1 : CHAPTER

Elementary Regression Analysis

In this chapter, we shall study three methods which are capable of generating estimates of statistical parameters in a wide variety of contexts. These are the method of moments, the method of least squares and the principle of maximum likelihood.

We shall study the methods only in relation to the simple linear regression model; and we shall show that each entails assumptions which may be more or less appropriate to the context in which we wish to apply the model.

In the case of the regression model, the three methods generate estimating equations which are formally identical; but this does not justify us in taking a casual approach to the statistical assumptions which sustain the model. To be casual in making our assumptions is to invite the danger of misinterpretation when the results of the estimation are in hand.

We begin with the method of moments, we shall proceed to the method of least squares, and we shall conclude with a brief treatment of the method of maximum likelihood.

Conditional Expectations

Let y be a continuously distributed random variable whose probability density function is $f(y)$. If we wish to predict the value of y without the help of any other information, then we might take its expected value which is defined by

$$E(y) = \int yf(y)dy.$$

The expected value is a so-called minimum-mean-square-error (m.m.s.e.) predictor. If π is the value of a prediction, then the mean-square error is given by

$$\begin{aligned} M &= \int (y-\pi)^2 f(y)dy \\ &= E\big\{(y-\pi)^2\big\} \\ &= E(y^2) - 2\pi E(y) + \pi^2; \end{aligned} \tag{1}$$

and, using the methods of calculus, it is easy to show that this quantity is minimised by taking $\pi = E(y)$.

Now let us imagine that y is statistically related to another random variable x whose value we have already observed. For the sake of argument, let us assume that we know the form of the joint distribution of x and y which is $f(x, y)$. Then the minimum-mean-square-error prediction of y is given by the conditional expectation

$$E(y|x) = \int y \frac{f(x,y)}{f(x)} dy \tag{2}$$

wherein

$$f(x) = \int f(x, y)dy \tag{3}$$

is the so-called marginal distribution of x. We may state this proposition formally in a way which will assist us in proving it:

(4) Let $\hat{y} = \hat{y}(x)$ be the conditional expectation of y given x which is also expressed as $\hat{y} = E(y|x)$. Then we have $E\{(y - \hat{y})^2\} \leq E\{(y - \pi)^2\}$, where $\pi = \pi(x)$ is any other function of x.

Proof. Consider

$$\begin{aligned} E\big\{(y-\pi)^2\big\} &= E\Big[\big\{(y-\hat{y}) + (\hat{y}-\pi)\big\}^2\Big] \\ &= E\big\{(y-\hat{y})^2\big\} + 2E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} + E\big\{(\hat{y}-\pi)^2\big\}. \end{aligned} \tag{5}$$

In the second term, there is

$$\begin{aligned} E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} &= \int_x \int_y (y-\hat{y})(\hat{y}-\pi) f(x,y)\partial y \partial x \\ &= \int_x \left\{ \int_y (y-\hat{y}) f(y|x)\partial y \right\} (\hat{y}-\pi) f(x)\partial x \\ &= 0. \end{aligned} \tag{6}$$

Here the second equality depends upon the factorisation $f(x, y) = f(y|x)f(x)$ which expresses the joint probability density function of x and y as the product of the conditional density function of y given x and the marginal density function of x. The final equality depends upon the fact that $\int (y - \hat{y})f(y|x)\partial y = E(y|x) - E(y|x) = 0$. Therefore $E\{(y - \pi)^2\} = E\{(y - \hat{y})^2\} + E\{(\hat{y} - \pi)^2\} \geq E\{(y - \hat{y})^2\}$, and the assertion is proved.

We might note that the definition of the conditional expectation implies that

$$\begin{aligned} E(xy) &= \int_x \int_y xyf(x,y)\partial y \partial x \\ &= \int_x x\Big\{\int_y yf(y|x)\partial y\Big\}f(x)\partial x \\ &= E(x\hat{y}). \end{aligned} \tag{7}$$

When the equation $E(xy) = E(x\hat{y})$ is rewritten as

$$E\big\{x(y-\hat{y})\big\} = 0, \tag{8}$$

it may be described as an orthogonality condition. This condition indicates that the prediction error $y - \hat{y}$ is uncorrelated with x. The result is intuitively appealing; for, if the error were correlated with x, then we should not be using the information of x efficiently in forming $\hat{y}$.

If the joint distribution of x and y is a normal distribution, then we can make rapid headway in finding an expression for the function $E(y|x)$. In the case of a normal distribution, we have

$$E(y|x) = \alpha + \beta x, \tag{9}$$

which is to say that the conditional expectation of y given x is a linear function of x. Equation (9) is described as a linear regression equation; and we shall explain this terminology later.

The object is to find expressions for α and β which are in terms of the first-order and second-order moments of the joint distribution. That is to say, we wish to express α and β in terms of the expectations $E(x)$, $E(y)$, the variances $V(x)$, $V(y)$ and the covariance $C(x,y)$.

Admittedly, if we had already pursued the theory of the Normal distribution to the extent of demonstrating that the regression equation is a linear equation, then we should have already discovered these expressions for α and β. However, our present purposes are best served by taking equation (9) as our starting point; and we are prepared to regard the linearity of the regression equation as an assumption in its own right rather than as a deduction from the assumption of a normal distribution.

Let us begin by multiplying equation (9) throughout by $f(x)$, and let us proceed to integrate with respect to x. This gives us the equation

$$E(y) = \alpha + \beta E(x), \tag{10}$$

whence

$$\alpha = E(y) - \beta E(x). \tag{11}$$

Equation (10) shows that the regression line passes through the point $E(x, y) = \{E(x), E(y)\}$ which is the expected value of the joint distribution.

By putting (11) into (9), we find that

$$E(y|x) = E(y) + \beta\big\{x - E(x)\big\}, \tag{12}$$

which shows how the conditional expectation of y differs from the unconditional expectation in proportion to the error of predicting x by taking its expected value.

Now let us multiply (9) by x and $f(x)$ and then integrate with respect to x to provide

$$E(xy) = \alpha E(x) + \beta E(x^2). \tag{13}$$

Multiplying (10) by $E(x)$ gives

$$E(x)E(y) = \alpha E(x) + \beta\big\{E(x)\big\}^2, \tag{14}$$

whence, on taking (14) from (13), we get

$$E(xy) - E(x)E(y) = \beta\Big[E(x^2) - \big\{E(x)\big\}^2\Big], \tag{15}$$

which implies that

$$\begin{aligned}
\beta &= \frac{E(xy) - E(x)E(y)}{E(x^2) - \big\{E(x)\big\}^2} \\
&= \frac{E\Big[\big\{x - E(x)\big\}\big\{y - E(y)\big\}\Big]}{E\Big[\big\{x - E(x)\big\}^2\Big]} \\
&= \frac{C(x, y)}{V(x)}.
\end{aligned} \tag{16}$$

Thus we have expressed α and β in terms of the moments $E(x)$, $E(y)$, $V(x)$ and $C(x, y)$ of the joint distribution of x and y.

Example. Let $x = \xi + \eta$ be an observed random variable which combines a signal component ξ and a noise component η. Imagine that the two components are uncorrelated with $C(\xi, \eta) = 0$, and let $V(\xi) = \sigma_\xi^2$ and $V(\eta) = \sigma_\eta^2$. The object is to extract the signal from the observation.

According to the formulae of (12) and (16), the expectation of the signal conditional upon the observation is

$$E(\xi|x) = E(\xi) + \frac{C(x,\xi)}{V(x)}\big\{x - E(x)\big\}. \tag{17}$$

Given that ξ and η are uncorrelated, it follows that

$$V(x) = V(\xi + \eta) = \sigma_\xi^2 + \sigma_\eta^2 \tag{18}$$

and that

$$C(x,\xi) = V(\xi) + C(\xi,\eta) = \sigma_\xi^2. \tag{19}$$

Therefore

$$E(\xi|x) = E(\xi) + \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\eta^2}\big\{x - E(x)\big\}. \tag{20}$$

Estimation by the Method of Moments

It is most unlikely that we should know the values of the various moments comprised in the formulae for the regression parameters. Nevertheless, we are often able to estimate them. Imagine that we have a sample of T observations on x and y : $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$. Then we can calculate the following empirical or sample moments:

$$\begin{aligned}
\bar{x} &= \frac{1}{T}\sum_{t=1}^{T} x_t, \\
\bar{y} &= \frac{1}{T}\sum_{t=1}^{T} y_t, \\
s_x^2 &= \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 = \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \bar{x}^2, \\
s_{xy} &= \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y}.
\end{aligned} \tag{21}$$

The method of moments suggests that, in order to estimate α and β, we should replace the moments in the formulae of (11) and (16) by the corresponding sample moments. Thus the estimates of α and β are

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \\
\hat{\beta} &= \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}.
\end{aligned} \tag{22}$$

The justification of the method is that, in many of the circumstances under which the sample is liable to be generated, we can expect the sample moments to converge to the true moments of the bivariate distribution, thereby causing the estimates of the parameters to converge likewise to the true values.

We should be precise about the meaning of convergence in this context. According to the concept of convergence which is used in mathematical analysis,

(23) A sequence of numbers $\{a_n\}$ is said to converge to a limit a if, for any arbitrarily small real number ϵ, there exists a corresponding integer N such that $|a_n - a| < \epsilon$ for all $n \geq N$.

This concept is not appropriate to the case of a stochastic sequence, such as a sequence of estimates. For, no matter how many observations N have been incorporated in the estimate a_N, there remains a possibility that, subsequently, an aberrant observation y_n will draw the estimate a_n beyond the bounds of $a \pm \epsilon$. We must adopt a criterion of convergence which allows for this possibility:

(24) A sequence of random variables $\{a_n\}$ is said to converge weakly in probability to a limit a if, for any ϵ, we have $\lim P(|a_n - a| > \epsilon) = 0$ as $n \to \infty$ or, equivalently, $\lim P(|a_n - a| \leq \epsilon) = 1$.

This means that, by increasing the size of the sample, we can make it virtually certain that a_n will 'fall within an epsilon of a.' It is conventional to describe a as the probability limit of a_n and to write $\text{plim}(a_n) = a$.

The virtue of this definition of convergence is that it does not presuppose that the random variable a_n has a finite variance or even a finite mean. However, if a_n does have finite moments, then we may use the concept of mean-square convergence.

(25) A sequence of random variables $\{a_n\}$ is said to converge in mean square to a limit a if $\lim(n \to \infty)E\{(a_n - a)^2\} = 0$.

We should note that

$$\begin{aligned} E\Big\{\big(a_n - a\big)^2\Big\} &= E\Big\{\Big(\big[a_n - E(a_n)\big] - \big[a - E(a_n)\big]\Big)^2\Big\} \\ &= V(a_n) + E\Big[\big\{a - E(a_n)\big\}^2\Big]; \end{aligned} \tag{26}$$

which is to say that the mean-square error of a_n is the sum of its variance and the square of its bias. If a_n is to converge in mean square to a, then both of these quantities must vanish.

Convergence in mean square is a stronger condition than convergence in probability in the sense that it implies the latter. Whenever an estimator

converges in probability to the value of the parameter which it purports to represent, then we say that it is a consistent estimator.

Regression and the Eugenic Movement

The theory of linear regression has its origins in the late 19th century when it was closely associated with the name of the English eugenicist Francis Galton (1822–1911).

Galton was concerned with the hereditibility of physical and mental characteristics; and he sought ways of improving the genetic quality of the human race. His disciple Karl Pearson, who espoused the same eugenic principles as Galton and who was a leading figure in the early development of statistical theory in Britain, placed Galton's contributions to science on a par with those of Charles Darwin who was Galton's cousin.

Since the 1930's, the science of eugenics has fallen into universal disrepute, and its close historical association with statistics has been largely forgotten. However it should be recalled that one of the premier journals of statistical theory, which now calls itself *Biometrika*, began life as *The Annals of Eugenics*. The thoughts which inspired the Eugenic Movement still arise, albeit that they are expressed, nowadays, in different guises.

One of Galton's studies which is best remembered concerns the relationship between the heights of fathers and the heights of their sons. The data which was gathered was plotted on a graph and was found to have a distribution which resembles a bivariate normal distribution.

It might be supposed that the best way to predict the height of a son is to take the height of the father. In fact, such a method would lead of a systematic over-estimation of the height of the sons if their fathers were above-average height. In the terminology of Galton, we actually witness a regression of the son's height towards "mediocrity ".

Galton's terminology suggests a somewhat unbalanced point of view. The phenomenon of regression is accompanied by a corresponding phenomenon of progression whereby fathers of less than average height are liable to have sons who are taller than themselves. Also, if the distribution of heights is to remain roughly the same from generation to generation and if it is not to loose its dispersion, then there are bound to be cases which conflict with the expectation of an overall reversion towards the mean.

A little reflection will go a long way toward explaining the phenomenon of reversion; for we need only consider the influence of the mother's height. If we imagine that, in general, men of above-average height show no marked tendency to marry tall women, then we might be prepared to attribute an average height to the mother, regardless of the father's height. If we acknowledge that the two parents are equally influential in determining the physical characteristics of their offspring, then we have a ready explanation of the tendency of heights

to revert to the mean. To the extent that tall people choose tall partners, we shall see a retardation of the tendency; and the characteristics of abnormal height will endure through a greater number of generations.

The Bivariate Normal Distribution

Most of the results in the theory of regression which we have developed so far can be obtained by examining the functional form of the bivariate normal distribution. Let x and y be the two variables. Let us denote their means by

$$E(x) = \mu_x, \qquad E(y) = \mu_y, \tag{27}$$

their variances by

$$V(x) = \sigma_x^2, \qquad V(y) = \sigma_y^2 \tag{28}$$

and their covariance by

$$C(x, y) = \rho\sigma_x\sigma_y. \tag{29}$$

Here

$$\rho = \frac{C(x, y)}{\sqrt{V(x)V(y)}}, \tag{30}$$

which is called the correlation coefficient of x and y, provides a measure of the relatedness of these variables.

The Cauchy–Schwarz inequality indicates that $-1 \leq \rho \leq 1$. If $\rho = 1$, then there is an exact positive linear relationship between the variables whereas, if $\rho = -1$, then there is an exact negative linear relationship, Neither of these extreme cases is admissible in the present context for, as we may see by examining the following formulae, they lead to the collapse of the bivariate distribution.

The bivariate distribution is specified by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp Q(x, y), \tag{31}$$

where

$$Q = \frac{-1}{2(1-\rho^2)} \left\{ \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \right\} \tag{32}$$

is a quadratic function of x and y.

The function can also be written as

$$Q = \frac{-1}{2(1-\rho^2)}\left\{\left(\frac{y-\mu_y}{\sigma_y} - \rho\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right\}. \tag{33}$$

Thus we have

$$f(x,y) = f(y|x)f(x), \tag{34}$$

where

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}, \tag{35}$$

and

$$f(y|x) = \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}}\exp\left\{-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho)^2}\right\}, \tag{36}$$

with

$$\mu_{y|x} = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x). \tag{37}$$

Equation (37) is the linear regression equation which specifies the value of $E(y|x) = \mu_{y|x}$ in terms of x; and it is simply the equation (12) in another notation. Equation (36) indicates that the variance of y about its conditional expectation is

$$V(y|x) = \sigma_y^2(1-\rho^2). \tag{38}$$

Since $(1-\rho^2) \leq 1$, it follows that variance of the conditional predictor $E(y|x)$ is less than that of the unconditional predictor $E(y)$ whenever $\rho \neq 0$—which is whenever there is a correlation between x and y. Moreover, as this correlation increases, the variance of the conditional predictor diminishes.

There is, of course, a perfect symmetry between the arguments x and y in the bivariate distribution. Thus, if we choose to factorise the joint probability density function as $f(x,y) = f(x|y)f(y)$, then, to obtain the relevant results, we need only interchange the x's and the y's in the formulae above.

We should take note of the fact that x and y will be statistically independent random variables that are uncorrelated with each other if and only if their joint distribution can be factorised as the product of their marginal distributions: $f(y,y) = f(y)f(y)$. In the absence of statistical independence, the joint distribution becomes the product of a conditional distribution and

a marginal distribution: $f(y, x) = f(y|x)f(x)$. The arguments of these two distributions will retain the properties of statistical independence. That is to say, the random variables $\varepsilon = y - \mu_{y|x}$ and $\nu = x - \mu_x$ are, by construction, statistically independent with $C(\varepsilon, \nu) = 0$.

Least-Squares Regression Analysis

Galton's analysis, which described the regression relationship between the heights of fathers and their sons, was an exercise in descriptive statistics which was wrought upon a given set of data. There can be no presumption that, for a different race of men living in a different environment, the same parameters would be uncovered. It is only as an experiment in thought that we may vary the value of the explanatory variable x and watch the concomitant variation of y. The heights of individual men are not subject to experimental manipulation.

Econometrics, in contrast to descriptive statistics, is often concerned with functional regression relationships which purport to describe the effects of manipulating the instruments of economic policy such as interest rates and rates of taxation. In such cases, it is no longer appropriate to attribute a statistical distribution to the explanatory variable x which now assumes the status of a control variable. Therefore it is necessary to derive the formulae of regression analysis from principles which make no reference to the joint distribution of the variables. The principle of least squares is appropriate to this purpose.

Before admitting this change of emphasis, we should offer some words of caution. For it seems that many of the errors of applied econometrics arise when an analyst imagines that, in fitting a regression equation, he has uncovered a causal connection.

The data which is used in inferring a regression relationship is part of an historical record of the evolution of the economy; and it is never certain that the same statistical relationships would have prevailed in other circumstances. Nor is it clear that they will prevail in the future.

An econometric analysis if often conducted with a view to guiding the actions of a regulatory agent. However, such actions are liable to alter the statistical relationships prevailing amongst economic variables. An assertion that a particular relationship will endure through time and that it will be unaffected by regulatory intercessions ought to be greeted with skepticism. Yet, in such matters, applied econometricians are often eager to suspend their disbelief.

To assist the application of the method of least squares, the regression equation, which has been defined by $E(y|x) = \alpha + \beta x$, can be written, alternatively, as

$$y = \alpha + x\beta + \varepsilon, \tag{39}$$

where $\varepsilon = y - E(y|x)$ is a random variable with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. This equation may be used to depict a functional relationship between an independent variable x and a dependent variable y. The relationship is affected by a disturbance ε which is independent of x and which might be taken to represent the effect of a large number of variables of minor importance which are not taken into account explicitly in describing the relationship.

Imagine that there is a sample of observations $(x_1, y_1), \ldots, (x_T, y_T)$ and that, from these data, we wish to estimate the parameters α and β. The principle of least squares suggests that we should do so by choosing the values which minimise the quantity

$$\begin{aligned} S &= \sum_{t=1}^{T} \varepsilon_t^2 \\ &= \sum_{t=1}^{T} (y_t - \alpha - x_t\beta)^2. \end{aligned} \tag{40}$$

This is the sum of squares of the vertical distances—measured parallel to the y-axis—of the data points from an interpolated regression line.

Differentiating the function S with respect to α and setting the results to zero for a minimum gives

$$\begin{aligned} -2\sum(y_t - \alpha - \beta x_t) = 0, \quad &\text{or, equivalently,} \\ \bar{y} - \alpha - \beta\bar{x} = 0. & \end{aligned} \tag{41}$$

This generates the following estimating equation for α:

$$\alpha(\beta) = \bar{y} - \beta\bar{x}. \tag{42}$$

Next, by differentiating with respect to β and setting the result to zero, we get

$$-2\sum x_t(y_t - \alpha - \beta x_t) = 0. \tag{43}$$

On substituting for α from (42) and eliminating the factor -2, this becomes

$$\sum x_t y_t - \sum x_t(\bar{y} - \beta\bar{x}) - \beta\sum x_t^2 = 0, \tag{44}$$

whence we get

$$\begin{aligned} \hat{\beta} &= \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} \\ &= \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}. \end{aligned} \tag{45}$$

This expression is identical to the one under (22) which we have derived from the method of moments. By putting $\hat{\beta}$ into the estimating equation for α under (42), we derive the same estimate $\hat{\alpha}$ for the intercept parameter as the one to be found under (22).

The method of least squares does not automatically provide an estimate of $\sigma^2 = E(\varepsilon_t^2)$. To obtain an estimate, we may invoke the method of moments which, in view of the fact that the regression residuals $e_t = y_t - \hat{\alpha} - \hat{\beta}x_t$ represent estimates of the corresponding values of ε_t, suggests an estimator in the form of

$$\tilde{\sigma}^2 = \frac{1}{T}\sum e_t^2. \tag{46}$$

In fact, this is a biased estimator with

$$E\left(\frac{\tilde{\sigma}^2}{T}\right) = \left(\frac{T-2}{T}\right)\sigma^2; \tag{47}$$

so it is common to adopt the unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum e_t^2}{T-2}. \tag{48}$$

We shall have occasion to demonstrate the unbiasedness of this estimator later. To understand the result on an intuitive level, one may recall that the unbiased estimator of the variance of a distribution, which is constructed from a random sample, is $\hat{\sigma}^2 = (T-1)^{-1}\sum(x_t - \hat{x})^2$. If the mean of the distribution μ were known and were used in place $\bar{x}$, then one should divide by T instead of $T-1$ to form $\hat{\sigma}^2 = T^{-1}\sum(x_t - \mu)^2$. The effect of using the datum $\bar{x}$ in place of the unknown mean μ would to reduce the measure of dispersion.To compensate, the measure is scaled by the factor $T/(T-1)$. In the context of the regression equation, where two parameters are estimated, the scale factor $T/(T-2)$ is used.

Properties of the Least-Squares Estimator

Now we shall reveal some of the properties of the least-squares estimators which follow from the assumptions which we have made so far. We shall also consider the likelihood that these assumptions will be fulfilled in practice, as well as some consequences of their violation.

We have assumed that the disturbance term ε is a random variable with

$$E(\varepsilon_t) = 0, \quad \text{and} \quad V(\varepsilon_t) = \sigma^2 \quad \text{for all } t. \tag{49}$$

We have avoided making statistical assumptions about x since we are unwilling to assume that its assembled values will manifest the sort of the regularities which are inherent in a statistical distribution. Therefore, we cannot express the assumption that ε is independent of x in terms of a joint distribution of these quantities; and, in particular, we should not assert that $C(x, \varepsilon) = 0$. However, if we are prepared to regard the x_t as predetermined values which have no effect on the ε_t, then we can say that

$$E(x_t\varepsilon_t) = x_t E(\varepsilon_t) = 0, \quad \text{for all } t. \tag{50}$$

In place of an assumption attributing a finite variance to x, we may assert that

$$\lim(T \to \infty)\frac{1}{T}\sum_{t=1}^{T} x_t^2 = m_{xx} < \infty. \tag{51}$$

For the random sequence $\{x_t\varepsilon_t\}$, we assert that

$$\operatorname{plim}(T \to \infty)\frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t = 0. \tag{52}$$

To see the effect of these assumptions, let us substitute the expression

$$y_t - \bar{y} = \beta(x_t - \bar{x}) + \varepsilon_t - \bar{\varepsilon} \tag{53}$$

in the expression for $\hat{\beta}$ found under (45). By rearranging the result, we have

$$\hat{\beta} = \beta + \frac{\sum(x_t - \bar{x})\varepsilon_t}{\sum(x_t - \bar{x})^2}. \tag{54}$$

The numerator of the second term on the RHS is obtained with the help of the identity

$$\begin{aligned}\sum(x_t - \bar{x})(\varepsilon_t - \bar{\varepsilon}) &= \sum(x_t\varepsilon_t - \bar{x}\varepsilon_t - x_t\bar{\varepsilon} + \bar{x}\bar{\varepsilon}) \\ &= \sum(x_t - \bar{x})\varepsilon_t.\end{aligned} \tag{55}$$

From the assumption under (50), it follows that

$$E\big\{(x_t - \bar{x})\varepsilon_t\big\} = (x_t - \bar{x})E(\varepsilon_t) = 0 \quad \text{for all } t. \tag{56}$$

Therefore

$$\begin{aligned}E(\hat{\beta}) &= \beta + \frac{\sum(x_t - \bar{x})E(\varepsilon_t)}{\sum(x_t - \bar{x})^2} \\ &= \beta;\end{aligned} \tag{57}$$

and $\hat{\beta}$ is seen to be an unbiased estimator of β.

The consistency of the estimator follows, likewise, from the assumptions under (51) and (52). Thus

$$\begin{aligned}\operatorname{plim}(\hat{\beta}) &= \beta + \frac{\operatorname{plim}\Big\{T^{-1}\sum(x_t - \bar{x})\varepsilon_t\Big\}}{\operatorname{plim}\Big\{T^{-1}\sum(x_t - \bar{x})^2\Big\}} \\ &= \beta;\end{aligned} \tag{58}$$

and $\hat{\beta}$ is seen to be a consistent estimator of β.

The consistency of $\hat{\beta}$ depends crucially upon the assumption that the disturbance term is independent of, or uncorrelated with, the explanatory variable or regressor x. In many econometric contexts, we should be particularly wary of this assumption. For, as we have suggested earlier, the disturbance term is liable to be compounded from the variables which have been omitted from the equation which explains y in terms of x. In a time-dependent context, these variables are liable to be correlated amongst themselves; and there may be scant justification for assuming that they are not likewise correlated with x.

There are other reasons of a more subtle nature for why the assumption of the independence of ε and x may be violated. The following example illustrates one of the classical problems of econometrics.

Example. In elementary macroeconomic theory, a simple model of the economy is postulated which comprises two equations:

$$y = c + i, \tag{59}$$

$$c = \alpha + \beta y + \varepsilon. \tag{60}$$

Here y stands for the gross product of the economy, which is also the income of consumers, i stands for investment and c stands for consumption. An additional identity $s = y - c$ or $s = i$, where s is savings, is also entailed. The disturbance term ε, which is omitted from the usual presentation in economics textbooks, is assumed to be independent of the variable i.

On substituting the consumption function of (60) into the income identity of (59) and rearranging the result, we find that

$$y = \frac{1}{1-\beta}\big(\alpha + i + \varepsilon\big), \tag{61}$$

from which

$$y_t - \bar{y} = \frac{1}{1-\beta}\big(i_t - \bar{i} + \varepsilon_t - \bar{\varepsilon}\big). \tag{62}$$

The ordinary least-squares estimator of the parameter β, which is called the marginal propensity to consume, gives rise to the following equation:

$$\hat{\beta} = \beta + \frac{\sum(y_t - \bar{y})\varepsilon_t}{\sum(y_t - \bar{y})^2}. \tag{63}$$

Equation (61), which shows that y is dependent on ε, suggests that $\hat{\beta}$ cannot be a consistent estimator of β.

To determine the probability limit of the estimator, we must assess the separate probability limits of the numerator and the denominator of the term on the RHS of (63).

The following results are available:

$$\begin{aligned} \lim \frac{1}{T}\sum_{t=1}^{T}(i_t - \bar{i})^2 &= m_{ii} = V(i), \\ \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 &= \frac{m_{ii} + \sigma^2}{(1-\beta)^2} = V(y), \\ \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})\varepsilon_t &= \frac{\sigma^2}{1-\beta} = C(y, \varepsilon). \end{aligned} \tag{64}$$

The results indicate that

$$\begin{aligned} \operatorname{plim}\ \hat{\beta} &= \beta + \frac{\sigma^2(1-\beta)}{m_{ii} + \sigma^2} \\ &= \frac{\beta m_{ii} + \sigma^2}{m_{ii} + \sigma^2}; \end{aligned} \tag{65}$$

and it can be seen that the limiting value of $\hat{\beta}$ has an upward bias which increases as the ratio σ^2/m_{ii} increases.

On the assumption that the model is valid, it is easy to understand why the parameter of the regression of c on y exceeds the value of the marginal propensity to consume. We can do so by considering the extreme cases.

Imagine, first, that $\sigma^2 = V(\varepsilon) = 0$. Then the only source of variation in y and c is the variation in i. In that case, the parameter of the regression of c on y will coincide with β. This is illustrated in Figure 1. Now imagine, instead, that i is constant and that the only variations in c and y are due ε which is disturbs consumption. Then the expected value of consumption is provided by the equation $c = y - i$ in which the coefficient associated with y is unity. Figure 2 illustrates this case. Assuming now that both $m_{ii} > 0$ and $\sigma^2 > 0$, it follows

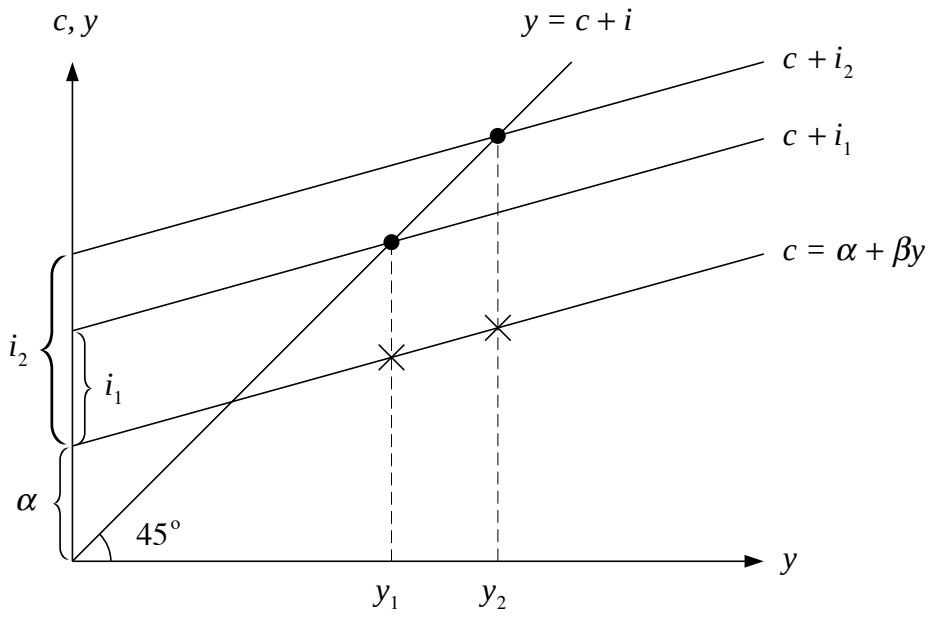

Figure 1. If the only source of variation in y is the variation in i, then the observations on y and c will delineate the consumption function.

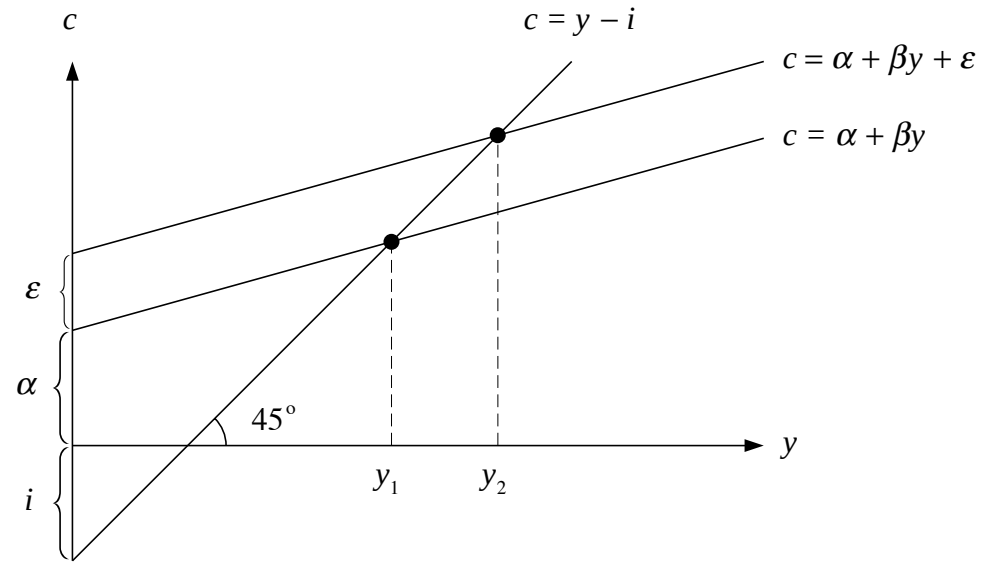

Figure 2. If the only source of variation in y are the disturbances to c, then the observations on y and c will line along a 45° line.

that the value of the regression parameter must lie somewhere in the interval $[\beta, 1]$.

Although it may be inappropriate for estimating the structural parameter β, the direct regression of c on y does provide the conditional expectation $E(c|y)$; and this endows it with a validity which it retains even if the Keynesian model of (59) and (60) is misspecified.

In fact, the simple Keynesian model of (59) and (60) is more an epigram than a serious scientific theory. Common sense dictates that we should give more credence to the estimate of the conditional expectation $E(c|y)$ than to a putative estimate of the marginal propensity to consume devised within the context of a doubtful model.

The Method of Maximum Likelihood

The method of maximum-likelihood constitutes a principle of estimation which may be applied to a wide variety of problems. One of the attractions of the method is that, granted the fulfilment of the assumptions on which it is based, it can be shown that the resulting estimates have optimal properties. In general, it can be shown that, at least in large samples, the variance of the resulting estimates is the least that can be achieved by any method.

The cost of using the method is precisely the need to make the assumptions which are necessary to sustain it. It is often difficult to assess, without a great deal of further analysis, the extent to which the desirable properties of the maximum-likelihood estimators survive when these assumptions are not fulfilled. In the case of the regression model, there is considerable knowledge on this account, some of which will be presented in later chapters.

The model to which we apply the method is the regression model with independently and identically distributed disturbances which follow a normal probability law. The probability density functions of the individual disturbances $\varepsilon_t; t = 1, \ldots, T$ are given by

$$N(\varepsilon_t; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right). \tag{66}$$

Since the ε's are assumed to be independently distributed, their joint probability density function (p.d.f.) is

$$\prod_{t=1}^{T} N(\varepsilon_t; 0, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(\frac{-1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon^2\right). \tag{67}$$

If we regard the elements $x_1, \ldots, x_T$ as a given set of numbers, then it follows that the conditional p.d.f. of the sample $y_1, \ldots, y_T$ is

$$f(y_1, \ldots, y_T|x_1, \ldots, x_T) = (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)\right\}. \tag{68}$$

The principle of maximum likelihood suggests that we should estimate α, β and σ^2 by choosing the values which maximise the probability measure which is attributed to the sample $y_1, \ldots, y_T$. That is to say, one chooses to regard the events which have generated the sample as the most likely of all the events which could have occurred.

Notice that, when α, β and σ^2 are the arguments of the function f rather than its parameters, and when $y_1, \ldots, y_T$ are data values rather than random variables, the function is no longer a probability density function. For this reason, we are apt to call it a likelihood function instead and to denote it by $L(\alpha, \beta, \sigma^2)$.

The log of the likelihood function, which has the same maximising values as the original function, is

$$\log L = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2. \tag{69}$$

It is clear that, given the value of σ^2, the likelihood is maximised by the values $\hat{\alpha}$ and $\hat{\beta}$ which minimise the sum of squares; and we already have expressions for $\hat{\alpha}$ and $\hat{\beta}$ under (42) and (45) respectively.

We may obtain the maximum-likelihood estimator for σ^2 from the following first-order condition:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2 = 0. \tag{70}$$

By multiplying throughout by $2\sigma^4/T$ and rearranging the result, we get the following estimating equation:

$$\sigma^2(\alpha, \beta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2. \tag{71}$$

By putting $\hat{\alpha}$ and $\hat{\beta}$ in place, we obtain the estimator $\tilde{\sigma}^2 = \sigma^2(\hat{\alpha}, \hat{\beta}) = T^{-1}\sum e_t$ already given under (46).

The General Theory of M-L Estimation

In order to derive an M-L estimator, we are bound to make an assumption about the functional form of the distribution which generates the data. However, the assumption can often be varied without affecting the form of the M-L estimator; and the general theory of maximum-likelihood estimation can be developed without reference to a specific distribution.

In fact, the M-L method is of such generality that it provides a model for most other methods of estimation. For the other methods tend to generate

estimators which can be depicted as approximations to the maximum-likelihood estimators, if they are not actually identical to the latter.

In order to reveal the important characteristics of the likelihood estimators, we should investigate the properties of the log-likelihood function itself.

Consider the case where θ is the sole parameter of a log-likelihood function $\log L(y;\theta)$ wherein $y = [y_1, \ldots, y_T]$ is a vector of sample elements. In seeking to estimate the parameter, we regard θ as an argument of the function whilst the elements of y are considered to be fixed. However, in analysing the statistical properties of the function, we restore the random character to the sample elements. The randomness is conveyed to the maximising value $\hat{\theta}$ which thereby acquires a distribution.

A fundamental result is that, as the sample size increases, the likelihood function divided by the sample size tends to stabilise in the sense that it converges in probability, at every point in its domain, to a constant function. In the process, the distribution of $\hat{\theta}$ becomes increasingly concentrated in the vicinity of the true parameter value θ_0. This accounts for the consistency of maximum-likelihood estimation.

To demonstrate the convergence of the log-likelihood function, we shall assume, as before, that the elements of $y = [y_1, \ldots, y_T]$ form a random sample. Then

$$L(y;\theta) = \prod_{t=1}^{T} f(y_t;\theta), \tag{72}$$

and therefore

$$\frac{1}{T}\log L(y;\theta) = \frac{1}{T}\sum_{t=1}^{T} \log f(y_t;\theta). \tag{73}$$

For any value of θ, this represents a sum of independently and identically distributed random variables. Therefore the law of large numbers can be applied to show that

$$\operatorname{plim}(T\to\infty)\frac{1}{T}\log L(y;\theta) = E\big\{\log f(y_t;\theta)\big\}. \tag{74}$$

The next step is to demonstrate that $E\{\log L(y;\theta_0)\} \geq E\{\log L(y;\theta)\}$, which is to say that the expected log-likelihood function, to which the sample likelihood function converges, is maximised by the true parameter value θ_0.

The first derivative of log-likelihood function is

$$\frac{d\log L(y;\theta)}{d\theta} = \frac{1}{L(y;\theta)}\frac{dL(y;\theta)}{d\theta}. \tag{75}$$

This is known as the score of the log-likelihood function at θ. Under conditions which allow the derivative and the integral to commute, the derivative of the expectation is the expectation of the derivative. Thus, from (75),

$$\frac{d}{d\theta}E\big\{\log L(y;\theta)\big\} = \int_y \left\{\frac{1}{L(y;\theta)}\frac{dL(y;\theta)}{d\theta}\right\} L(y;\theta_0)dy, \tag{76}$$

where θ_0 is the true value of θ and $L(y,\theta_0)$ is the probability density function of y. When $\theta = \theta_0$, the expression on the RHS simplifies in consequence of the cancellation of $L(y,\theta)$ in the denominator with $L(y,\theta_0)$ in the numerator. Then we get

$$\int_y \frac{dL(y;\theta_0)}{d\theta}dy = \frac{d}{d\theta}\int_y L(y;\theta_0)dy = 0, \tag{77}$$

where the final equality follows from the fact that the integral is unity, which implies that its derivative is zero. Thus

$$\frac{d}{d\theta}E\big\{\log L(y;\theta_0)\big\} = E\left\{\frac{d\log L(y;\theta_0)}{d\theta}\right\} = 0; \tag{78}$$

and this is a first-order condition which indicates that the $E\{\log L(y;\theta)/T\}$ is maximised at the true parameter value θ_0.

Given that the $\log L(y;\theta)/T$ converges to $E\{\log L(y;\theta)/T\}$, it follows, by some simple analytic arguments, that the maximising value of the former must converge to the maximising value of the latter: which is to say that $\hat{\theta}$ must converge to θ_0.

Now let us differentiate (75) in respect to θ and take expectations. Provided that the order of these operations can be interchanged, then

$$\frac{d}{d\theta}\int_y \frac{d\log L(y;\theta)}{d\theta}L(y;\theta)dy = \frac{d^2}{d\theta^2}\int_y L(y;\theta)dy = 0, \tag{79}$$

where the final equality follows in the same way as that of (77). The LHS can be expressed as

$$\int_y \frac{d^2\log L(y;\theta)}{d\theta^2}L(y;\theta)dy + \int_y \frac{d\log L(y;\theta)}{d\theta}\frac{dL(y;\theta)}{d\theta}dy = 0 \tag{80}$$

and, on substituting from (75) into the second term, this becomes

$$\int_y \frac{d^2\log L(y;\theta)}{d\theta^2}L(y;\theta)dy + \int_y \left\{\frac{d\log L(y;\theta)}{d\theta}\right\}^2 L(y;\theta)dy = 0. \tag{81}$$

Therefore, when $\theta = \theta_0$, we get

$$E\left\{-\frac{d^2 \log L(y;\theta_0)}{d\theta^2}\right\} = E\left[\left\{\frac{d \log L(y;\theta_0)}{d\theta}\right\}^2\right] = \Phi. \tag{82}$$

This measure is know as Fisher's Information. Since (78) indicates that the score $d \log L(y;\theta_0)/d\theta$ has an expected value of zero, it follows that Fisher's Information represents the variance of the score at θ_0.

Clearly, the information measure increases with the size of the sample. To obtain a measure of the information about θ which is contained, on average, in a single observation, we may define $\phi = \Phi/T$

The importance of the information measure Φ is that its inverse provides an approximation to the variance of the maximum-likelihood estimator which become increasingly accurate as the sample size increases. Indeed, this is the explanation of the terminology. The famous Cramèr–Rao theorem indicates that the inverse of the information measure provides a lower bound for the variance of any unbiased estimator of θ. The fact that the asymptotic variance of the maximum-likelihood estimator attains this bound, as we shall proceed to show, is the proof of the estimator's efficiency.

The Asymptotic Distribution of the M-L Estimator

The asymptotic distribution of the maximum-likelihood estimator is established under the assumption that the log-likelihood function obeys certain regularity conditions. Some of these conditions are not readily explicable without a context. Therefore, instead of itemising the conditions, we shall make an overall assumption which is appropriate to our own purposes but which is stronger than is strictly necessary. We shall image that $\log L(y;\theta)$ is an analytic function which can be represented by a Taylor-series expansion about the point θ_0:

$$\begin{aligned}\log L(\theta) = \log L(\theta_0) + \frac{d \log L(\theta_0)}{d\theta}(\theta - \theta_0) + \frac{1}{2}\frac{d^2 \log L(\theta_0)}{d\theta^2}(\theta - \theta_0)^2 \\ + \frac{1}{3!}\frac{d^3 \log L(\theta_0)}{d\theta^3}(\theta - \theta_0)^3 + \cdots.\end{aligned} \tag{83}$$

In pursuing the asymptotic distribution of the maximum-likelihood estimator, we can concentrate upon a quadratic approximation which is based the first three terms of this expansion. The reason is that, as we have shown, the distribution of the estimator becomes increasingly concentrated in the vicinity of the true parameter value as the size of the sample increases. Therefore the quadratic approximation becomes increasingly accurate for the range of values

of θ which we are liable to consider. It follows that, amongst the regularity conditions, there must be at least the provision that the derivatives of the function are finite-valued up to the third order.

The quadratic approximation to the function, taken at the point θ_0, is

$$\log L(\theta) = \log L(\theta_0) + \frac{d\log L(\theta_0)}{d\theta}(\theta-\theta_0) + \frac{1}{2}\frac{d^2\log L(\theta_0)}{d\theta^2}(\theta-\theta_0)^2. \tag{84}$$

Its derivative with respect to θ is

$$\frac{d\log L(\theta)}{d\theta} = \frac{d\log L(\theta_0)}{d\theta} + \frac{d^2\log L(\theta_0)}{d\theta^2}(\theta-\theta_0). \tag{85}$$

By setting $\theta = \hat{\theta}$ and by using the fact that $d\log L(\hat{\theta})/d\theta = 0$, which follows from the definition of the maximum-likelihood estimator, we find that

$$\sqrt{T}(\hat{\theta}-\theta_0) = \left\{-\frac{1}{T}\frac{d^2\log L(\theta_0)}{d\theta^2}\right\}^{-1}\left\{\frac{1}{\sqrt{T}}\frac{d\log L(\theta_0)}{d\theta}\right\}. \tag{86}$$

The argument which establishes the limiting distribution of $\sqrt{T}(\hat{\theta}-\theta_0)$ has two strands. First, the law of large numbers is invoked in to show that

$$-\frac{1}{T}\frac{d^2\log L(y;\theta_0)}{d\theta^2} = -\frac{1}{T}\sum_t \frac{d^2\log f(y_t;\theta_0)}{d\theta^2} \tag{87}$$

must converge to its expected value which is the information measure $\phi = \Phi/T$. Next, the central limit theorem is invoked to show that

$$\frac{1}{\sqrt{T}}\frac{d\log L(y;\theta_0)}{d\theta} = \frac{1}{\sqrt{T}}\sum_t \frac{d\log f(y_t;\theta_0)}{d\theta} \tag{88}$$

has a limiting normal distribution which is $N(0,\phi)$. This result depends crucially on the fact that $\Phi = T\phi$ is the variance of $d\log L(y;\theta_0)/d\theta$. Thus the limiting distribution of the quantity $\sqrt{T}(\hat{\theta}-\theta_0)$ is the normal $N(0,\phi^{-1})$ distribution, since this is the distribution of ϕ^{-1} times an $N(0,\phi)$ variable.

Within this argument, the device of scaling $\hat{\theta}$ by $\sqrt{T}$ has the purpose of preventing the variance from vanishing, and the distribution from collapsing, as the sample size increases indefinitely. Having completed the argument, we can remove the scale factor; and the conclusion which is to be drawn is the following:

(89) Let $\hat{\theta}$ be the maximum-likelihood estimator obtained by solving the equation $d\log L(y,\theta)/d\theta = 0$, and let θ_0 be the true value of

the parameter. Then $\hat{\theta}$ is distributed approximately according to the distribution $N(\theta_0, \Phi^{-1})$, where Φ^{-1} is the inverse of Fisher's measure of information.

In establishing these results, we have considered only the case where a single parameter is to estimated. This has enabled us to proceed without the panoply of vectors and matrices. Nevertheless, nothing essential has been omitted from our arguments. In the case where θ is a vector of k elements, we define the information matrix to be the matrix whose elements are the variances and covariances of the elements of the score vector. Thus the generic element of the information matrix, in the ijth position, is

$$E\left\{-\frac{\partial^2 \log L(\theta_0)}{\partial\theta_i\partial\theta_j}\right\} = E\left\{\frac{\partial \log L(\theta_0)}{\partial\theta_i}.\frac{\partial \log L(\theta_0)}{\partial\theta_j}\right\}. \tag{90}$$