CHAPTER 10

The Theory of Inference

This chapter summarises some results in the classical theory of statistical inference which depends heavily on the method of maximum-likelihood estimation.

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. Thus, the estimates are statistically consistent and their asymptotic distributions have the least possible variance.

Springing from the asymptotic theory of maximum-likelihood estimation is a powerful theory of hypothesis testing which makes use of a collection of alternative, but asymptotically equivalent, test statistics which are the Wald statistic, the likelihood-ratio statistic and the Lagrangean multiplier statistic.

The practical virtue of the method of maximum likelihood is that it often leads directly to a set of estimating equations which could have been derived more laboriously and more doubtfully from other principles of estimation. In other words, the method can be used as a vehicle for reaching the objectives of estimation.

When the estimating equations are in hand, one is often inclined to discard some of the original assumptions which have been used in their derivation. The assumptions might be unrealistic and that they might not be crucial to the validity of the estimation procedure. In that case, one is inclined to describe the estimates as quasi maximum-likelihood estimates.

Principles of Estimation

Let $Y' = [y_1, \ldots, y_T]$ be a data matrix comprising T realisations of a random vector y whose marginal probability density function $f(y;\theta)$ is characterised by the parameter vector $\theta = [\theta_1, \ldots, \theta_k]'$. Then any function $\hat{\theta} = \hat{\theta}(Y)$ of the data which purports to provide a useful approximation to parameter vector is called a point estimator.

The joint probability density function of the elements of Y can be expressed as the product

$$\begin{aligned} L(Y;\theta) &= f(y_T|y_{T-1}, \ldots, y_1) \cdots f(y_2|y_1)f(y_1) \\ &= f(y_1)\prod_{t=2}^{T} f(y_t|y_{t-1}, \ldots, y_1), \end{aligned} \tag{25.1}$$

where $f(y_t|y_{t-1}, \ldots, y_1)$ is the conditional probability density function of y_t given the preceeding values $y_{t-1}, \ldots, y_1$ and $f(y_1)$ is the marginal probability density function of the initial vector y_1. In classical theory, the vectors of the sequence

$y_1, \ldots, y_T$ are assumed to be independently and identically distributed, which enables us to write

$$\begin{aligned} L(Y;\theta) &= f(y_T)\cdots f(y_2)f(y_1) \\ &= \prod_{t=1}^{T} f(y_t) \end{aligned} \tag{25.2}$$

in place of (25.1).

The set $\mathcal{S}$ comprising all possible values of the data matrix Y is called the sample space, and the set $\mathcal{A}$ of all values of θ which conform to whatever restrictions have been postulated is called the admissible parameter space. A point estimator is therefore a function which associates with every value Y in $\mathcal{S}$ a unique value $\hat{\theta}$ in $\mathcal{A}$.

There are numerous principles which can be used in constructing estimators. The principle of maximum-likelihood estimation is a fundamental one. The idea is that we should estimate θ by choosing the value which maximises the probability measure attributed to Y. Thus

(25.3) A maximum-likelihood estimate $\hat{\theta} = \hat{\theta}(Y)$ is an element of the admissible parameter space for which $L(Y;\hat{\theta}) \geq L(Y;\theta)$ for every $\theta \in \mathcal{A}$.

Another common principle of estimation is the method of moments. In many cases, it will be possible to estimate the moments of the density function $f(y)$ in a straightforward manner. If the parameter vector θ is expressible as a function of these moments, then an estimator can be constructed which uses the same function and which replaces the moments by their estimates.

We shall concentrate primarily on the method of maximum likelihood which is widely applicable, and we shall demonstrate that maximum-likelihood estimators have certain optimal properties. Usually, we are able to justify the estimators which are derived from other principles by showing that, as the size of the data sample increases, they tend to approximate to the corresponding maximum-likelihood estimators with increasing accuracy.

Identifiability

Before examining the properties of maximum-likelihood estimators in detail, we should consider some preconditions which must be satisfied before any reasonable inferences can be made about the parameter θ. We can estimate θ only if its particular value is somehow reflected in the realised value of Y. Therefore, a basic requirement is that distinct values of θ should lead to distinct probability density functions. Thus we may declare that

(25.4) The parameter values in $\mathcal{A}$ are identifiable if, for any two distinct values $\theta_1, \theta_2 \in \mathcal{A}$, we have $L(Y;\theta_1) \neq L(Y;\theta_2)$ for all Y in a subset of $\mathcal{S}$ which has a nonzero probability measure in respect of either of the distributions implied by θ_1, θ_2.

There are numerous ways of comparing the values $L(Y;\theta_1)$ and $L(Y;\theta_2)$ over the set $\mathcal{S}$. However, the requirement of (25.4) would certainly be fulfilled if the measure

$$\int_{\mathcal{S}} \Big\{ \log L(Y;\theta_1) - \log L(Y;\theta_2) \Big\} L(Y;\theta_2) dY \tag{25.5}$$

were nonzero for all values of θ_1, θ_2 which are distinct.

A concept which may sometimes serve in place of identifiability is that of unbiased estimability. We say that

(25.6) The parameter θ is unbiasedly estimable if and only if there exists some function $\hat{\theta} = \hat{\theta}(Y)$ such that $E(\hat{\theta}) = \theta$.

A parameter which is unbiasedly estimable is certainly identifiable according to the criterion previous criterion (25.4); for if $\theta_1 = E(\hat{\theta}|\theta_1) = \int \hat{\theta} L(Y;\theta_1) dY$ and $\theta_2 = E(\hat{\theta}|\theta_2) = \int \hat{\theta} L(Y;\theta_2) dY$ are distinct values, then it must be true that $L(Y;\theta_1) \neq L(Y;\theta_2)$ over a measureable set in $\mathcal{S}$. Unfortunately, the concept of unbiased estimability is of limited use since it is often difficult, if not impossible, to prove that an unbiased estimator exists. Indeed, there are cases where none of the estimators which are worth considering have finite moments of any order.

The criterion of identifiability under (25.4) may be too stringent, for it is difficult to talk broadly of the generality of values in $\mathcal{A}$. It may be that some elements of $\mathcal{A}$ are identifiable whilst others are not. Therefore, in the main, we have to be content with saying that

(25.7) The parameter vector $\theta_0 \in \mathcal{A}$ is identifiable if there exists no other $\theta \in \mathcal{A}$ such that $L(Y;\theta) = L(Y;\theta_0)$ with a probability of 1 when Y is regarded as a random variable. If $L(Y;\theta_0) = L(Y;\theta_1)$ with a probability of 1, then θ_0, θ_1 are observationally equivalent.

By concentrating our attention on the point θ_0, we can put out of mind the pitfalls which may be lurking elsewhere in the parameter space $\mathcal{A}$.

Our object must be to establish necessary and sufficient conditions for identifiability which can be checked easily. For this purpose, it is useful to consider the so-called information integral. Imagine, therefore, that $L(Y;\theta_0)$ is the probability density function of the process which generates the data, and let $L(Y;\theta)$ be construed as a function of $\theta \in \mathcal{A}$. Then the information integral is defined as the function

$$\begin{aligned} H(\theta;\theta_0) &= \int_{\mathcal{S}} \log\left\{\frac{L(Y;\theta)}{L(Y;\theta_0)}\right\} L(Y;\theta_0) dY \\ &= E\left[\log\left\{\frac{L(Y;\theta)}{L(Y;\theta_0)}\right\}\right]. \end{aligned} \tag{25.8}$$

This function, which is an instance of the function under (25.5), provides a measure of the extent to which the statistical implications of θ differ from those of θ_0.

The expectation is formed under the presumption that θ_0 is the true value. It is straightforward to show that

$$H(\theta;\theta_0) \leq 0 \quad \text{with} \quad H(\theta;\theta_0) = 0 \quad \text{when} \quad \theta = \theta_0. \tag{25.9}$$

Proof. It is clear that $H(\theta_0, \theta_0) = 0$. To show that $H(\theta, \theta_0) \leq 0$, we may employ Jensen's inequality which indicates that, if $x \sim f(x)$ is a random variable and $g(x)$ is a strictly concave function, then $E\{g(x)\} < g\{E(x)\}$. This result, which is little more than a statement that $\lambda g(x_1) + (1-\lambda)g(x_2) < g\{\lambda x_1 + (1-\lambda)x_2\}$ when $0 < \lambda < 1$, is proved by Rao [421]. Noting that $\log(z)$ is a strictly concave function, we find that

$$\begin{aligned} H(\theta,\theta_0) &= E\left[\log\left\{\frac{L(Y;\theta)}{L(Y;\theta_0)}\right\}\right] \\ &\leq \log\left[E\left\{\frac{L(Y;\theta)}{L(Y;\theta_0)}\right\}\right] \\ &= \log\int_{\mathcal{S}}\left\{\frac{L(Y;\theta)}{L(Y;\theta_0)}\right\}L(Y;\theta_0)dY \\ &= \log 1 = 0. \end{aligned} \tag{25.10}$$

It follows, from the definition of the information measure and from the conditions under (25.9), that

(25.11) The parameter vector θ_0 is identifiable if and only if there is no other vector θ sharing the maximum information measure. Equivalently, θ_0 is identifiable if and only if the equation $H(\theta;\theta_0) = 0$ has the unique solution $\theta = \theta_0$.

The condition for the identifiability of θ_0 is therefore the condition that $H(\theta;\theta_0)$ should achieve a unique global maximum at this point. Conditions for global maximisation are hard to come by. The conditions for local maximisation and therefore for local identifiability are more accessible. In saying that θ_0 is locally identified, we mean that there is no other point in the neighbourhood sharing the maximum information measure. Thus

(25.12) If $H(\theta,\theta_0)$ has continuous first and second derivatives in an open neighbourhood of the parameter point θ_0, then a necessary and sufficient condition for the local identifiability of θ_0, is that $\partial H/\partial\theta = 0$ and that $\partial\{\partial H/\partial\theta\}'/\partial\theta$ is negative definite at this point.

The points in $\mathcal{A}$ in whose neighbourhood the derivatives are continuous may be described as regular points. Usually, we can make assumptions which guarantee that the irregular points of $\mathcal{A}$, where the derivatives are discontinuous, constitute a set of measure zero.

The Information Matrix

The condition for identifiability given under (25.12) can be expressed in terms of a classical statistical construct known as Fisher's Information Matrix. In order to demonstrate this connection, we need to derive a series of fundamental identities which are used throughout the development of the theory of estimation. First let us consider the identity

$$\frac{\partial L(Y;\theta)}{\partial\theta} = \frac{\partial \log L(Y;\theta)}{\partial\theta} L(Y;\theta). \tag{25.13}$$

This comes from rearranging the equation $\partial \log L/\partial\theta = (1/L)\partial L/\partial\theta$. Next we may consider the condition

$$1 = \int_{\mathcal{S}} L(Y;\theta)dY. \tag{25.14}$$

Differentiating under the integral with respect to θ and using (25.13) gives a further useful identity:

$$0 = \int_{\mathcal{S}} \frac{\partial L(Y;\theta)}{\partial\theta} dY = \int_{\mathcal{S}} \frac{\partial \log L(Y;\theta)}{\partial\theta} L(Y;\theta)dY. \tag{25.15}$$

Setting $\theta = \theta_0$ in this equation gives the condition

$$E\left\{\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\} = 0. \tag{25.16}$$

Differentiating (25.15) with the help of (25.13) gives

$$0 = \int_{\mathcal{S}} \left[\frac{\partial(\partial \log L(Y;\theta)/\partial\theta)'}{\partial\theta} + \left\{\frac{\partial \log L(Y;\theta)}{\partial\theta}\right\}' \left\{\frac{\partial \log L(Y;\theta)}{\partial\theta}\right\}\right] L(Y;\theta)dY. \tag{25.17}$$

Setting $\theta = \theta_0$ in the latter serves to show that

$$\begin{aligned} E\left[\left\{\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\}' \left\{\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\}\right] &= -E\left[\frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial\theta}\right] \\ &= Q(\theta_0). \end{aligned} \tag{25.18}$$

Also, in the light of equation (25.16), we can interpret the first expression of (25.18) as the dispersion matrix of the derivative $\partial \log L(Y;\theta)/\partial\theta$ evaluated at $\theta = \theta_0$; and thus we can write

$$Q(\theta_0) = D\left(\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right). \tag{25.19}$$

The matrix $Q(\theta_0)$ is known as Fisher's Information Matrix.

The information matrix is, in fact, the negative of the matrix of second derivatives of the information measure $H(\theta;\theta_0)$ at the point $\theta = \theta_0$. Consider the first derivative of the measure:

$$\begin{aligned} \frac{\partial H(\theta;\theta_0)}{\partial\theta} &= \int_{\mathcal{S}} \frac{\partial}{\partial\theta}\Big\{\log L(Y;\theta) - \log L(Y;\theta_0)\Big\} L(Y;\theta_0)dY \\ &= \int_{\mathcal{S}} \frac{\partial \log L(Y;\theta)}{\partial\theta} L(Y;\theta_0)dY. \end{aligned} \tag{25.20}$$

Setting $\theta = \theta_0$ delivers the identity under (25.16); and this reflects the fact that θ_0 is a stationary point of the function. Differentiating a second time and setting $\theta = \theta_0$ gives

$$\begin{aligned}\frac{\partial(\partial H(\theta_0;\theta_0)/\partial\theta)'}{\partial\theta} &= E\left[\frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial\theta}\right] \\ &= -Q(\theta_0).\end{aligned} \tag{25.21}$$

This is the negative of Fisher's Information Matrix. In view of the statement under (25.12), we may conclude that

(25.22) The parameter vector θ_0 is identifiable if the information matrix $Q(\theta_0)$ is positive definite.

The Efficiency of Estimation

To be of any worth, a estimator must possess a probability distribution which is closely concentrated around the true value of the unknown parameter. The easiest way of characterising such a distribution is in terms its moments. However, as we have already indicated, these moments might not exist. Nevertheless, it is usually the case that, as the size of the sample increases, an estimator will converge in probability upon a random variable whose distribution has well-defined moments. We must content ourselves, in the main, with analysing such limiting distributions. For the moment, we shall imagine that our estimator $\hat{\theta} = \hat{\theta}(Y)$ is unbiased and that it has a finite variance.

For an unbiased estimator, the natural measure of concentration is the variance. For any given sample, there is a bound below which the variance of an unbiased estimator cannot be reduced.

(25.23) Let $L(Y;\theta_0)$ be the density function of the sample Y. If $\hat{\theta} = \hat{\theta}(Y)$ is an unbiased estimator of θ, and if q is any vector of the same order, then we have $V(q'\hat{\theta}) \geq q'Q(\theta_0)q$, where $Q(\theta_0)$ is the information matrix specified in (25.18) and (25.19). This is the Cramér–Rao inequality.

Proof. Let us consider the condition which asserts that $\hat{\theta} = \hat{\theta}(Y)$ is an unbiased estimator:

$$\begin{aligned}E\left\{\hat{\theta}(Y)\right\} &= \int_{\mathcal{S}} \hat{\theta}(Y)L(Y;\theta_0)dY \\ &= \theta_0.\end{aligned} \tag{25.24}$$

The derivative is

$$\begin{aligned}\frac{\partial E\{\hat{\theta}(Y)\}}{\partial\theta} &= \int_{\mathcal{S}} \hat{\theta}(Y)\frac{\partial \log L(Y;\theta_0)}{\partial\theta}L(Y;\theta_0)dY \\ &= E\left\{\hat{\theta}(Y)\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\} = I.\end{aligned} \tag{25.25}$$

Now a pair of random vectors a, b have a covariance of $C(a,b) = E(ab')$ when $E(b) = 0$. Therefore, since $E\{\partial \log L(Y;\theta_0)/\partial\theta\} = 0$, it follows that the final equality under (25.25) can be written as

$$C\left(\hat{\theta}, \frac{\partial \log L(\theta_0)}{\partial\theta}\right) = I. \tag{25.26}$$

The joint dispersion matrix of $\hat{\theta}$ and $\partial \log L(Y;\theta_0)/\partial\theta$ is

$$\begin{aligned} D\begin{bmatrix} \hat{\theta} \\ \left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}\right)' \end{bmatrix} &= \begin{bmatrix} D(\hat{\theta}) & C\left(\hat{\theta}, \dfrac{\partial \log L(\theta_0)}{\partial\theta}\right) \\ C\left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}, \hat{\theta}\right) & D\left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}\right) \end{bmatrix} \\ &= \begin{bmatrix} D(\hat{\theta}) & I \\ I & Q(\theta_0) \end{bmatrix}. \end{aligned} \tag{25.27}$$

This is a positive-semidefinite matrix. It follows that

$$[\, q' \quad -q'Q^{-1}(\theta_0)\,] \begin{bmatrix} D(\hat{\theta}) & I \\ I & Q(\theta_0) \end{bmatrix} \begin{bmatrix} q \\ -Q^{-1}(\theta_0)q \end{bmatrix} = q'D(\hat{\theta})q - q'Q^{-1}(\theta_0)q \geq 0. \tag{25.28}$$

Using $q'D(\hat{\theta})q = V(q'\hat{\theta})$, we can write this inequality as $V(q'\hat{\theta}) \geq q'Q(\theta_0)q$ which is the desired result.

Now consider the case where $\hat{\theta}$ attains the minimum variance bound. Then $V(q'\hat{\theta}) - q'Q^{-1}(\theta_0)q = 0$ or equivalently

$$[\, q' \quad -q'Q^{-1}(\theta_0)\,] D \begin{bmatrix} \hat{\theta} \\ \left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}\right)' \end{bmatrix} \begin{bmatrix} q \\ -Q^{-1}(\theta_0)q \end{bmatrix} = 0. \tag{25.29}$$

But this is equivalent to the condition that

$$[\, q' \quad -q'Q^{-1}(\theta_0)\,] \begin{bmatrix} \hat{\theta} - E(\hat{\theta}) \\ \left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}\right)' - E\left(\dfrac{\partial \log L(\theta_0)}{\partial\theta}\right)' \end{bmatrix} = 0, \tag{25.30}$$

whence, using the condition of unbiasedness $E(\hat{\theta}) = \theta_0$ and the condition $E\{\partial \log L(\theta_0)/\partial\theta\} = 0$ from (25.16), we get

$$q'(\hat{\theta} - \theta_0) - q'Q^{-1}(\theta_0)\left(\frac{\partial \log L(\theta_0)}{\partial\theta}\right)' = 0. \tag{25.31}$$

Since this holds for all q, we must have $\hat{\theta} - \theta_0 = Q^{-1}(\theta_0)(\partial \log L(\theta_0)/\partial\theta)'$. What we have shown is that

(25.32) Subject to regularity conditions, there exists an unbiased estimator $\hat{\theta}(Y)$ whose variance attains the Cramér–Rao minimum-variance bound if and only if $\partial \log L(Y;\theta)/\partial\theta$ can be expressed in the form

$$\left(\frac{\partial \log L}{\partial \theta}\right)' = -E\left\{\frac{\partial(\partial \log L/\partial\theta)'}{\partial\theta}\right\}(\hat{\theta}-\theta).$$

This is, in fact, a rather strong requirement; and therefore it is only in exceptional circumstances that the minimum-variance bound can be attained. However, as we shall see shortly, whenever the regularity conditions are satisfied, the variance associated with the limiting distribution of the maximum-likelihood estimates invariably attains the bound. Indeed, the equation

$$(\hat{\theta}-\theta) = -\left[E\left\{\frac{\partial(\partial \log L/\partial\theta)'}{\partial\theta}\right\}\right]^{-1}\left(\frac{\partial \log L}{\partial \theta}\right)' \tag{25.33}$$

is the prototype of a form of asymptotic equation which the maximum-likelihood estimators satisfy in the limit when the sample size becomes indefinitely large.

Unrestricted Maximum-Likelihood Estimation

(25.34) If $\hat{\theta}$ is the maximum-likelihood estimator obtained by solving the equation $\partial \log L(Y;\theta)/\partial\theta = 0$, and if θ_0 is the true parameter value, then $\sqrt{T}(\hat{\theta}-\theta_0)$, has the limiting distribution $N(0, M^{-1})$ where

$$\begin{aligned} M &= -\frac{1}{T}E\left\{\frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial\theta}\right\} \\ &= \frac{1}{T}E\left[\left\{\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\}'\left\{\frac{\partial \log L(Y;\theta_0)}{\partial\theta}\right\}\right] \\ &= \frac{1}{T}Q(\theta_0). \end{aligned}$$

Proof. It follows from the mean-value theorem that

$$\begin{aligned} \frac{\partial \log L(Y;\theta_0)}{\partial\theta} &= \frac{\partial \log L(Y;\hat{\theta})}{\partial\theta} \\ &\quad + (\theta_0-\hat{\theta})'\frac{\partial(\partial \log L(Y;\theta_*)/\partial\theta)'}{\partial\theta}, \end{aligned} \tag{25.35}$$

where θ_* is a value subject to the condition $\|\theta_*-\theta_0\| \leq \|\hat{\theta}-\theta_0\|$, which is to say that it lies between $\hat{\theta}$ and θ_0. By the definition of $\hat{\theta}$, we have $\partial \log L(Y;\hat{\theta})/\partial\theta = 0$, so the above expression can be rearranged to give

$$\sqrt{T}(\hat{\theta}-\theta_0) = -\left\{\frac{1}{T}\frac{\partial(\partial \log L(Y;\theta_*)/\partial\theta)'}{\partial\theta}\right\}^{-1}\left\{\frac{1}{\sqrt{T}}\frac{\partial \log L(Y;\hat{\theta})}{\partial\theta}\right\}'. \tag{25.36}$$

Now $\hat{\theta} \xrightarrow{P} \theta_0$, which denotes the consistency of the maximum-likelihood estimator, implies that $\theta_* \xrightarrow{P} \theta_0$. Therefore, in the limit, both factors on the RHS of (25.36) may be evaluated at θ_0; and we may use the following results:

(25.37) (i) By the law of large numbers, the term

$$\frac{1}{T}\frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial\theta} = \frac{1}{T}\sum_t \frac{\partial(\partial \log f(y_t;\theta_0)/\partial\theta)'}{\partial\theta}$$

converges to its expected value of M,

(ii) By the central limit theorem, the term

$$\frac{1}{\sqrt{T}}\frac{\partial \log L(Y;\theta_0)}{\partial\theta} = \frac{1}{\sqrt{T}}\sum_t \frac{\partial \log f(y_t;\theta_0)}{\partial\theta}$$

has a limiting normal distribution $N(0, M)$.

It follows immediately that $\sqrt{T}(\hat{\theta} - \theta_0)$ tends in distribution to a random variable $M^{-1}\eta$ where $\eta \sim N(0, M)$; and, therefore, we conclude that $\sqrt{T}(\hat{\theta} - \theta_0)$ has the limiting distribution $N(0, M^{-1})$. Equivalently, $\sqrt{T}(\hat{\theta} - \theta_0)$ tends in distribution to a random variable $\phi^\diamond = (Z'Z)^{-1}Z'\varepsilon$ where $\varepsilon \sim N(0, I)$ is a standard normal vector and where $Z'Z = M$. Finally, we may recognise that the equivalence of the two expressions for M follows from equation (25.18).

It is apparent that the asymptotic form of the maximum-likelihood estimator is identical to that of a least-squares regression estimator of the parameter ϕ in the distribution $N(\varepsilon; Z\phi, I)$. We can exploit this least-squares analogy to demonstrate that

(25.38) If $\hat{\theta}$ is the maximum-likelihood estimator obtained by solving the equation $\partial \log L(Y;\theta)/\partial\theta = 0$, and if θ_0 is the true parameter value, then the quantity

$$-\sqrt{T}(\hat{\theta} - \theta_0)'\left\{\frac{1}{T}\frac{\partial(\partial \log L(Y;\hat{\theta})/\partial\theta)'}{\partial\theta}\right\}\sqrt{T}(\hat{\theta} - \theta_0)$$

has a limiting distribution which is identical to that of the variate $\phi^{\diamond\prime}Z'Z\phi^\diamond = \varepsilon'Z(Z'Z)^{-1}Z'\varepsilon = \varepsilon'P\varepsilon \sim \chi^2(k)$.

This result can be used in testing an hypothesis relating to the vector θ_0. The theory of least-squares estimation, which is expounded in chapter 4, will help us to devise tests relating to subsets of the elements of θ_0.

Restricted Maximum-Likelihood Estimation

Often we wish to consider a model which can be expressed in terms of the likelihood function $L(Y;\theta)$ where $\theta \in \mathcal{R}^k$ is subject to a set of restrictions in the form of a vector function $r(\theta) = 0$ of $j < k$ elements. These restrictions will have the effect of confining θ to some subset $\mathcal{A} \subset \mathcal{R}^k$. One approach to estimating θ, which may be fruitful, is to reexpress the restrictions in the form of $\theta = \theta(\alpha)$ where α is an vector of $k - j$ unrestricted elements. Once we have an estimate $\hat{\alpha}$ of the unrestricted elements, we can obtain a restricted estimate of θ in the form of $\theta^* = \theta(\hat{\alpha})$. The alternative approach is to maximise the function $L(Y;\theta)$ with respect to θ subject to the restrictions. Our criterion function is then

$$L^* = \log L(Y;\theta) - \lambda' r(\theta), \tag{25.39}$$

where λ is a $j \times 1$ vector of Lagrangean multipliers corresponding to the j restrictions.

The first-order conditions for maximisation are

$$\begin{aligned} \frac{\partial \log L(Y;\theta)}{\partial \theta} - \lambda' R(\theta) &= 0, \\ r(\theta) &= 0, \end{aligned} \tag{25.40}$$

where $R(\theta) = \partial r(\theta)/\partial\theta$ is a $j \times k$ matrix of the derivatives of the restrictions with respect to the unknown parameters. The solution of the equations (25.40) is the restricted maximum-likelihood estimator θ^*. The equations are liable to be nonlinear so that, in order to investigate the properties of the estimator, we must rely upon a Taylor-series expansion to provide the appropriate linear approximation. As the sample size increases, the linear approximation should become increasingly valid.

Consider the following expansion about the true value θ_0 of the first derivative of the log-likelihood function at θ^* :

$$\begin{aligned} \frac{\partial \log L(Y;\theta^*)}{\partial \theta} = \frac{\partial \log L(Y;\theta_0)}{\partial \theta} \\ + (\theta^* - \theta_0)' \frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial \theta} + \zeta'. \end{aligned} \tag{25.41}$$

Here ζ stands for the higher-order terms. Also consider the expansion

$$\begin{aligned} r(\theta^*) &= r(\theta_0) + R(\theta_0)(\theta^* - \theta_0) - \xi \\ &= R(\theta_0)(\theta^* - \theta_0) - \xi. \end{aligned} \tag{25.42}$$

On substituting the RHS of (25.41) in place of $\partial \log L(Y;\theta)/\partial\theta$ in (25.40) and on dividing the resulting expressions by $\sqrt{T}$, we get, after some minor manipulations,

$$\begin{aligned} -\left\{ \frac{1}{T} \frac{\partial(\partial \log L(Y;\theta_0)/\partial\theta)'}{\partial \theta} \right\} \sqrt{T}(\theta^* - \theta_0) + R' \frac{\lambda}{\sqrt{T}} \\ = \frac{1}{\sqrt{T}} \left\{ \frac{\partial \log L(Y;\theta_0)}{\partial \theta} \right\}' + \frac{1}{\sqrt{T}} \zeta. \end{aligned} \tag{25.43}$$

When this is combined with the equation

$$\sqrt{T}R(\theta_0)(\theta^* - \theta_0) = \sqrt{T}\xi \tag{25.44}$$

which comes from (25.42), we obtain the following representation of the first-order conditions of (25.40):

$$\begin{aligned}\begin{bmatrix} -\dfrac{1}{T}\dfrac{\partial(\partial \log L(\theta_0)/\partial\theta)'}{\partial\theta} & R'(\theta_0) \\ R(\theta_0) & 0 \end{bmatrix} \begin{bmatrix} \sqrt{T}(\theta^* - \theta_0) \\ \dfrac{\lambda}{\sqrt{T}} \end{bmatrix} \\ = \begin{bmatrix} \dfrac{1}{\sqrt{T}} \left\{ \dfrac{\partial \log L(\theta_0)}{\partial\theta} \right\}' \\ 0 \end{bmatrix} + \begin{bmatrix} \dfrac{\zeta}{\sqrt{T}} \\ \sqrt{T}\xi \end{bmatrix}.\end{aligned} \tag{25.45}$$

As the sample size T increases, the terms involving the first and the second derivatives of the log-likelihood function tend to their probability limits. Given that the restricted estimate θ^* is consistent, the remainder terms $\zeta/\sqrt{T}$ and $\sqrt{T}\xi$ will tend in probability to zero. To find the limiting distribution of the estimator, we use again the two results under (25.37) concerning the central limit theorem and the law of large numbers. It follows that the vectors $\sqrt{T}(\theta^* - \theta_0)$ and $\lambda/\sqrt{T}$ have a limiting normal distribution which is identical to the distribution of the vectors ϕ^* and μ which are determined by the linear system

$$\begin{bmatrix} Z'Z & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \phi^* \\ \mu \end{bmatrix} = \begin{bmatrix} Z'\varepsilon \\ 0 \end{bmatrix}, \tag{25.46}$$

wherein Z is such that $Z'Z = M$ and $\varepsilon \sim N(0, I)$ is a vector with a standard normal distribution, and where $R = R(\theta_0)$.

The solutions for ϕ^* and μ are obtained from the equations

$$\begin{bmatrix} C_1 & C_2 \\ C_2' & C_3 \end{bmatrix} \begin{bmatrix} Z'\varepsilon \\ 0 \end{bmatrix} = \begin{bmatrix} \phi^* \\ \mu \end{bmatrix}. \tag{25.47}$$

The elements of the partitioned matrix are defined by the following identities:

$$\begin{aligned} &\text{(i)} \quad Z'ZC_1 + R'C_2' = I, && \text{(ii)} \quad Z'ZC_2 + R'C_3 = 0, \\ &\text{(iii)} \quad RC_1 = 0, && \text{(iv)} \quad RC_2 = I. \end{aligned} \tag{25.48}$$

From these conditions, we may easily obtain the following identities:

$$\begin{aligned} &\text{(i)} \quad C_1Z'ZC_1 = C_1, && \text{(ii)} \quad C_1Z'ZC_2 = 0, \\ &\text{(iii)} \quad C_2'Z'ZC_2 = -C_3. \end{aligned} \tag{25.49}$$

Using the latter, we may confirm that the dispersion matrix of ϕ^* and μ is given by

$$D\begin{bmatrix}\phi^* \\ \mu\end{bmatrix} = \begin{bmatrix}C_1 & C_2 \\ C_2' & C_3\end{bmatrix}\begin{bmatrix}Z'Z & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}C_1 & C_2 \\ C_2' & C_3\end{bmatrix} = \begin{bmatrix}C_1 & 0 \\ 0 & -C_3\end{bmatrix}. \tag{25.50}$$

Since the systems under (25.45) and (25.46) are equivalent asymptotically, we may draw the following conclusions:

(25.51) If θ^* is the restricted maximum-likelihood estimator and θ_0 is the true value of the parameter, then $\sqrt{T}(\theta^* - \theta_0)$ has a limiting normal distribution $N(0, C_1)$ which is the same as the distribution of the random variable $\phi^* = C_1 Z'\varepsilon \sim N(0, C_1)$. If λ is the Lagrangean multiplier associated the restrictions, then $\lambda/\sqrt{T}$ has a limiting normal distribution $N(0, -C_3)$ which is the same as the distribution of the random variable $\mu = C_2' Z'\varepsilon \sim N(0, -C_3)$.

We can exploit these results in order to establish an asymptotic result which relates the restricted and the unrestricted maximum-likelihood estimators. Consider the vectors $\phi^\diamond = (Z'Z)^{-1}Z'\varepsilon$ and $\phi^* = C_1 Z'\varepsilon$. From these we may construct

$$-Z(\phi^* - \phi^\diamond) = (P - ZC_1Z)\varepsilon, \tag{25.52}$$

where $P = Z(Z'Z)^{-1}Z$ is a symmetric idempotent matrix such that $P = P' = P^2$ and $PZ = Z$. We find that

$$\begin{aligned}(\phi^* - \phi^\diamond)'Z'Z(\phi^* - \phi^\diamond) &= \varepsilon'(P - ZC_1Z')'(P - ZC_1Z')\varepsilon \\ &= \varepsilon'(P - ZC_1Z')\varepsilon.\end{aligned} \tag{25.53}$$

Now consider the identity

$$\varepsilon' P\varepsilon = \varepsilon'(P - ZC_1Z')\varepsilon + \varepsilon' ZC_1Z'\varepsilon. \tag{25.54}$$

Since $P_\diamond = (P - ZC_1Z')$ and $P_* = ZC_1Z'$ are symmetric idempotent matrices with $P_\diamond P_* = 0$, and given that $\text{Rank}(P) = k$ and $\text{Rank}(ZC_1Z') = \text{Rank}(C_1) = j$, we can apply Cochran's theorem to show that equation (25.5) represents the decomposition of a chi-square variate. Thus

$$\begin{aligned}&\varepsilon'(P - ZC_1Z')\varepsilon \sim \chi^2(j), \\ &\varepsilon' ZC_1Z'\varepsilon \sim \chi^2(k - j), \\ &\varepsilon' P\varepsilon \sim \chi^2(k).\end{aligned} \tag{25.55}$$

We can conclude that

(25.56) If $\hat{\theta}$ and θ^* are, respectively, the restricted maximum-likelihood estimate and the unrestricted maximum-likelihood estimate, then the quantity

$$-\sqrt{T}(\theta^* - \hat{\theta})' \left\{ \frac{1}{T} \frac{\partial(\partial \log L(Y;\theta^*)/\partial\theta)'}{\partial\theta} \right\} \sqrt{T}(\theta^* - \hat{\theta})$$

has a limiting distribution which is identical to that of the variate $(\phi^* - \phi^\diamond)'Z'Z(\phi^* - \phi^\diamond) = \varepsilon'(P - ZC_1Z')\varepsilon \sim \chi^2(j)$.

Tests of the Restrictions

Three closely related methods are available for testing the hypothesis that $\theta_0 \in \mathcal{A}$, where $\mathcal{A} = \{\theta; r(\theta) = 0\}$ is the parameter set defined by the restrictions. These are the likelihood-ratio test, the Wald test and the Lagrangean-multiplier test. They are based respectively on the measures

$$-\sqrt{T}(\theta^* - \hat{\theta})' \left\{ \frac{1}{T} \frac{\partial(\partial \log L(\hat{\theta})/\partial\theta)'}{\partial\theta} \right\} \sqrt{T}(\theta^* - \hat{\theta}), \tag{25.57}$$

$$-\sqrt{T}r'(\hat{\theta}) \left[R(\hat{\theta}) \left\{ \frac{1}{T} \frac{\partial(\partial \log L(\hat{\theta})/\partial\theta)'}{\partial\theta} \right\}^{-1} R'(\hat{\theta}) \right]^{-1} \sqrt{T}r(\hat{\theta}) \tag{25.58}$$

and

$$\begin{aligned} &-\frac{\lambda'}{\sqrt{T}} R(\theta^*) \left\{ \frac{1}{T} \frac{\partial(\partial \log L(\theta^*)/\partial\theta)'}{\partial\theta} \right\}^{-1} R'(\theta^*) \frac{\lambda'}{\sqrt{T}} \\ &= -\frac{1}{\sqrt{T}} \left\{ \frac{\partial \log L(\theta^*)}{\partial\theta} \right\} \left\{ \frac{1}{T} \frac{\partial(\partial \log L(\theta^*)/\partial\theta)'}{\partial\theta} \right\}^{-1} \frac{1}{\sqrt{T}} \left\{ \frac{\partial \log L(\theta^*)}{\partial\theta} \right\}', \end{aligned} \tag{25.59}$$

wherein θ^* is the restricted maximum-likelihood estimator and $\hat{\theta}$ is the unrestricted estimator. These statistics are asymptotically equivalent and they share the same limiting distribution.

The ideas which give rise to these statistics are easily explained. The likelihood-ratio statistic in the form given under (25.57) embodies a measure of the proximity of the estimator θ^*, which incorporates the information of the restrictions, and the estimator $\hat{\theta}$, which freely reflects the information of the sample data in Y. If θ^* is remote from $\hat{\theta}$ then doubt will be cast upon the validity of restrictions. The limiting distribution of the statistic is given in (25.56) above.

The likelihood ratio itself, from which our statistic is derived remotely, is defined as

$$\kappa = \frac{\max\{\theta \in \mathcal{A}\} L(Y;\theta)}{\max\{\theta \in \mathcal{R}^k\} L(Y;\theta)} = \frac{L(Y;\theta^*)}{L(Y;\hat{\theta})}. \tag{25.60}$$

By taking the logarithm, we get

$$-2\log\kappa = 2\log L(Y;\hat{\theta}) - 2\log L(Y;\theta^*). \tag{25.61}$$

To show how this form relates to the measure under (25.57), we may take the Taylor's series expansion of $\log L(Y;\theta^*)$ about the point of the unrestricted estimator $\hat{\theta}$. This gives

$$\begin{aligned}\log L(Y;\theta^*) &\approx \log L(Y;\hat{\theta}) + \frac{\partial \log L(Y;\hat{\theta})}{\partial\theta}(\theta^* - \hat{\theta}) \\ &\quad + \frac{1}{2}(\theta^* - \hat{\theta})'\frac{\partial(\partial \log L(Y;\hat{\theta})/\partial\theta)'}{\partial\theta}(\theta^* - \hat{\theta}) \\ &\approx \log L(Y;\hat{\theta}) + \frac{1}{2}(\theta^* - \hat{\theta})'\frac{\partial(\partial \log L(Y;\hat{\theta})/\partial\theta)'}{\partial\theta}(\theta^* - \hat{\theta}).\end{aligned} \tag{25.62}$$

The second expression follows by virtue of the fact that $\partial \log L(Y;\hat{\theta})/\partial\theta = 0$, since $\hat{\theta}$ satisfies the first-order condition for maximising $\log L(Y;\theta)$. Hence

$$\begin{aligned}-2\log\kappa &\approx -(\theta^* - \hat{\theta})'\frac{\partial(\partial \log L(Y;\hat{\theta})/\partial\theta)'}{\partial\theta}(\theta^* - \hat{\theta}) \\ &\approx -\sqrt{T}(\theta^* - \hat{\theta})'\left\{\frac{1}{T}\frac{\partial(\partial \log L(Y;\hat{\theta})/\partial\theta)'}{\partial\theta}\right\}\sqrt{T}(\theta^* - \hat{\theta}).\end{aligned} \tag{25.63}$$

The Wald statistic under (25.58) measures the extent to which the unrestricted estimator $\hat{\theta}$ fails to satisfy the restrictions $r(\theta) = 0$. If its value is significant, then doubt will be cast, once more, upon the validity of the restrictions.

The Lagrange multiplier statistic uses λ to measure the strength of the constraint which must be imposed to ensure that the estimator θ^* obeys the restrictions. The alternative form of the statistic is obtained using the equality

$$\frac{\partial \log L(Y;\theta^*)}{\partial\theta} = \lambda' R(\theta^*), \tag{25.64}$$

which comes from the first-order conditions (25.40). The quantity $\partial \log L(Y;\theta)/\partial\theta$ is know as the score vector which accounts for the alternative description of the Lagrangean-multiplier statistic as the score statistic.

Our choice of a statistic for testing the validity of the restrictions will be influenced by the relative ease with which we can obtain the restricted and unrestricted estimates. If both $\hat{\theta}$ and θ^* are readily available, then we might use the likelihood-ratio statistic. If the unrestricted estimator $\hat{\theta}$ is available and we wish to test the validity of the restrictions $r(\theta) = 0$ before imposing them upon our estimates, then we should use the Wald statistic to perform a test of specification. If only the restricted estimator θ^* is available, then we should test the validity of the restrictions using the Lagrangean-multiplier statistic. This is a test of testing whether θ^* embodies a misspecification.

We wish to demonstrate that these three statistics are equivalent asymptotically and to show that they have the same limiting χ^2 distribution. To begin, let us recall that the limiting distribution of $\sqrt{T}(\hat{\theta} - \theta_0)$ is the same as the distribution of vector $\phi^\diamond = (Z'Z)^{-1}Z'\varepsilon$, and that the limiting distribution of $\sqrt{T}(\theta^* - \theta_0)$ is the same as the distribution of vector $\phi^* = C_1 Z'\varepsilon$. Then it is straightforward to demonstrate the following:

(25.65) (i) The likelihood ratio under (25.57) has a limiting distribution which is identical to that of $(\phi^* - \phi^\diamond)'Z'Z(\phi^* - \phi^\diamond)$,

(ii) The Wald statistic under (25.58) has a limiting distribution which is identical to that of $\phi^{\diamond\prime}R'\{R(Z'Z)^{-1}R'\}^{-1}\phi^\diamond$,

(iii) The Lagrange multiplier statistic under (25.59) has a limiting distribution which is identical to that of $\mu'R(Z'Z)^{-1}R'\mu$.

In order to demonstrate the asymptotic equivalence of the three statistics, it only remains to show that

$$\begin{aligned}(\phi^* - \phi^\diamond)'Z'Z(\phi^* - \phi^\diamond) &= -\phi^{\diamond\prime}R'C_3R\phi^\diamond \\ &= -\mu'C_3^{-1}\mu,\end{aligned} \tag{25.66}$$

and that

$$-C_3 = \{R(Z'Z)^{-1}R'\}^{-1}. \tag{25.67}$$

To demonstrate the equalities in (25.66), we make use of the identities in (25.48). First, we may postmultiply (25.48)(ii) by R and transpose the result to give

$$R'C_3R = -R'C_2'Z'Z. \tag{25.68}$$

Next, by postmultiplying (25.48)(i) by $Z'Z$ and rearranging, we get

$$Z'(I - ZC_1Z')Z = R'C_2Z'Z. \tag{25.69}$$

Taking these two results together, we get

$$-R'C_3R = Z'(I - ZC_1Z')Z. \tag{25.70}$$

Now, from (25.47), we get $\phi^* = C_1Z'\varepsilon$ and we also have $\phi^\diamond = (Z'Z)^{-1}Z'\varepsilon$; so, using (25.70), we can establish the first equality in (25.66).

To help in establishing the second equality of (25.66), we premultiply the expression in (25.48)(ii) by $Z'(Z'Z)^{-1}$ and transpose the result to give

$$C_2'Z' = -C_3R(Z'Z)^{-1}Z'. \tag{25.71}$$

Using this result in the expression for μ given by (25.47), we find that

$$\begin{aligned}\mu &= C_2'Z'\varepsilon \\ &= -C_3R(Z'Z)^{-1}Z'\varepsilon \\ &= -C_3R\phi^\diamond.\end{aligned} \tag{25.72}$$

The second equality follows immediately.

Finally, we must demonstrate the identity of (25.67). For this, we premultiply (25.48)(ii) by $R(Z'Z)^{-1}$ to give

$$RC_2 + \{R(Z'Z)^{-1}R'\}C_3 = 0. \tag{25.73}$$

The result follows from using $RC_2 = I$ from (25.48)(iv).

Having established that the three statistics are asymptotically equivalent, it remains to determine their common limiting distribution. We know that the $j \times 1$ vector μ of (25.47) has the distribution $N(0, -C_3)$. Therefore it follows that

$$-\mu' C_3^{-1} \mu \sim \chi^2(j). \tag{25.74}$$

Thus the limiting distribution of the three statistics is a chi-square with j degrees of freedom.

Bibliography

[6] Aitchison, J., and S.D. Silvey, (1958), Maximum-Likelihood Estimation of Parameters Subject to Restraints, *Annals of Mathematical Statistics,* **29,** 813–828.

[7] Aitchison, J., and S.D. Silvey, (1960), Maximum-Likelihood Estimation Procedures and Associated Tests of Significance, *Journal of the Royal Statistical Society, Series B,* **22,** 154–171.

[33] Bar-Shalom, Y., (1971), On the Asymptotic-Likelihood Estimate Obtained from Dependent Observations, *Journal of the Royal Statistical Society, Series B,* **33,** 72–77.

[57] Bhat, B.R., (1974), On the Method of Maximum Likelihood for Dependent Observations, *Journal of the Royal Statistical Society, Series B,* **36,** 48–53.

[101] Chernoff, H., (1954), On the Distribution of the Likelihood Ratio, *Annals of Mathematical Statistics,* **25,** 573–578.

[214] Godfrey, L.G., (1988), *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and other Approaches, Econometric Society Monographs No. 16,* Cambridge University Press, Cambridge.

[265] Huzurbazar, V.S., (1948), The Likelihood Equation, Consistency and the Maxima of the Likelihood Function, *Annals of Eugenics,* **14,** 185–200.

[420] Rao, C.R., (1961), Apparent Anomalies and Irregularities in Maximum Likelihood Estimation, *Bulletin, Institut International de Statistique,* **38,** 439–453.

[421] Rao, C.R., (1973), *Linear Statistical Inference and its Applications, Second Edition,* John Wiley and Sons, New York.

[460] Silvey, S.D., (1961), A Note on Maximum Likelihood in the Case of Dependent Random Variables, *Journal of the Royal Statistical Society, Series B,* **23,** 444–452.

[461] Silvey S.D., (1970), *Statistical Inference,* Chapman Hall, London.

[502] Wald, A., (1943), Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large, *Transactions of the American Mathematical Society,* **54,** 462–482.

[503] Wald, A., (1949), A Note on the Consistency of the Maximum Likelihood Estimator, *Annals of Mathematical Statistics,* **20,** 595–601.