2 The Maximum Likelihood Estimation Method: Practical Issues

> *Part of the attraction of the ML theory is that it does offer estimator and test procedures that are technically almost universally applicable, provided one has a reasonably precise model. . . . There is a much better reason for using likelihood theory in that it provides a coherent framework for statistical inference in general.*
> —Cramer 1986, p. 8

The maximum likelihood estimation method is a way of obtaining estimators of a model when a specific distributional assumption is made about the vector of sample observations. Unlike least squares methods which use only the first two moments, maximum likelihood incorporates all the information in a model by working with the complete joint distribution of the observations.

Since the classical inferences for the models in chapters 3 through 6 depend heavily on maximum likelihood estimation, we provide a review of the method with a focus on related practical issues. For a more complete analysis of the maximum likelihood estimation method, readers are referred to Cramer 1986, Judge et al. 1982, Davidson and MacKinnon 1993, and Harvey 1990, from which this chapter draws.

2.1 Maximum Likelihood Estimation and the Covariance Matrix of $\hat{\theta}_{ML}$

A statistical model with a k-dimensional vector of parameters, θ, specifies a joint distribution for a vector of observations $\tilde{y}_T = [y_1 \quad y_2 \quad \ldots \quad y_T]'$:

$$\text{Joint Density Function:} \quad p(\tilde{y}_T|\theta), \tag{2.1}$$

which, in the discrete case, provides us with the probabilities of obtaining a particular set of values for $\tilde{y}_T$, given θ. The joint density therefore is a function of $\tilde{y}_T$ given θ.

In econometric practice, we have a realization of the $\tilde{y}_T$ vector, or the sample data, and we do not know the parameter vector θ of the underlying statistical model. In this case, the joint density in (2.1) is a function of θ given $\tilde{y}_T$, and it is called the likelihood function:

$$\text{Likelihood Function:} \quad L(\theta|\tilde{y}_T), \tag{2.2}$$

which is functionally equivalent to (2.1). Different values for θ result in different values for the likelihood function in (2.2). The likelihood function specifies the plausibility or likelihood of the data given the parameter vector θ. In the maximum likelihood estimation method, we are interested in choosing parameter estimates so as to maximize the probability of having generated the

Coherent : logical and well organized

observed sample, by maximizing the log of the above likelihood function:

$$\hat{\theta}_{ML} = \text{Argmax} \ln L(\theta|\tilde{y}_T), \tag{2.3}$$

where ln L refers to the log likelihood function.

Maximizing the log likelihood function instead of the likelihood function itself enables us to estimate directly the asymptotic covariance matrix, $\text{Cov}(\hat{\theta}_{ML})$, of the maximum likelihood estimate, $\hat{\theta}_{ML}$. The expectation of the second derivatives of the log likelihood function provides us with the information matrix $I(\theta)$:

$$I(\theta) = -E\left[\frac{\partial^2 \ln L(\theta|\tilde{y}_T)}{\partial\theta\partial\theta'}\right], \tag{2.4}$$

which summarizes the amount of information in the sample. The inverse of this information matrix provides us with the lower bound for the covariance matrix of an unbiased estimator $\tilde{\theta}$, known as the Cramer-Rao inequality:

$$\text{Cov}(\tilde{\theta}) - I(\theta)^{-1} \quad \text{is positive semidefinite.} \tag{2.5}$$

In addition, it can be shown that the maximum likelihood estimator $\hat{\theta}_{ML}$ has the following asymptotic normal distribution:

$$\sqrt{T}(\hat{\theta}_{ML} - \theta) \longrightarrow N(0, (\bar{H})^{-1}), \tag{2.6}$$

where

$$-\frac{1}{T}\frac{\partial^2 lnL(\theta|\tilde{y}_T)}{\partial\theta\partial\theta'} \to \bar{H} = \lim \frac{1}{T}I(\theta).$$

For an easy proof of the Cramer-Rao inequality and the asymptotic normality of the maximum likelihood estimator, refer to Harvey 1990. Equation (2.6) suggests that the maximum likelihood estimator is consistent and asymptotically efficient in the sense that its covariance matrix reaches the Cramer-Rao lower bound. Equation (2.6) also provides us with an idea of how to estimate the covariance matrix of the maximum likelihood estimator, using the inverse of the negative of the second derivative of the log likelihood function (Hessian) evaluated at $\hat{\theta}_{ML}$:

$$\text{Cov}(\hat{\theta}_{ML}) = \left[-\frac{\partial^2 \ln L(\theta|\tilde{y}_T)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}_{ML}}\right]^{-1} \tag{2.7}$$

2.2 The Prediction Error Decomposition and the Likelihood Function

For maximum likelihood estimation, we need to derive the joint density function or the likelihood function (they are functionally equivalent) for the vector $\tilde{y}_T$, given a statistical model. For independent observations, its derivation is straightforward:

$$L(\theta \mid \tilde{y}_T) = \prod_{t=1}^{T} p(y_t \mid \theta), \tag{2.8}$$

where $p(y_t \mid \theta)$ is the marginal density of an individual observation. For dependent observations, products of the conditional densities allow us to achieve the same goal:

$$L(\theta \mid \tilde{y}_T) = \prod_{t=2}^{T} p(y_t \mid \tilde{y}_{t-1}, \theta)p(y_1 \mid \theta), \tag{2.9}$$

where $\tilde{y}_t = [\ y_1 \ \ \ldots \ \ y_t\ ]$, $p(y_t \mid \tilde{y}_{t-1}, \theta)$, $t = 2, 3, \ldots, T$, is the conditional density, and $p(y_1 \mid \theta)$ is the marginal density of y_1. Notice that for the first observation, we have no information on which to condition.

However, the derivation of the conditional densities used in (2.9) for dependent observations may not always be straightforward. Consider, for example, the following unobserved-components model with normality assumptions:

$$y_t = x_t + e_t, e_t \sim \text{i.i.d.}N(0, \sigma_e^2) \quad t = 1, 2, \ldots, T, \tag{2.10}$$

$$x_t = \delta + \phi x_{t-1} + v_t, v_t \sim \text{i.i.d.}N(0, \sigma_v^2), \tag{2.11}$$

where e_t and v_t are independent and $\mid \phi \mid < 1$. The conditional density $p(y_t \mid \tilde{y}_{t-1}, \theta)$ is not directly obtained from the statistical model, where $\theta = [\delta \quad \sigma_e^2 \quad \sigma_v^2 \quad \phi]'$ in the present case. Taking advantage of the normality assumption, the vector of observations $\tilde{y}_T$ can be represented by the following multivariate normal distribution:

$$\tilde{y}_T \sim N(\mu, \Omega), \tag{2.12}$$

with the likelihood function:

$$L(\theta \mid \tilde{y}_T) = (2\pi)^{-\frac{T}{2}} \mid \Omega \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\tilde{y}_T - \mu)'\Omega^{-1}(\tilde{y}_T - \mu)\}, \tag{2.13}$$

where all elements of μ and Ω are complicated functions of δ, σ_e, σ_v^2, and ϕ. Even when μ and Ω can be specified explicitly, maximizing the log of the likelihood function with respect to unknown parameters would be troublesome because of the inversion of the $T \times T$ matrix Ω. Harvey (1980) provides a solution to these difficulties, based on the prediction error decomposition obtained from applying the triangular factorization of the Ω matrix in (2.12).

Note that, for the $T \times T$ positive-definite, symmetric matrix Ω, there exists a unique triangular factorization of the following form:

$$\Omega = AfA', \tag{2.14}$$

where f is a diagonal matrix with positive elements and A is a lower triangular matrix with the following forms:

$$f = \begin{bmatrix} f_1 & 0 & 0 & \dots & 0 \\ 0 & f_2 & 0 & \dots & 0 \\ 0 & 0 & f_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & f_T \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ a_{21} & 1 & 0 & \dots & 0 \\ a_{31} & a_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{T1} & a_{T2} & a_{T3} & \dots & 1 \end{bmatrix},$$

and where $f_t > 0$ for all t. Substituting (2.14) into (2.13), we have:

$$\begin{aligned} L(\theta \mid \tilde{y}_T) &= (2\pi)^{-\frac{T}{2}} \mid AfA' \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\tilde{y}_T - \mu)'(AfA')^{-1}(\tilde{y}_T - \mu)\} \\ &= (2\pi)^{-\frac{T}{2}} \mid f \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}\eta' f^{-1}\eta\} \\ &= (2\pi)^{-\frac{T}{2}} \prod_{t=1}^{T} f_t^{-\frac{1}{2}} \exp\{-\frac{1}{2}\sum_{t=1}^{T}\eta_t' f_t^{-1}\eta_t\} \\ &= \prod_{t=1}^{T}\left[\frac{1}{\sqrt{2\pi f_t}} \exp\{-\frac{1}{2}\frac{\eta_t^2}{f_t}\}\right], \end{aligned} \tag{2.15}$$

where $\eta = A^{-1}(\tilde{y}_T - \mu)$ and η_t is the t-th element of the $T \times 1$ vector η. Because A is a lower trangular with ones in its diagonal elements, one can easily show that the t-th element of η can be rewritten as:

$$\eta_t = y_t - y_{t|t-1}, \tag{2.16}$$

where $y_{t|t-1}$ is the prediction of y_t conditional on $\tilde{y}_{t-1} = [\, y_1 \quad \dots \quad y_{t-1}\,]'$, which is information up to $t-1$, since we have:

$$y_{t|t-1} = \sum_{i=1}^{t-1} a_{t,i}^{*} y_i, \quad t = 2, 3, \dots, T, \tag{2.17}$$

where $a_{t,i}^{*}$ is the (t, i)-th element of A^{-1}. Notice that the argument in the bracket of the last line in (2.15) is a normal density function of y_t conditional on past information:

$$y_t \mid \tilde{y}_{t-1} \sim N(y_{t|t-1}, f_t), \tag{2.18}$$

where f_t is interepreted as the variance of the prediction error $\eta_t = y_t - y_{t|t-1}$.

To summarize, equations (2.15) and (2.18) suggest that when the observations are normally distributed, insofar as we have the prediction errors and their variances, the log likelihood value can be easily calculated. Thus, the success of maximum likelihood estimation for a complicated dynamic time series model with dependent observations may depend on the availability of the prediction errors and their variances. For the unobserved component model specified in equations (2.10) and (2.11), for example, the Kalman filter introduced in chapter 3 provides us with η_t and f_t used in the last line of equation (2.15). With normality assumptions, equation (2.15) is a general approach to deriving the likelihood function for dependent observations of a dynamic time series. For a multivariate case, it is easy to see that (2.15) is replaced by:

$$L(\theta \mid \tilde{y}_T) = \prod_{t=1}^{T}\left[\frac{1}{\sqrt{(2\pi)^n |f_t|}} \exp\{-\frac{1}{2}\eta_t' f_t^{-1}\eta_t\}\right], \tag{2.15$'$}$$

where n is the dimension of y_t, η_t is $n \times 1$, and f_t is $n \times n$.

2.3 Parameter Constraints and the Covariance Matrix of $\hat{\theta}_{ML}$

2.3.1 Constrained Optimization

The maximum likelihood estimator, $\hat{\theta}_{ML}$, can be obtained by setting the first derivative of the log likelihood function to 0:

$$\frac{\partial \ln L(\theta \mid \tilde{y}_T)}{\partial \theta} = 0. \tag{2.19}$$

In most cases, however, a closed-form solution for $\hat{\theta}_{ML}$ is not available. Thus, in general, we resort to a nonlinear numerical optimization procedure to maximize the log likelihood function. Given initial estimates (θ^{j-1}) of the parameters, new estimates (θ^{j}) are obtained using the information provided by the first derivatives (and sometimes, depending upon the algorithms employed, the second derivatives) of the log likelihood function evaluated at θ^{j-1}. New estimates are obtained such that the log likelihood value evaluated at the revised estimates is larger than that at the initial estimates. This process may be iterated until convergence is achieved to obtain the value of the parameters that maximize the log likelihood function. In some cases, the maximum may not be unique. For specific and easy expositions of various algorithms for numerical optimization, readers are referred to chapter 4 of Harvey 1990.

When numerical optimization is employed to maximize the log likelihood function with respect to θ, the computer searches over the parameter space that ranges between negative infinity and positive infinity. But some of the parameters may have to be constrained to lie in an interval. For example, if one of the elements in θ is a probability (p), then it must be constrained such that $0 < p < 1$. In general, such constraints may be imposed by the following transformations of a vector ψ that ranges between negative infinity and positive infinity:

$$\theta = g(\psi), \tag{2.20}$$

where $g(.)$ is a continuous function. Then the log likelihood function may be considered a function of ψ:

$$\ln L(\theta \mid \tilde{y}_T) = \ln L(g(\psi) \mid \tilde{y}_T) = \ln L(\psi \mid \tilde{y}_T), \tag{2.21}$$

and the unconstrained numerical optimization may be applied with respect to ψ.

For example, if θ_j, the j-th element of θ, represents a variance, then $\theta_j > 0$. Then we may use the transformations

$$\theta_j = \psi_j^2 \quad \text{or} \quad \theta_j = \exp(\psi_i).$$

If θ_j represents a probability term, then $0 < \theta_j < 1$. The transformation we may employ is

$$\theta_j = \frac{1}{1 + \exp(\psi_j^{-1})}.$$

If θ_j represents an autoregressive parameter in an AR(1) model, then we may want to constrain the parameter within the stationary region $-1 < \psi_j < +1$:

$$\theta_j = \frac{\psi_j}{1 + \mid \psi_j \mid}.$$

If $\theta = [\, \phi_1 \quad \phi_2 \,]'$, where ϕ_1 and ϕ_2 are the autoregressive coefficients of the model in an AR(2) model, we may want to constrain the values of ϕ_1 and ϕ_2 within the stationary region (roots of $(1 - \phi_1 L - \phi_2 L^2) = 0$ lie outside the unit circle). In this case, we may employ the transformation

$$z_1 = \frac{\psi_1}{1 + \mid \psi_1 \mid}, \qquad z_2 = \frac{\psi_2}{1 + \mid \psi_2 \mid},$$

$$==> \phi_1 = z_1 + z_2, \qquad \phi_2 = -1 * z_1 * z_2.$$

Notice, however, that the recommended procedure in fact imposes the further restriction that the roots of the AR(2) polynomial are real. Finally, consider the following generalized autoregressive conditional hetereskedasticity (GARCH)(1,1) model:

$$h_t = a_0 + a_1 e_{t-1}^2 + a_2 h_{t-1}.$$

We generally want $a_1 > 0$, $a_2 > 0$, and $0 < a_1 + a_2 < 1$. The following transformations achieve this goal:

$$a_1 = \frac{\exp(\psi_1)}{1 + \exp(\psi_1) + \exp(\psi_2)}, \qquad a_2 = \frac{\exp(\psi_2)}{1 + \exp(\psi_1) + \exp(\psi_2)}.$$

2.3.2 Constrained Optimization and the Covariance Matrix of $\hat{\theta}_{ML}$

In section 2.3.1, we noted that applying *unconstrained* optimization to the log likelihood function in (2.18) with respect to ψ is equivalent to applying *constrained* optimization with respect to θ, the parameter of interest to us. Unconstrained optimization then results in $\hat{\psi}_{ML}$ and $\text{Cov}(\hat{\psi}_{ML})$, the maximum likelihood (ML) estimate of ψ and its covariance matrix. But we actually want the parameter estimates and the covariance matrix for θ. As $\theta = g(\psi)$, the ML estimate for θ is easily obtained by

$$\hat{\theta}_{ML} = g(\hat{\psi}_{ML}). \tag{2.22}$$

We can also obtain $\mathrm{Cov}(\hat{\theta}_{ML})$ based on $\mathrm{Cov}(\hat{\psi}_{ML})$ and $g(.)$ in the following way:

$$\mathrm{Cov}(\hat{\theta}_{ML}) = \left(\frac{\partial g(\hat{\psi}_{ML})}{\partial \psi}\right) \mathrm{Cov}(\hat{\psi}_{ML}) \left(\frac{\partial g(\hat{\psi}_{ML})}{\partial \psi}\right)'. \tag{2.23}$$

The following provides a proof of equation (2.23). Differentiating the log likelihood function $\ln L(\theta) = \ln L(g(\psi))$ with respect to ψ, we get

$$\frac{\partial \ln L(g(\psi))}{\partial \psi} = \frac{\partial \ln L(g(\psi))}{\partial \theta}\left(\frac{\partial g(\psi)}{\partial \psi}\right), \tag{2.24}$$

and then differentiating again,

$$\begin{aligned}&\left(\frac{\partial^2 \ln L(g(\psi))}{\partial \psi \partial \psi'}\right)\\ &= \left(\frac{\partial g(\psi)}{\partial \psi}\right)' \frac{\partial^2 \ln L(g(\psi))}{\partial \theta \partial \theta'}\left(\frac{\partial g(\psi)}{\partial \psi}\right) + \frac{\partial \ln L(g(\psi))}{\partial \psi}\left(\frac{\partial^2 g(\psi)}{\partial \psi \partial \psi'}\right).\end{aligned} \tag{2.25}$$

As

$$\frac{\partial \ln L(g(\psi_{ML}))}{\partial \theta} = 0,$$

(2.25) is written as

$$\begin{aligned}&\left(\frac{\partial^2 \ln L(g(\hat{\psi}_{ML}))}{\partial \psi \partial \psi'}\right)\\ &= \left(\frac{\partial g(\hat{\psi}_{ML})}{\partial \psi}\right)' \frac{\partial^2 \ln L(g(\hat{\psi}_{ML}))}{\partial \theta \partial \theta'}\left(\frac{\partial g(\hat{\psi}_{ML})}{\partial \psi}\right)\end{aligned} \tag{2.26}$$

Multiplying both sides of (2.26) by -1 and then taking the inverse of both sides, we have

$$\begin{aligned}&\left(-\frac{\partial^2 \ln L(g(\hat{\psi}_{ML}))}{\partial \psi \partial \psi'}\right)^{-1}\\ &= \left(\frac{\partial g(\hat{\psi}_{ML})}{\partial \psi}\right)^{-1}\left(-\frac{\partial^2 \ln L(g(\hat{\psi}_{ML}))}{\partial \theta \partial \theta'}\right)^{-1}\left(\frac{\partial g(\hat{\psi}_{ML})'}{\partial \psi}\right)^{-1}.\end{aligned} \tag{2.27}$$

Arranging the terms in (2.27) and noting that

$$\mathrm{Cov}(\hat{\theta}_{ML}) = (-\frac{\partial^2 \ln L(\hat{\theta}_{ML})}{\partial \theta \partial \theta'})^{-1}$$

and

$$\mathrm{Cov}(\hat{\psi}_{ML}) = (-\frac{\partial^2 \ln L(\hat{\psi}_{ML})}{\partial \psi \partial \psi'})^{-1},$$

we get equation (2.23).

References

Cramer, J. S. 1986. *Econometric Applications of Maximum Likelihood Methods*, Cambridge: Cambridge University Press.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* Oxford, UK: Oxford University Press.

Harvey, Andrew C. 1981. *Time Series Models.* Oxford: Philip Allan and Humanities Press.

Harvey, Andrew C. 1990. *The Econometric Analysis of Time Series.* 2nd ed. Cambridge: MIT Press.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lutkepohl, and T.-C. Lee. 1982. *Introduction to the Theory and Practice of Econometrics.* New York: John Wiley & Sons.