Econ 620

Maximum Likelihood Estimation (MLE)

Definition of MLE

- Consider a parametric model in which the **joint distribution** of $Y = (y_1, y_2, \cdots, y_n)$ has a density $\ell(Y;\theta)$ with respect to a dominating measure μ, where $\theta \in \Theta \subset R^P$.

Definition 1 *A maximum likelihood estimator of θ is a solution to the maximization problem*

$$\max_{\theta\in\Theta} \ell(y;\theta)$$

- Note that the solution to an optimization problem is invariant to a strictly monotone increasing transformation of the objective function, a MLE can be obtained as a solution to the following problem;

$$\max_{\theta\in\Theta} \log \ell(y;\theta) = \max_{\theta\in\Theta} L(y;\theta)$$

Proposition 2 *(Sufficient condition for existence) If the parameter space Θ is compact and if the likelihood function $\theta \mapsto \ell(y;\theta)$ is continuous on Θ, then there exists a MLE.*

Proposition 3 *(Sufficient condition for uniqueness of MLE) If the parameter space Θ is convex and if the likelihood function $\theta \mapsto \ell(y;\theta)$ is strictly concave in θ, then the MLE is unique when it exists.*

- If the observations on Y are i.i.d. with density $f(y_i;\theta)$ for each observation, then we can write the likelihood function as

$$\ell(y;\theta) = \prod_{i=1}^{n} f(y_i;\theta) \Rightarrow L(y;\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$$

Properties of MLE

Proposition 4 *(Functional invariance of MLE) Suppose a bijective function $g : \Theta \to \Lambda$ where $\Lambda \subset R^q$ and $\widehat{\theta}$ is a MLE of θ, then $\widehat{\lambda} = g\left(\widehat{\theta}\right)$ is a MLE of $\lambda \in \Lambda$.*

$\Rightarrow$ By definition of MLE, we have

$$\widehat{\theta} \in \Theta \text{ and } \ell\left(y;\widehat{\theta}\right) \geq \ell(y;\theta), \forall\theta \in \Theta$$

or equivalently,

$$\widehat{\lambda} \in \Lambda \text{ and } \ell\left(y;g^{-1}\left(\widehat{\lambda}\right)\right) \geq \ell\left(y;g^{-1}(\lambda)\right), \forall\lambda \in \Lambda$$

which implies that $\widehat{\lambda} = g\left(\widehat{\theta}\right)$ is a MLE of λ in a model with density $\ell\left(y;g^{-1}(\lambda)\right)$.

Proposition 5 *(Relationship with sufficiency) MLE is a function of every sufficient statistic.*

$\Rightarrow$ Let $S(Y)$ be a sufficient statistic. From the factorization theorem of a sufficient statistic, the density function can be written as $\ell(y;\theta) = \Psi(S(y);\theta)\, h(y)$, *i.e.*, $L(y;\theta) = \log\Psi(S(y);\theta) + \log h(y)$. Hence maximizing $\ell(y;\theta)$ with respect to θ is equivalent to maximizing $\log\Psi(S(y);\theta)$ with respect to θ. Therefore, MLE depends on Y through $S(Y)$.

- To discuss asymptotic properties of MLE, which are why we study and use MLE in practice, we need some so-called **regularity conditions**. These conditions are to be checked not to be granted before we use MLE. It is difficult, mostly impossible, to check in practice, though.

Regularity Conditions

1. The variables $Y_i, i = 1, 2, \cdots$ are independent and identically distributed with density $f(y;\theta)$.
2. The parameter space Θ is compact.
3. The true but unknown parameter value θ_0 is identified, i.e.
$$\theta_0 = \arg\max_{\theta\in\Theta} E_{\theta_0} \log f(Y_i;\theta)$$
4. The likelihood function
$$L(y;\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$$
is continuous in θ.
5. $E_{\theta_0} \log f(Y;\theta)$ exists.
6. The log-likelihood function is such that $\frac{1}{n}L(y;\theta)$ converges almost surely (in probability) to $E_{\theta_0} \log f(Y_i;\theta)$ **uniformly** in $\theta \in \Theta$, *i.e.*,
$$\sup_{\theta\in\Theta} \left| \frac{1}{n} L(y;\theta) - E_{\theta_0} \log f(Y_i;\theta) \right| < \delta \text{ almost surely (in probability) for some } \delta > 0.$$

Proposition 6 *Under 1 - 6, there exists a sequence of MLE's converging almost surely (in probability) to the true parameter value θ_0. That is, MLE is a consistent estimator.*

$\Rightarrow$ 1 and 2 ensure the existence of MLE $\widehat{\theta}_n$. It is obtained by maximizing $L(y;\theta)$ or equivalently, $\frac{1}{n}L(y;\theta)$. Since $\frac{1}{n}L(y;\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta)$ can be interpreted as the sample mean of the random variables $\log f(y_i;\theta)$, which are i.i.d., the objective function converges almost surely (in probability) to $E_{\theta_0} \log f(Y;\theta)$ by the strong(weak) law of large numbers. Furthermore, the uniform strong law of large numbers implies that the solution to $\frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta)$, $\widehat{\theta}_n$, converges to the solution to the limit problem

$$\max_{\theta\in\Theta} E_{\theta_0} \log f(Y;\theta)$$

i.e.,

$$\max_{\theta\in\Theta} \int_{\mathcal{Y}} \log f(y;\theta) f(y;\theta_0)\, dy$$

Now, note that the identifiability condition 3 ensures the convergence of $\widehat{\theta}_n$ to θ_0.

More regularity conditions for asymptotic distribution

2'. $\theta_0 \in Int(\Theta)$.

7. The log-likelihood function $L(y;\theta)$ is twice continuously differentiable in a neighborhood of θ_0.
8. Integration and differential operators are interchangeable.
9. The matrix
$$\mathcal{I}(\theta_0) = E_{\theta_0}\left(-\frac{\partial^2 \log f(Y;\theta_0)}{\partial\theta\partial\theta'} \right)$$
called information matrix, exists and non-singular.

- The additional assumptions enables us to use differential method to obtain MLE and its asymptotic distribution.

Lemma 7

$$E_{\theta_0} \frac{\partial \log f(Y;\theta_0)}{\partial\theta} = 0.$$

$\Rightarrow$

$$E_{\theta_0}\frac{\partial \log f(Y;\theta_0)}{\partial \theta} = \int \frac{\partial \log f(y;\theta_0)}{\partial \theta} f(y;\theta_0)\, dy$$
$$= \int \frac{1}{f(y;\theta_0)} \frac{\partial f(y;\theta_0)}{\partial \theta} f(y;\theta_0)\, dy = \int \frac{\partial f(y;\theta_0)}{\partial \theta} dy$$

However,

$$\int f(y;\theta_0)\, dy = 1 \text{ by definition.}$$

Hence, differentiating with respect to θ gives

$$\frac{\partial}{\partial \theta}\int f(y;\theta_0)\, dy = \int \frac{\partial f(y;\theta_0)}{\partial \theta} dy = 0$$

Lemma 8

$$E_{\theta_0}\left(\frac{\partial \log f(Y;\theta_0)}{\partial \theta}\frac{\partial \log f(Y;\theta_0)}{\partial \theta'}\right) = E_{\theta_0}\left(-\frac{\partial^2 \log f(Y;\theta_0)}{\partial \theta \partial \theta'}\right)$$

$\Rightarrow$

$$E_{\theta_0}\left(\frac{\partial^2 \log f(Y;\theta_0)}{\partial \theta \partial \theta'}\right)$$
$$= \int \frac{\partial^2 \log f(y;\theta_0)}{\partial \theta \partial \theta'} f(y;\theta_0)\, dy = \int \frac{\partial}{\partial \theta}\left(\frac{\partial \log f(y;\theta_0)}{\partial \theta'}\right) f(y;\theta_0)\, dy$$
$$= \int \frac{\partial}{\partial \theta}\left(\frac{1}{f(y;\theta_0)}\frac{\partial f(y;\theta_0)}{\partial \theta'}\right) f(y;\theta_0)\, dy$$
$$= \int \left[-\frac{1}{(f(y;\theta_0))^2}\frac{\partial f(y;\theta_0)}{\partial \theta}\frac{\partial f(y;\theta_0)}{\partial \theta'} + \frac{1}{f(y;\theta_0)}\frac{\partial^2 f(y;\theta_0)}{\partial \theta \partial \theta'}\right] f(y;\theta_0)\, dy$$
$$= -\int \left[\frac{1}{f(y;\theta_0)}\frac{\partial f(y;\theta_0)}{\partial \theta}\right]\left[\frac{1}{f(y;\theta_0)}\frac{\partial f(y;\theta_0)}{\partial \theta'}\right] f(y;\theta_0)\, dy + \int \frac{\partial^2 f(y;\theta_0)}{\partial \theta \partial \theta'} dy$$
$$= -\int \frac{\partial \log f(Y;\theta_0)}{\partial \theta}\frac{\partial \log f(Y;\theta_0)}{\partial \theta'} f(y;\theta_0)\, dy = -E_{\theta_0}\left(\frac{\partial \log f(Y;\theta_0)}{\partial \theta}\frac{\partial \log f(Y;\theta_0)}{\partial \theta'}\right)$$

The last line follows from the fact that $\int \frac{\partial^2 f(y;\theta_0)}{\partial \theta \partial \theta'} dy = 0$.

Proposition 9 *Under 1,2', 3 - 9, a sequence of MLE, $\widehat{\theta}_n$, satisfies*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

$\Rightarrow$ A Taylor series expansion of the first order condition around the true value of θ, θ_0, yields

$$\frac{\partial L\left(\widehat{\theta}_n\right)}{\partial \theta} = \frac{\partial L(\theta_0)}{\partial \theta} + \frac{\partial^2 L(\theta^*)}{\partial \theta \partial \theta'}\left(\widehat{\theta}_n - \theta_0\right)$$

where θ^* is on the line segment connecting $\widehat{\theta}_n$ and θ_0. From the first order condition, we have

$$0 = \frac{\partial L(\theta_0)}{\partial \theta} + \frac{\partial^2 L(\theta^*)}{\partial \theta \partial \theta'}\left(\widehat{\theta}_n - \theta_0\right)$$

Therefore,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -\left(\frac{1}{n}\frac{\partial^2 L(\theta^*)}{\partial \theta \partial \theta'}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial \theta}$$

As $n \to \infty$,

$$-\frac{1}{n}\frac{\partial^2 L(\theta^*)}{\partial \theta \partial \theta'} = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2 \log f(Y_i;\theta^*)}{\partial \theta \partial \theta'}$$

converges almost surely to

$$\mathcal{I}(\theta_0) = E_{\theta_0}\left(-\frac{\partial^2 \log f(Y;\theta_0)}{\partial\theta\partial\theta'}\right)$$

by the strong law of large numbers and the fact that $\theta^* \xrightarrow{a.s.} \theta_0$. Moreover,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial\theta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(Y;\theta_0)}{\partial\theta} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial \log f(Y;\theta_0)}{\partial\theta} - E_{\theta_0}\frac{\partial \log f(Y;\theta_0)}{\partial\theta}\right)\end{aligned}$$

which converges in distribution to

$$N(0, \mathcal{I}(\theta_0))$$

by the central limit theorem. We have used Lemma 7 and Lemma 8 here to get the asymptotic distribution of $\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial\theta}$. Then,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

- The asymptotic distribution, itself is useless since we have to evaluate the information matrix at true value of parameter. However, we can consistently estimate the asymptotic variance of MLE by evaluating the information matrix at MLE, i.e.,

  $$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}\left(\widehat{\theta}_n\right)^{-1}\right)$$

  In other expression which is slightly misleading but commonly used in practice is

  $$\widehat{\theta}_n \xrightarrow{d} N\left(\theta_0, \frac{1}{n}\mathcal{I}\left(\widehat{\theta}_n\right)^{-1}\right) = N\left(\theta_0, \left(n\mathcal{I}\left(\widehat{\theta}_n\right)\right)^{-1}\right)$$

  where $n\mathcal{I}\left(\widehat{\theta}_n\right) = -\frac{\partial^2 L\left(\widehat{\theta}_n\right)}{\partial\theta\partial\theta'}$. We can also use the approximation that

  $$n\mathcal{I}\left(\widehat{\theta}_n\right) = \sum_{i=1}^{n}\frac{\partial \log f\left(y_i;\widehat{\theta}_n\right)}{\partial\theta}\frac{\partial \log f\left(y_i;\widehat{\theta}_n\right)}{\partial\theta'}$$

Proposition 10 *Let g be a continuously differentiable function of $\theta \in R^p$ with values in R^q. Then, under the assumptions of Proposition 9,*

(i) $g\left(\widehat{\theta}_n\right)$ *converges almost surely to* $g(\theta_0)$.

(ii) $\sqrt{n}\left(g\left(\widehat{\theta}_n\right) - g(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{dg(\theta_0)}{d\theta'}\mathcal{I}(\theta_0)^{-1}\frac{dg(\theta_0)}{d\theta}\right)$

$\Rightarrow$ The first claim is straight application of Slutsky theorem. For the second claim, we do a Taylor expansion of $g\left(\widehat{\theta}_n\right)$ around θ_0 to get

$$g\left(\widehat{\theta}_n\right) = g(\theta_0) + \frac{dg(\theta^*)}{d\theta}\left(\widehat{\theta}_n - \theta_0\right)$$

Hence,

$$\sqrt{n}\left(g\left(\widehat{\theta}_n\right) - g(\theta_0)\right) = \frac{dg(\theta^*)}{d\theta}\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$$

Note that, as $n \to \infty$, we have

$$\begin{aligned}&\frac{dg(\theta^*)}{d\theta} \xrightarrow{a.s.} \frac{dg(\theta_0)}{d\theta} \text{ and} \\ &\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)\end{aligned}$$

The claim follows immediately.