APPENDIX 1

The Method of Maximum Likelihood

The Method of Maximum Likelihood

The method of maximum-likelihood constitutes a principle of estimation which can 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 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 method will be applied to 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

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

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

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

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

$$f(y_1, \ldots, y_T | x_1, \ldots, x_T) = (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \alpha - \beta x_t)\right\}. \tag{3}$$

The principle of maximum likelihood suggests that α, β and σ^2 should be estimated by choosing the values which maximise the probability measure that 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 that 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, it called a likelihood function instead and it is denoted it by $L(\alpha, \beta, \sigma^2)$.

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

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

It is clear that, given the value of σ^2, the likelihood is maximised by the values $\hat{\alpha}$ and $\hat{\beta}$ that minimise the sum of squares; and expressions for $\hat{\alpha}$ and $\hat{\beta}$ have been given already under (1.42) and (1.45) respectively.

The maximum-likelihood estimator for σ^2 can be obtained from the following first-order condition:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2 = 0. \tag{5}$$

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

$$\sigma^2(\alpha, \beta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2. \tag{6}$$

By putting $\hat{\alpha}$ and $\hat{\beta}$ in place, the estimator $\tilde{\sigma}^2 = \sigma^2(\hat{\alpha}, \hat{\beta}) = T^{-1}\sum e_t$ is obtained, whic had been given already under (1.46).

The General Theory of M-L Estimation

In order to derive an M-L estimator, it is necessary 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 that 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

$$L(y;\theta) = \prod_{t=1}^{T} f(y_t;\theta), \tag{7}$$

and therefore

$$\frac{1}{T}\log L(y;\theta) = \frac{1}{T}\sum_{t=1}^{T}\log f(y_t;\theta). \tag{8}$$

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

$$\operatorname{plim}(T\to\infty)\frac{1}{T}\log L(y;\theta) = E\big\{\log f(y_t;\theta)\big\}. \tag{9}$$

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

$$\frac{d\log L(y;\theta)}{d\theta} = \frac{1}{L(y;\theta)}\frac{dL(y;\theta)}{d\theta}. \tag{10}$$

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 (10),

$$\frac{d}{d\theta}E\big\{\log L(y;\theta)\big\} = \int_y \left\{\frac{1}{L(y;\theta)}\frac{dL(y;\theta)}{d\theta}\right\} L(y;\theta_0)dy, \tag{11}$$

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

$$\int_y \frac{dL(y;\theta_0)}{d\theta}dy = \frac{d}{d\theta}\int_y L(y;\theta_0)dy = 0, \tag{12}$$

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

$$\frac{d}{d\theta}E\big\{\log L(y;\theta_0)\big\} = E\left\{\frac{d\log L(y;\theta_0)}{d\theta}\right\} = 0; \tag{13}$$

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 (8) in respect to θ and take expectations. Provided that the order of these operations can be interchanged, then

$$\frac{d}{d\theta}\int_y \frac{d\log L(y;\theta)}{d\theta}L(y;\theta)dy = \frac{d^2}{d\theta^2}\int_y L(y;\theta)dy = 0, \tag{14}$$

where the final equality follows in the same way as that of (11). The LHS can be expressed as

$$\int_y \frac{d^2\log L(y;\theta)}{d\theta^2}L(y;\theta)dy + \int_y \frac{d\log L(y;\theta)}{d\theta}\frac{dL(y;\theta)}{d\theta}dy = 0 \tag{15}$$

and, on substituting from (11) into the second term, this becomes

$$\int_y \frac{d^2\log L(y;\theta)}{d\theta^2}L(y;\theta)dy + \int_y \left\{\frac{d\log L(y;\theta)}{d\theta}\right\}^2 L(y;\theta)dy = 0. \tag{16}$$

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

$$E\left\{-\frac{d^2 \log L(y;\theta_0)}{d\theta^2}\right\} = E\left[\left\{\frac{d \log L(y;\theta_0)}{d\theta}\right\}^2\right] = \Phi. \tag{17}$$

This measure is know as Fisher's Information. Since (12) 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:

$$\begin{aligned}\log L(\theta) = \log L(\theta_0) + \frac{d \log L(\theta_0)}{d\theta}(\theta - \theta_0) + \frac{1}{2}\frac{d^2 \log L(\theta_0)}{d\theta^2}(\theta - \theta_0)^2 \\ + \frac{1}{3!}\frac{d^3 \log L(\theta_0)}{d\theta^3}(\theta - \theta_0)^3 + \cdots.\end{aligned} \tag{18}$$

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

$$\log L(\theta) = \log L(\theta_0) + \frac{d\log L(\theta_0)}{d\theta}(\theta - \theta_0) + \frac{1}{2}\frac{d^2\log L(\theta_0)}{d\theta^2}(\theta - \theta_0)^2. \tag{19}$$

Its derivative with respect to θ is

$$\frac{d\log L(\theta)}{d\theta} = \frac{d\log L(\theta_0)}{d\theta} + \frac{d^2\log L(\theta_0)}{d\theta^2}(\theta - \theta_0). \tag{20}$$

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

$$\sqrt{T}(\hat{\theta} - \theta_0) = \left\{-\frac{1}{T}\frac{d^2\log L(\theta_0)}{d\theta^2}\right\}^{-1}\left\{\frac{1}{\sqrt{T}}\frac{d\log L(\theta_0)}{d\theta}\right\}. \tag{21}$$

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

$$-\frac{1}{T}\frac{d^2\log L(y;\theta_0)}{d\theta^2} = -\frac{1}{T}\sum_t \frac{d^2\log f(y_t;\theta_0)}{d\theta^2} \tag{22}$$

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

$$\frac{1}{\sqrt{T}}\frac{d\log L(y;\theta_0)}{d\theta} = \frac{1}{\sqrt{T}}\sum_t \frac{d\log f(y_t;\theta_0)}{d\theta} \tag{22}$$

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:

(23) 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

$$E\left\{-\frac{\partial^2 \log L(\theta_0)}{\partial\theta_i \partial\theta_j}\right\} = E\left\{\frac{\partial \log L(\theta_0)}{\partial\theta_i}.\frac{\partial \log L(\theta_0)}{\partial\theta_j}\right\}. \tag{24}$$