 = 0$).[2] Formally we test the null hypothesis

$$H_0 : b_i = 0$$

... hypothesis

$$H_1 : b_i \neq 0$$

... may be outlined as follows. From the formulae of ... established in the preceding section, we compute

... that the standard error test is formally equivalent ... Chapter 8 we deal with the F statistic, which is ... joint tests of significance, i.e. tests of the signifi- ... generalised to allow for any *a priori* value for the

their standard error

$$s(\hat{b}_1) = \sqrt{\operatorname{var}(\hat{b}_1)} = \sqrt{\hat{\sigma}_u^2/\Sigma x^2} = \sqrt{\frac{\Sigma e^2}{(n-2)\Sigma x^2}}$$

$$s(\hat{b}_0) = \sqrt{\operatorname{var}(\hat{b}_0)} = \sqrt{\frac{\hat{\sigma}_u^2 \Sigma X^2}{n\Sigma x^2}} = \sqrt{\frac{(\Sigma e^2)\Sigma X^2}{(n-2)n\Sigma x^2}}$$

We next compare the standard deviations with the numerical values of $\hat{b}_0$ and $\hat{b}_1$. If the standard error is smaller than half the numerical value of the parameter estimate (that is if $s(\hat{b}_i) < (\hat{b}_i/2)$, we conclude that this estimate is statistically significant. This means that we reject the null hypothesis (we reject the hypothesis that the true population parameter $b_i = 0$), which is equivalent to accepting that the true population parameter b_i is different from zero. If, on the other hand, the standard error of the parameter estimate is greater than half its numerical value (that is if $s(\hat{b}_i) > (\hat{b}_i/2)$, we conclude that the least squares estimate is not statistically significant. This means that we accept the null hypothesis that the true parameter $b_1 = 0$. In arriving at the conclusion regarding the significance or nonsignificance of $\hat{b}$ we have been using a two-tail test at the 5 per cent level of significance (see Appendix I).

Economic interpretation of the 'standard-error test'

The procedure outlined above provides a rule of thumb for deciding whether the estimates $\hat{b}_0$ and $\hat{b}_1$ are statistically reliable. The acceptance or rejection of the null hypothesis has a definite economic meaning. Namely the acceptance of the null hypothesis $b_1 = 0$ implies that the explanatory variable to which this estimate relates does not in fact influence the dependent variable Y and should not be included in the function, since the conducted test provided evidence that changes in X leave Y unaffected. In other words acceptance of H_0 implies that the relationship between Y and X is in fact $Y = b_0 + (0)(X) = b_0$, i.e. there is no relationship between Y and X.[1]

Geometric interpretation of the 'standard-error test'

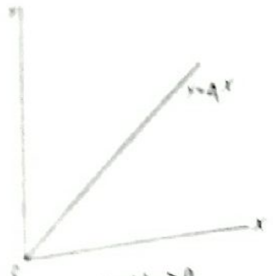
We said that b_0 is the intercept of the regression line on the Y-axis, and b_1 measures the slope of the regression line.

(1) If, when conducting the above test, we find that $s(\hat{b}_0) > \hat{b}_0/2$ and accept the null hypothesis $b_0 = 0$, then the regression line passes through the origin of the axes (figure 5.3), since the relationship between Y and X is actually

$$Y_i = 0 + b_1 X_i = b_1 X_i$$

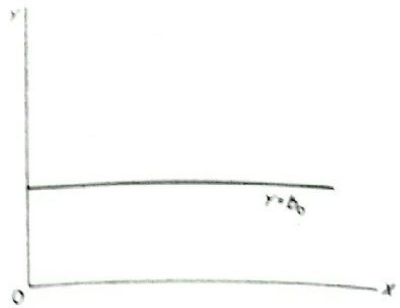
(2) Similarly, if from the test we find $s(\hat{b}_1) > \hat{b}_1/2$, we would accept the null hypothesis that $b_1 = 0$. This would imply that the relationship between Y

[1] Note that in this section we assumed a two-tail test of significance, conducted at the 5 per cent level of significance; that is, we allowed our conclusion to be wrong five times out of one hundred. See Appendix I, p. 563.



$b_0 = 0$ and $b_1 > 0$

Figure 5.3



$b_0 > 0$ and $b_1 = 0$

Figure 5.4

and X is in fact $Y = b_0$. The slope of the regression line would be equal to zero, that is the regression line would in this case be parallel to the X-axis (figure 5.4).

To facilitate the comparison of the standard errors of the estimates to their numerical value it is convenient to print the standard errors in parentheses under the parameter estimates to which they refer.

Example. The standard errors of the coefficients of the supply function, estimated in Chapter 4, are

$$s(\hat{b}_1) = \sqrt{\hat{\sigma}_u^2 / \Sigma x^2} = \sqrt{38.4/48} \approx 0.9$$

$$s(\hat{b}_0) = \sqrt{(\hat{\sigma}_u^2)(\Sigma X^2)/n\Sigma x^2} = \sqrt{(38.4)(1{,}020)/(12)(48)} \approx 8.3$$

(given $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-2) = 383.98/10 \approx 38.4$).

We may present the results of our regression in the compact form

$$\hat{Y}_i = \underset{(8.3)}{33.75} + \underset{(0.9)}{3.25}\, X_i$$

In this form a quick test of the significance of the estimates can be carried out by inspection. Clearly

$$s(\hat{b}_0) < \hat{b}_0/2 \quad \text{and} \quad s(\hat{b}_1) < \hat{b}_1/2$$

Thus both $\hat{b}_0$ and $\hat{b}_1$ are significantly different from zero at the 5 per cent level of significance (in the context of a two-tail test of significance).

We may state the statistical significance of the estimates with one of the following equivalent ways: (1) The estimates are significantly different from zero; or (2) the estimates are statistically significant; or (3) we reject the 'null hypothesis'.

Of course, each of the above statements must be accompanied by the level of significance with which the decision is made (see section 5.2.9, and Appendix I).

5.2.9. THE Z TEST OF THE LEAST-SQUARES ESTIMATES

This test is based on the Standard Normal Distribution (or Gauss Standard Normal Curve) (see Appendix I). It is applicable only if (a) the population

variance is known, or (b) the population variance is unknown, and provided that the sample with which we work is sufficiently large ($n > 30$). If these conditions cannot be fulfilled we apply the student's t test, which is explained in the next section.

In econometric applications the population variance of Y is the variance of u, σ_u^2, which is unknown. However, if we have a large sample ($n > 30$) we may still use the Standard Normal Distribution and perform the Z test (approximately) since the sample estimate of the variance s^2, is a satisfactory approximation to the unknown population variance, σ^2, for large n (see Appendix I).

The Z test may be outlined as follows. We want to test the null hypothesis

$$H_0 : b_i = 0$$

against the alternative hypothesis

$$H_1 : b_i \neq 0$$

We have established that under certain assumptions regarding the values of u (namely $u \sim N(0, \sigma_u^2)$) the least squares estimates $\hat{b}_0$ and $\hat{b}_1$ have the following normal distributions

$$\hat{b}_0 \sim N\left(b_0, \sigma_{(\hat{b}_0)} = \sqrt{\sigma_u^2 \frac{\Sigma X^2}{n \Sigma x^2}}\right)$$

$$\hat{b}_1 \sim N\left(b_1, \sigma_{(\hat{b}_1)} = \sqrt{\sigma_u^2 \frac{1}{\Sigma x^2}}\right)$$

The normal distributions above can be standardised, that is they can be transformed into the units of the standard normal variable Z, which has zero mean and unit variance, $Z \sim N(0, 1)$, through the transformation formula

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

where X_i = the value of the variable which we want to normalise (transform into standard Z units)
μ = the mean of the distribution of the variable
σ = the standard deviation of the variable.

In the case of the distribution of the least squares estimates $\hat{b}_0$ and $\hat{b}_1$, the above transformation formula assumes the form:

$$Z = \frac{\hat{b}_0 - b_0}{\sigma(\hat{b}_0)} = \frac{\hat{b}_0 - b_0}{\sqrt{\sigma_u^2 \Sigma X_i^2 / n \Sigma x_i^2}} \sim N(0, 1) \qquad \text{for } \hat{b}_0$$

$$Z = \frac{\hat{b}_1 - b_1}{\sigma(\hat{b}_1)} = \frac{\hat{b}_1 - b_1}{\sqrt{\sigma_u^2 / \Sigma x_i^2}} \sim N(0, 1) \qquad \text{for } \hat{b}_1$$

With the above transformation formulae we may conduct tests of any hypothesis concerning the true value of the population parameter b. Suppose we want to test the null hypothesis that the true parameter b_1 is equal to a certain value b_1^*. Formally we wish to test the null hypothesis

$$H_0 : b_1 = b_1^*$$

against the alternative hypothesis

$$H_1 : b_1 \neq b_1^*$$

We substitute $b_1 = b_1^*$ into the above formula, and given the estimate $\hat{b}_1$ and its standard error $\sigma_{(\hat{b}_1)}$, we compute the Z^* value

$$Z^* = \frac{\hat{b}_1 - b_1^*}{\sigma_{(\hat{b}_1)}}$$

Given this 'empirical' or 'sample value' or 'observed value' of Z^*, we may calculate (from the Standard Normal distribution table on page 659, the probability of getting the estimate $\hat{b}_1$ if our basic hypothesis ($b_1 = b_1^*$) is true, as follows.

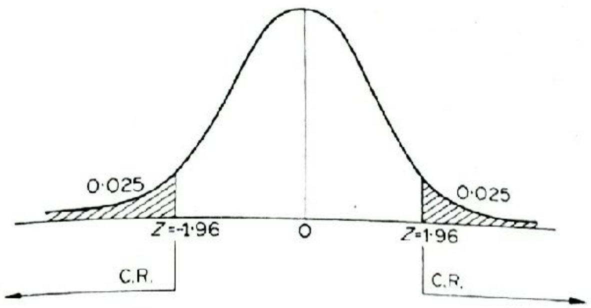
We choose a level of significance[1] for deciding whether to accept or reject our hypothesis. It is customary in econometric research to choose the 5 per cent or the 1 per cent level of significance. This means that in making our decision we allow (tolerate) five times out of a hundred to be 'wrong', that is, to reject the hypothesis when it is actually true.

In applied econometric work it has become customary to perform a two-tail test.[2] That is we choose as our critical region (C.R.) both tails of the Standard Normal distribution, and in particular that part of each tail which corresponds to half the probability of the chosen level of significance. For example, if we choose the 5 per cent level of significance, each tail will include the area (probability) 0.025 (figure 5.5). From the Standard Normal distribution table (on p. 659) we find the critical values of Z, which correspond to the probability 0.025 at each end of the curve ($Z_1 = -1.96$ and $Z_2 = 1.96$). Our final step is to compare the empirical (observed) Z^* with the above critical values of Z.

If the empirical Z^* falls in the critical region, (that is if $Z^* > 1{\cdot}96$ or $Z^* < -1{\cdot}96$) we reject our hypothesis that the true value of b is b^*, because the probability of observing the empirical Z^* (if our hypothesis were true) is very small (smaller than 0·025). Or, to put it in another way, it is improbable that such Z^* would be observed, if our basic hypothesis, H_0, were true. If, on the contrary, the sample value of Z^* falls outside the chosen critical region (that

[1] Level of significance is the probability of making the 'wrong' decision, that is the probability of rejecting the hypothesis when it is actually true or the probability of committing a type I error. See Appendix I.

[2] The choice of a two-tail test implies no *a priori* knowledge regarding the sign of the coefficient whose significance is being tested. However, a one-tail test would be more appropriate in the majority of econometric applications, since economic theory does usually provide us with *a priori* expectations regarding the sign of the coefficients of economic relations.



If the observed value Z^* falls in the shaded area we reject the null hypothesis H_0.

Figure 5.5. A two-tail test at the 5 per cent level of significance.

is $-1{\cdot}96 < Z^* < 1{\cdot}96$) we accept our basic hypothesis, $H_0(b_1 = b_1^*)$, because the probability of observing Z^* (if the hypothesis is true) is large.

For example, suppose $\hat{b} = 29{\cdot}48$, $\sigma_{(\hat{b}_1)} = 36{\cdot}0$ and we want to test the hypothesis $H_0 : b_1 = 25{\cdot}0$. From the Z transformation formula we get

$$Z^* = \frac{\hat{b}_1 - b_1}{\sigma_{(\hat{b}_1)}} = \frac{29{\cdot}48 - 25{\cdot}0}{36{\cdot}0} = 0{\cdot}12$$

Since Z^* does not fall in the critical region ($Z^* < 1{\cdot}96$) we accept our hypothesis that $b = 25{\cdot}0$, because the probability of observing such a value of Z^* is large (larger than 0·05).

In applied econometrics it has become customary to test the hypothesis that the true population parameter is zero. That is, the typical form of the null hypothesis in econometrics is

$$H_0 : b_i = 0$$

and is tested against the alternative hypothesis

$$H_1 : b_i \neq 0$$

The meaning and implications of this hypothesis have been examined in the preceding section. We may summarise the discussion as follows. If we reject the null hypothesis, we say that the empirical coefficient $\hat{b}_1$ is statistically significant, or, it is significantly different from zero. If we accept the null hypothesis, then $\hat{b}_1$ is not significant and there is probably no linear relation between X and Y in the population.

To carry out the test of the above null hypothesis we set $b = 0$ in the Z transformation formula

$$Z^* = \frac{\hat{b} - b}{\sigma_{(\hat{b})}} = \frac{\hat{b} - 0}{\sigma_{(\hat{b})}} = \frac{\hat{b}}{\sigma_{(\hat{b})}}$$

Thus in the case of the test for the null hypothesis $H_0 : b_1 = 0$ the procedure of the Z test reduces to the simple step of dividing the estimated value of the parameter (b_1) by its standard deviation and then comparing the resulting Z^*

value with the theoretical (tabular) values of Z, which define the critical region of our test. The theoretical values of Z are obtained from the Standard Normal curve table (p. 659).

Given that for the 5 per cent level of significance (or the 95 per cent confidence level) the critical value of Z is 1·96, we can take this critical value as approximately equal to 2·0, and perform the rough test which was outlined in the previous section, and can now be explained in some detail. We said there that if $\sigma_{(\hat{b}_i)} > b_i/2$ we reject the null hypothesis. From the preceding discussion we concluded that if $Z^* > 2$ we reject the null hypothesis. These two statements are identical, because from the formula $Z^* = \hat{b}_i/\sigma_{(\hat{b}_i)}$ it is obvious that Z^* can be greater than 2 only if the numerator $\hat{b}_i$ is at least twice the value of the denominator, that is only if $\hat{b} > 2\sigma_{(\hat{b})}$, or $\hat{b}_i/2 > \sigma_{(\hat{b}_1)}$. Thus the statements:

(a) we reject the null hypothesis if $Z^* > 2$; and

(b) we reject the null hypothesis if $\sigma_{(\hat{b})} < \hat{b}/2$

are two different ways of saying the same thing. We stress that these statements assume a two-tail test conducted at the 5 per cent level of significance.

Example Suppose that we have estimated the following supply function from a sample of 700 observations ($n = 700$)

$$\underset{(20)}{Y = 100} + \underset{(1\cdot 5)}{4\cdot 00}\, X$$

We will conduct the Z-test for the slope estimate $\hat{b}_1 = 4$, given its standard error 1·5.

Since the sample is large, the estimated standard deviation of the parameters is a good approximation of the true standard deviation of these parameters. Therefore we may apply the Z test for finding the statistical significance of the estimates $\hat{b}_0$ and $\hat{b}_1$ (see Appendix I):

Null Hypothesis: $b_1 = 0$

Alternative Hypothesis: $b_1 \neq 0$

Computing the Z^* value, we find

$$Z^* = \frac{\hat{b}_1}{\sigma(\hat{b}_1)} = \frac{4}{1\cdot 5} = 2\cdot 66$$

Since the theoretical (tabular) value of Z (at the 5 per cent level of significance) is 1·96, $Z^* > Z$.

On the evidence of our sample we conclude that it is highly improbable that the true slope b_1 is equal to zero. Our regression estimate is statistically significant.

5.2.10. THE STUDENT'S t TEST

We said that the Z test can be applied in the following cases only. Firstly, if the true population variance is known, irrespective of the sample size. Secondly, if the true variance of the estimates is unknown, provided the size of the sample is sufficiently large ($n > 30$), because in this case the sample estimate of the variance is a satisfactory approximation of the unknown population variance.[1]

[1] See T. Yamane, *Statistics*, 2nd ed., Harper & Row, Japan 1967, pp. 514–16.

In econometric applications the true variances of the estimates, $\sigma^2_{\hat{b}_0}$ and $\sigma^2_{\hat{b}_1}$, are unknown, because they involve the true variance of the random term, σ^2_u, which of course is unknown. We may, however, use the unbiased estimate $\hat{\sigma}^2_u = \Sigma e^2/(n-K)$ and obtain estimates of the variances of the coefficients, $s^2_{\hat{b}_0}$ and $s^2_{\hat{b}_1}$. If the sample is sufficiently large ($n > 30$) these estimates are adequate for the application of the Z transformation. However, in practice the sample is rarely sufficiently large. When the sample is small ($n < 30$) *and provided that the population of the parameters is normal*, we can apply another transformation, based on the Student's t distribution. (See Appendix I.)

The general formula which transforms the values of any variable X into t units is similar to the Z transformation, but the t value depends in addition on the number of degrees of freedom and it includes the variance estimates s^2_X instead of the true variance. (See Appendix I.) In the formation of the t statistic the true variance σ^2_X is eliminated and we are left with a formula which includes its unbiased estimate $\hat{\sigma}^2_X$. (See Yamane, *Statistics*, pp. 517–19.) The t transformation formula (t statistic) is

$$t = \frac{X_i - \mu}{s_X} \qquad \text{with } n-1 \text{ degrees of freedom}$$

where μ = value of the population mean

s^2_X = sample estimate of the population variance

$$s^2_X = \Sigma(X_i - \bar{X})^2/(n-1)$$

n = sample size.

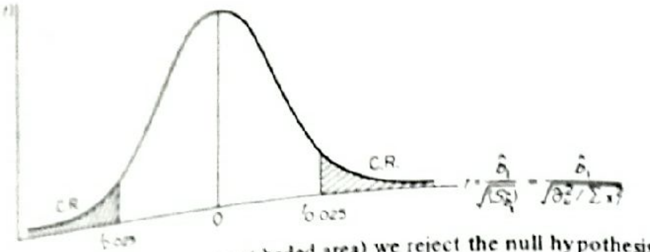
The sampling distribution in this case, that is the distribution of the sample means, is $\bar{X} \sim N(\mu, s^2_{\bar{X}})$ and the transformation statistic is $(\bar{X} - \mu)/\sqrt{s^2_X/n}$, and has a t distribution with $(n-1)$ degrees of freedom.

The t distribution is always symmetric, with mean equal to zero and variance $(n-1)/(n-3)$, which approaches unity when n is large. Clearly as n increases, the t distribution approaches the Standard Normal distribution $Z \sim N(0, 1)$.

The probabilities of the t distribution have been tabulated by W. S. Gosset, who wrote under the pseudonym *Student* which gave the name to the t distribution. The t distribution is reproduced in table 2 of Appendix IV (p. 660).

To perform a two-tailed test we must (a) define the null and alternative hypotheses, (b) choose the desired level of significance (5 per cent or 1 per cent customarily), (c) define the number of degrees of freedom. With this information we can define the critical region, that is the critical values of t which divide the total set of values of t in two regions, the acceptance and the rejection regions. We can define the acceptance region for t as follows. Assume we want to test the null hypothesis

$$H_0: \mu = \mu_0$$



If the observed t^* falls in the critical region (shaded area) we reject the null hypothesis H_0.

Figure 5.7. A two-tail t test of the null hypothesis at the 5 per cent level of significance.

The t test can be performed in an approximate way by simple inspection. From the t table we see that the value of t changes very slowly when the degrees of freedom $(n-K)$ are more than 8. For example $t_{0.025}$ takes values between 2·30 (when $n-K=8$) and 1·96 (when $n-K=\infty$). The change from 2·30 to 1·96 is obviously very slow. Consequently we can ignore the degrees of freedom (when $n-K>8$) and say that the critical value of $t_{0.025}$ is 2. Thus the two-tail test of the null hypothesis (at 5 per cent level of significance) reduces to the following rule.

If the observed t^* is greater than 2 (or smaller than −2), we reject the null hypothesis.

If, on the other hand, the observed t^* is smaller than 2 (but greater than −2), we accept the null hypothesis.

Given that $t^* = \frac{\hat{b}_i}{s_{(\hat{b}_i)}}$, the sample value of t^* would be greater than 2 if the relevant estimate ($\hat{b}_0$ or $\hat{b}_1$) is at least twice its standard deviation. In other words

$$t^* > 2 \quad \text{if} \quad \hat{b}_i > 2s_{(\hat{b}_i)} \quad \text{or} \quad s_{(b_i)} < \hat{b}_i/2$$

Thus we see that the statements: (a) we reject the null hypothesis if $t^* > t_{0.025}$, and (b) we reject the null hypothesis is $s_{(\hat{b})} < \hat{b}/2$ are essentially the same. We repeat that this is an approximation to the formal t test and is valid only for $(n-K) > 8$.

Although the two-tail test is traditionally applied in testing the regression coefficients, a one-tail test would be appropriate in the majority of cases since economic theory provides us with *a priori* expectations regarding the sign of coefficients of economic relationships.

Example 1. Suppose that from a sample of size $n = 20$ we estimate the following consumption function

$$\hat{C} = 100 + 0{\cdot}70\,Y$$
$$(75{\cdot}5)\quad(0{\cdot}21)$$

The figures in brackets are the standard errors of the coefficients $\hat{b}_0 = 100$ and $\hat{b}_1 = 0{\cdot}70$. Since $n < 30$ we cannot apply the Z test. However, given, the stochastic assumptions about the values of u, the estimates are normally distributed, and hence we may apply the t test. For b_1 we have

$$t^* = \frac{\hat{b}_1}{\hat{\sigma}_{(\hat{b}_1)}} = \frac{0{\cdot}70}{0{\cdot}21} \approx 3{\cdot}3$$

We wish to test the hypothesis

$$H_0 : \hat{b}_1 = 0$$

against the alternative hypothesis

$$H_1 : \hat{b}_1 \neq 0$$

The critical values of t for $(n-K) = 18$ degrees of freedom are

$$t_1 = -t_{0.025} = -2{\cdot}10 \quad \text{and} \quad t_2 = +t_{0.025} = +2{\cdot}10$$

The relevant critical region is shown in figure 5.8.

Since $t^* > t_{0.025}$ we reject the null hypothesis and conclude that $\hat{b}_1$ is different from zero.

Example 2. The standard errors and the t values for the coefficients of the supply function estimated in Chapter 4 are given below. The regression is $\hat{Y} = 33{\cdot}75 + 3{\cdot}25X$.

(a) The standard errors are

$$s_{\hat{b}_0} = \sqrt{\hat{\sigma}_u^2 \frac{\Sigma X_i^2}{n\,\Sigma x_i^2}} = 8{\cdot}28 \quad \text{and} \quad s_{\hat{b}_1} = \sqrt{\hat{\sigma}_u^2 \frac{1}{\Sigma x_i^2}} = 0{\cdot}89$$

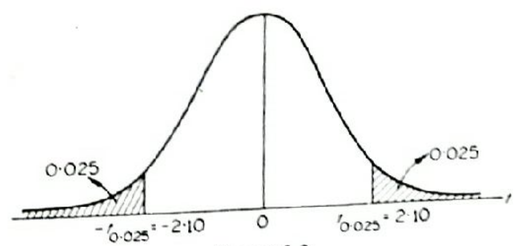


Figure 5.8

(b) The t values for the two parameter estimates are

$$t_{(\hat{b}_0)} = \frac{\hat{b}_0}{s_{\hat{b}_0}} = 4{\cdot}07$$

$$t_{(\hat{b}_1)} = \frac{\hat{b}_1}{s_{\hat{b}_1}} = 3{\cdot}62$$

Clearly both estimates are statistically significant.

5.3. CONFIDENCE INTERVALS FOR b_0 AND b_1

Rejection of the null hypothesis does not mean that our estimate $\hat{b}_i$ is 'the' correct estimate of the true population parameter b_i. It simply means that our estimate comes from a sample drawn from a population whose parameter b_i is different from zero.

In order to define how close to the estimate the true parameter lies, we must construct confidence intervals for the true parameter, in other words we must establish limiting values around the estimate within which the true parameter is expected to lie with a certain 'degree of confidence'. In this respect we say that with a given probability the population parameter will be within the defined *confidence interval* or *confidence limits*.

We choose a probability in advance and refer to it as *the confidence level* (or confidence coefficient). It is customary in econometrics to choose the 95 per cent confidence level. This means that in repeated sampling the confidence limits, computed from the sample, would include the true population parameter in 95 per cent of the cases. In the other 5 per cent of the cases the population parameter will fall outside the confidence limits.

5.3.1 CONFIDENCE INTERVAL FROM THE STANDARD NORMAL DISTRIBUTION

It has already been mentioned that the Z distribution may be employed either if we know the true standard deviation $\sigma_{(\hat{b}_i)}$, or when we have a large sample ($n > 30$) because, for large samples, the sample standard deviation, s, is a reasonably good estimate of the unknown population standard deviation.

The Z statistic for $\hat{b}_i$ is

$$Z = \frac{\hat{b}_i - b_i}{\sigma_{(\hat{b}_i)}}$$

Our first task is to choose a confidence coefficient, say 95 per cent. We next look at the standard normal table and find that the probability of the value of Z lying between -1.96 and 1.96 is 0.95. This may be written as follows

$$P\{-1.96 < Z < +1.96\} = 0.95$$

Substituting $Z = (\hat{b}_i - b_i)/\sigma_{(\hat{b}_i)}$ and rearranging slightly, we get

$$P\left\{-1.96 < \frac{\hat{b}_i - b_i}{\sigma_{(\hat{b}_i)}} < 1.96\right\} = 0.95$$

$$P\{\hat{b}_i - 1.96(\sigma_{\hat{b}_i}) < b_i < \hat{b}_i + 1.96(\sigma_{\hat{b}_i})\} = 0.95$$

Thus the 95 per cent confidence interval for b_i is

$$\boxed{\hat{b}_i - 1.96(\sigma_{\hat{b}_i}) < b_i < \hat{b}_i + 1.96(\sigma_{\hat{b}_i})}$$

or

$$\boxed{b_i = \hat{b}_i \pm (1.96) \cdot (\sigma_{\hat{b}_i})}$$

The meaning of the confidence interval is that the unknown population parameter, b_i, will lie within the defined limits 95 times out of 100.

For example, if $\hat{b} = 8.4$ and $\sigma_{(\hat{b})} = 2.2$, choosing a value of 95 per cent for the confidence coefficient, we find the confidence interval

$$b = 8.4 \pm 1.96\,(2.2)$$

or

$$8.4 - 1.96\,(2.2) < b < 8.4 + 1.96\,(2.2)$$

and

$$4.1 < b < 12.7$$

Thus from our single sample estimate we infer that the (unknown) true population parameter will lie between 4.1 and 12.7, with a probability of 95 per cent.

5.3.2. CONFIDENCE INTERVAL FROM THE STUDENT'S t DISTRIBUTION

The procedure for constructing a confidence interval with the t distribution is similar to the one outlined earlier with the main difference that in this case we must take into account the degrees of freedom.

The t statistic for $\hat{b}_i$ is

$$t = \frac{\hat{b}_i - b_i}{s_{(\hat{b}_i)}} \qquad \text{with } (n - K) \text{ degrees of freedom}$$

We first choose the 95 per cent confidence level (or any other confidence level) and we find from the t table the value of $\pm t_{0.025}$ with $(n - K)$ degrees of freedom. This implies that the probability of t lying between $-t_{0.025}$ and $+t_{0.025}$ is 0.95 (with $n - K$ degrees of freedom). Consequently we may write

$$P\{-t_{0.025} < t < +t_{0.025}\} = 0.95$$

Substituting $t = (\hat{b}_i - b_i)/s_{(\hat{b}_i)}$ in the above expression, we find

$$P\left\{-t_{0.025} < \frac{\hat{b}_i - b_i}{s_{(\hat{b}_i)}} < +t_{0.025}\right\} = 0.95$$

$$P\{\hat{b}_i - t_{0.025}(s_{\hat{b}_i}) < b_i < \hat{b}_i + t_{0.025}(s_{\hat{b}_i})\} = 0.95$$

Thus the 95 per cent confidence interval for b, when we use a small sample for its estimation, is

$$\boxed{\hat{b}_i - t_{0.025}(s_{\hat{b}_i}) < b_i < \hat{b}_i + t_{0.025}(s_{\hat{b}_i})} \qquad \text{with } (n - K) \text{ degrees of freedom}$$

or

$$\boxed{b_i = \hat{b}_i \pm t_{0.025}(s_{\hat{b}_i})} \qquad \text{with } (n - K) \text{ degrees of freedom}$$

The meaning of the 95 per cent confidence interval is that there is a 0.95 probability of including the true value of the population parameter in the interval $\hat{b} \pm t_{0.025}$ (with $n - K$ degrees of freedom). For example, suppose we have estimated the following regression line from a sample of 20 observations.

$$\hat{Y} = 128.5 + 2.88X \qquad (n - K) = 20 - 2 = 18.$$
$$(38.2) \quad (0.85)$$

We next find the observed value of the Z statistic

$$Z^* = \frac{\omega_r - \omega_\rho}{\sigma_\omega} = \frac{(0.9730) - (0.6932)}{0.2} = 1.28$$

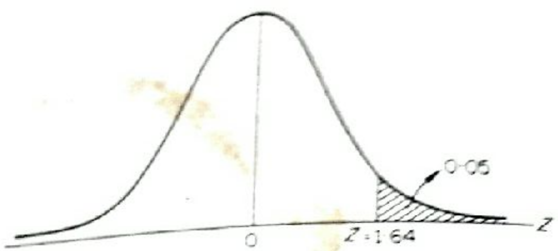


Figure 5.10

We will choose for our test the upper tail of the standard normal distribution, since the alternative hypothesis is $H_1 : \rho > 0.60$. (See Appendix I.) From the table of the Standard Normal curve we find that $Z = 1.64$ (at 5 per cent level of significance (figure 5.10)). Since $Z^* < Z$ we *cannot reject* the hypothesis that the population ρ is 0.60: We accept H_0. This implies that our estimate of r is not significantly different from 0.60, at the 5 per cent level of significance.

For a test of the R^2 in multiple regression see Chapter 8.

5.4.3. TEST OF SIGNIFICANCE OF THE RANK CORRELATION COEFFICIENT r'

The statistical significance of Spearman's rank correlation coefficient can be tested by the following procedure.

If the population ρ is zero, the distribution of r' can be approximated with a normal curve having the mean 0 and the standard deviation $1/\sqrt{n-1}$, that is

$$r' \sim N\left(0, \sigma_{r'} = \frac{1}{\sqrt{n-1}}\right)$$

The null and alternative hypotheses are

$$H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \neq 0$$

Given the form of the alternative hypothesis, we apply a two-tail test. The Z statistic can be used in this test provided that the sample is large. We estimate

$$Z^* = \frac{r'}{\sigma_{r'}} = r'\sqrt{n-1}$$

We next compare Z^* with the tabular values of $Z = \pm 1.96$, which define the critical region of a two tail test at the 5 per cent level of significance. If $-1.96 < Z^* < 1.96$, we accept H_0. Otherwise we reject H_0. This procedure is equivalent to the rule:

(a) We reject H_0 if $r' < \frac{-1.96}{\sqrt{n-1}}$, or if $r' > \frac{1.96}{\sqrt{n-1}}$

(b) We accept H_0 if $\frac{-1.96}{\sqrt{n-1}} < r' < \frac{1.96}{\sqrt{n-1}}$

5.5 A NOTE ON THE IMPORTANCE OF THE STATISTICAL TESTS OF SIGNIFICANCE

There is no general agreement among econometricians as to which of the two statistical criteria is more important: a high r^2, or low standard errors of the estimates.

Statistical criteria acquire great importance when one follows the experimental approach in investigating any particular problem. We said that in this approach the research takes the form of a process of computing various models with various combinations of the relevant variables, and then trying to decide which is preferable. The choice would not be difficult if one of the models produced a higher r^2 and lower standard errors. However, this is not usually the case. In most applications we obtain a high r^2, while some parameters have high standard errors. In this event some econometricians tend to attribute great importance to r^2, and to accept the parameter estimates, despite the fact that some of them are statistically insignificant. Others suggest that acceptance or rejection of the estimates which are not statistically significant depends on the aim of the model in any particular situation.

The majority of writers seem to agree that r^2 is a more important criterion when the model is to be used for forecasting, while the standard errors acquire a greater importance when the purpose of the research is the explanation (analysis) of economic phenomena and the estimation of reliable values of particular economic parameters.

A high r^2 has a clear merit only when combined with significant estimates (low standard errors). When high r^2 and low standard errors are not found contemporaneously in any particular study the researcher should be very careful in his interpretation and acceptance of the results. Priority should always be given to the fulfilment of the economic *a priori* criteria (sign and size of the estimates). Only when the economic criteria are satisfied should one proceed with the application of the first-order and second-order tests of significance.

5.6 SUMMARY OF THE ESTIMATION PROCEDURE OF OLS APPLIED TO THE TWO-VARIABLE MODEL

The estimation procedure of OLS may be expressed in a sequence of five steps, which greatly simplify the computations involved.