# Chapter 1

# Maximum Likelihood Estimation

The word "estimation" is an euphemism of "educated guess". What is a reasonable way to make a guess? The maximum likelihood principle is arguably one of the the most important ideas in the history of mathematical statistics. It is intuitive at the first glance, but it took centuries to formalize. Milestones were set by Sir Ronald Fisher (1890–1962) and his contemporaries.

In econometrics, the most familiar OLS estimator is indeed a special case of MLE when the error term is normally distributed. When the support of the dependence variable is not continuous, most estimators are constructed by assuming an underlying conditional density model and then estimated by MLE. LIML (limited information maximum likelihood) is 2SLS counterpart in the likelihood world.

## 1.1 Parametric Model

A parametric model is a complete specification of the distribution. Once the parameter is given, the distribution function is determined. Instead, a semiparametric model only gives a few features rather than a complete description of the distribution.

**Example 1.1.** Semiparametric model: If we know $Y \sim i.i.d.\left(\mu, \sigma^2\right)$, we can estimate $\mu, \sigma^2$ by method of moments.

Parametric model: If we assume $Y \sim N\left(\mu, \sigma^2\right)$, the model has only two parameters $\mu$ and $\sigma^2$.

**Definition 1.1. Parametric model**. The distribution of the data $\mathbf{Y} = (Y_1, ..., Y_n)$ is known up to a finite dimensional parameter.

Let $\Theta$ be the parameter space a researcher specifies.

**Definition 1.2.** A model is **correctly specified**, if the true DGP is $f(\mathbf{Y} \mid \theta_0)$ for some $\theta_0 \in \Theta$. Otherwise, the model is **misspecified**.

## 1.2 Likelihood

In this chapter we will mostly talk about unconditional models. The results can be carried over to conditional models. To keep the setting simple, let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be i.i.d. The **likelihood** of the sample under a hypothesized value of $\theta \in \Theta$ is

$$L(\theta; \mathbf{Y}) = f(\mathbf{Y}; \theta) = \prod_{i=1}^{n} f(Y_i; \theta_0).$$

*Remark* 1.1. In the above expression $L(\theta; \mathbf{Y})$ and $f(\mathbf{Y}; \theta)$ are equal, but they have different interpretations. The former emphasizes that when the data $\mathbf{Y}$ is provided $L$ is a function of the parameter $\theta$. The latter is the opposite: given the parameter $\theta$, the joint density is a function of the realized value $\mathbf{Y}$.

*Remark* 1.2. In this note I adhere to the frequentist view where $\theta$ is an unknown constant, and I thus prefer using ";" to separate the two arguments. However, some textbooks (including Hansen's) write $L(\theta|\mathbf{Y})$ and $f(\mathbf{Y}|\theta)$, which I don't like because "|" delivers a confusing connotation of conditional density (for example $f(Y|X)$) as if $\theta$ is treated as a random variable. I would adopt "|" only in the Bayesian world, where the posterior density of the random $\theta$ is updated via the Bayes Theorem

$$L(\theta|\mathbf{Y}) = \frac{f(\mathbf{Y}|\theta)\,\pi(\theta)}{f(\mathbf{Y})},$$

give the prior distribution $\pi(\theta)$.

**Example 1.2.** Conditional model: the conditioning variable can be viewed as if it is fixed and the randomness comes from the error term only. In the linear regression model $Y_i = X_i'\beta + \varepsilon_i$, the conditioning variable is $X_i$. The condition $E[\varepsilon_i|X_I] = 0$ together with a full rank $E[X_i X_i']$ can help to identify $\beta$. This is semiparametric model. However, if we assume $f(\varepsilon_i \mid X_i) \sim N(0, \sigma^2)$, then this is a conditional parametric model as it completely describes $f(Y_i \mid X_i)$ with the likelihood $L\left((\beta, \sigma^2); \mathbf{Y}|\mathbf{X}\right)$.

In practice, we work with the log-likelihood, which is more convenient. The **log-likelihood** is

$$\ell_n(\theta) = \frac{1}{n}\log L(\theta; \mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}\log f(Y_i; \theta).$$

Here, we put $1/n$ to average the log-likelihood. This scaling factor does not change the estimation at all. The MLE estimator is then defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\ell_n(\theta).$$

To justify the likelihood principle, consider the population version of the

$$\ell(\theta) = E[\log f(Y; \theta)]$$

**Theorem 1.1.** *When model is correctly specified, $\theta_0$ is the maximizer.*

*Proof.* The Kullback-Leibler distance

$$\begin{aligned} E[\log f(Y;\theta_0)] - E[\log f(Y;\theta)] &= E[\log(f(Y;\theta_0)/f(Y;\theta))] = -E[\log(f(Y;\theta)/f(Y;\theta_0))] \\ &\geq -\log E[f(Y;\theta)/f(Y;\theta_0)] = 0, \end{aligned}$$

where the inequality holds by Jensen's inequality for the convex function $-\log(\cdot)$. □

## 1.3 Score, Hessian, and Information

Define the **score** as

$$\psi_n(\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(Y_i; \theta)$$

and the **Hessian** as

$$\mathscr{H}_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y_i;\theta).$$

Moreover, define the **efficient score**

$$\psi_0 = \frac{\partial}{\partial\theta} \log f(Y_i;\theta_0)$$

as the first derivative of the log-likelihood of a representative observation $Y_i$ evaluated at $\theta_0$.

**Theorem 1.2.** *If the model is correctly specified, the support of $Y$ does not depend on $\theta$, and $\theta_0$ is in the interior of $\Theta$, then $E[\psi_0] = 0$.*

*Proof.* By the Leibniz integral rule,[1]

$$\begin{aligned} E(\psi(\theta)) &= E\left[\frac{\partial}{\partial\theta} \log f(Y;\theta)\right] = \int \frac{\partial}{\partial\theta} \log f(Y;\theta)\, dF(Y;\theta_0) \\ &= \frac{\partial}{\partial\theta} \int \log f(Y;\theta)\, dF(Y;\theta_0) = \frac{\partial}{\partial\theta} E[\log f(Y;\theta)]. \end{aligned}$$

Evaluate $\frac{\partial}{\partial\theta} E[\log f(Y_i;\theta)]$ at $\theta_0$, we have $\frac{\partial}{\partial\theta} E[\log f(Y_i;\theta_0)] = 0$ as $E[\log f(Y_i;\theta_0)]$ is maximized $\theta_0$, which is an interior. □

The **Fisher information matrix**

$$\mathscr{I}_0 = E\left[\psi_0\psi_0'\right]$$

is the variance of the efficient score. The **expected Hessian** is

$$\mathscr{H}_0 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y;\theta_0)\right].$$

Since $E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y;\theta_0)\right]$ is negative definite, in this definition the expected Hessian is positive-definite.

**Theorem 1.3.** *If the model is correctly specified, we have $\mathscr{I}_0 = \mathscr{H}_0$, the **information matrix equality**.*[2]

*Proof.* Start with the Hessian:

$$\begin{aligned} E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(\theta_0)\right] &= E\left[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'} \log f(\theta_0)\right] = E\left[\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta'} f(\theta)}{f(\theta)}\Big|_{\theta=\theta_0}\right] \\ &= -E\left[\frac{\frac{\partial}{\partial\theta} f(\theta)\,\frac{\partial}{\partial\theta'} f(\theta)}{f^2(\theta)}\Big|_{\theta=\theta_0}\right] + E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'} f(\theta)}{f(\theta)}\Big|_{\theta=\theta_0}\right]. \end{aligned}$$

[1] Leibniz integral rule (from undergraduate calculus): For an integral of the form $\int_{a(t)}^{b(t)} g(t,z)\, dz$, the derivative

$$\frac{d}{dt}\int_{a(t)}^{b(t)} g(t,z)\, dz = g(t,b(t)) \cdot \frac{d}{dt} b(t) - g(t,a(t)) \cdot \frac{d}{dt} a(t) + \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} g(t,z)\, dz.$$

[2] Some textbooks may define the expected Hessian as $\mathscr{H}_0 = E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y;\theta_0)\right]$. This is also valid if we rewrite the information matrix equality as $\mathscr{I}_0 + \mathscr{H}_0 = 0$.

The first term:

$$E\left[\frac{\frac{\partial}{\partial\theta}f(\theta)\frac{\partial}{\partial\theta'}f(\theta)}{f^2(\theta_0)}\right]=E\left[\frac{\partial}{\partial\theta}\log f(\theta_0)\frac{\partial}{\partial\theta'}\log f(\theta_0)\right]=E\left[\psi_0\psi_0'\right]=\mathscr{I}_0.$$

The second term: $E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta)}\right]=\int\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta)}f(\theta_0)\,dx$. If the model is correctly specified, evaluated at $\theta=\theta_0$ the $f(\theta_0)$ cancels out, leading to

$$\int\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta_0)}{f(\theta_0)}f(\theta_0)\,dy=\int\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta_0)\,dy=\frac{\partial^2}{\partial\theta\partial\theta'}\int f(\theta_0)\,dy=\frac{\partial^2}{\partial\theta\partial\theta'}1=0.$$

We complete the proof. □

Notice that the information matrix equality holds only when the model is correctly specified. It fails when the model is misspecified.

## 1.4 Cramér-Rao Lower Bound

**Theorem 1.4.** *Suppose the model is correctly specified, the support of $Y$ does not depend on $\theta$, and $\theta_0$ is in the interior of $\Theta$. If $\widetilde{\theta}$ is unbiased estimator, then $var(\widetilde{\theta})\geq(n\mathscr{I}_0)^{-1}$.*

*Proof.* Because of unbiasedness,

$$\theta=E_\theta\left[\widetilde{\theta}\right]=\int\widetilde{\theta}f(\mathbf{Y};\theta)\,d\mathbf{y}$$

for any $\theta\in\Theta$. $\mathbf{Y}$ here is for the entire sample, $f(\mathbf{Y};\theta)=f(Y_1,...,Y_n;\theta)=\prod_{i=1}^n f(Y_i;\theta)$. Take derivative at the two sides. The LHS is

$$\frac{\partial\theta}{\partial\theta'}=\mathbf{I}_p.$$

The RHS:

$$\begin{aligned}\frac{\partial}{\partial\theta'}\int\widetilde{\theta}f(\mathbf{Y};\theta)\,d\mathbf{y}&=\int\widetilde{\theta}\frac{\partial}{\partial\theta'}f(\mathbf{Y};\theta)\,d\mathbf{y}\\&=\int\widetilde{\theta}\frac{\frac{\partial}{\partial\theta'}f(\mathbf{Y};\theta)}{f(\mathbf{Y};\theta)}f(\mathbf{Y};\theta)\,d\mathbf{y}\\&=\int\widetilde{\theta}\frac{\partial}{\partial\theta'}\log f(\mathbf{Y};\theta)f(\mathbf{Y};\theta)\,d\mathbf{y}\\&=\int\widetilde{\theta}\psi_n(\theta)f(\mathbf{Y};\theta)\,d\mathbf{y}\end{aligned}$$

Evaluate at the true $\theta_0$, and due to i.i.d. data

$$\mathbf{I}_p=\int\widetilde{\theta}\psi_n(\theta_0)f(\mathbf{Y};\theta_0)\,d\mathbf{y}=E\left[\widetilde{\theta}\psi_n(\theta_0)\right]=E\left[\left(\widetilde{\theta}-\theta_0\right)\psi_n(\theta_0)\right]$$

where the last equality holds by $E[\theta_0\psi_n(\theta_0)]=\theta_0E[\psi_n(\theta_0)]=\theta_0E[n\psi_0]=0$. We thus have

$$var\begin{pmatrix}\widetilde{\theta}-\theta_0\\ \psi_n(\theta_0)\end{pmatrix}=\begin{bmatrix}\boldsymbol{V}&\mathbf{I}_p\\ \mathbf{I}_p&n\mathscr{I}_0\end{bmatrix}.$$

Pre- and post-multiply $\left[\begin{array}{cc} \boldsymbol{I}_p & -(n\mathscr{I}_0)^{-1} \end{array}\right]$, we have

$$\left[\begin{array}{cc} \boldsymbol{I}_p & -(n\mathscr{I}_0)^{-1} \end{array}\right] \left[\begin{array}{cc} \boldsymbol{V} & \boldsymbol{I}_p \\ \boldsymbol{I}_p & n\mathscr{I}_0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{I}_p \\ -(n\mathscr{I}_0)^{-1} \end{array}\right] = \boldsymbol{V} - (n\mathscr{I}_0)^{-1} \geq 0.$$

□

The Cramér-Rao Lower Bound is a lower bound. It may not reachable. When it is reached, an estimator is **Cramér-Rao efficient** if it is unbiased and the variance is $(n\mathscr{I}_0)^{-1}$.

**Example 1.3.** Normal distribution: Let $\gamma = \sigma^2$. Then the log-likelihood is

$$\log \ell_n(Y; \mu, \gamma) = -\frac{n}{2}\log\gamma - \frac{n}{2}\log\pi - \frac{1}{2\gamma}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

The score and Hessian are

$$\psi_n(\mu, \gamma) = \begin{cases} \frac{1}{\gamma}\sum_{i=1}^{n}(Y_i - \mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \end{cases}$$

$$\mathscr{H}_n(\mu, \gamma) = \left[\begin{array}{cc} \frac{n}{\gamma} & \frac{1}{2\gamma^2}\sum_{i=1}^{n}(Y_i - \mu) \\ \frac{1}{2\gamma^2}\sum_{i=1}^{n}(Y_i - \mu) & -\frac{n}{2\gamma^2} + \frac{1}{\gamma^3}\sum_{i=1}^{n}(Y_i - \mu)^2 \end{array}\right].$$

The expected Hessian is

$$E\left[\mathscr{H}_n(\mu, \gamma)\right] = \left[\begin{array}{cc} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{array}\right].$$

The inverse $\left[\begin{array}{cc} \frac{\gamma}{n} & 0 \\ 0 & 2\frac{\gamma^2}{n} \end{array}\right]$ is the lower bound.

Check: the variance of the sample mean:$var\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right) = \frac{\gamma}{n}$. The sample mean is Cramér-Rao efficient.

Regarding the variance estimator, notice

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{n-1}Y'\left(I - \frac{1}{n}1_n 1_n'\right)Y$$

is the unbiased estimator as $E\left(S_n^2\right) = \gamma$. Since

$$(n-1)\frac{s_n^2}{\gamma} = \left(\frac{Y}{\sigma}\right)'\left(I - \frac{1}{n}1_n 1_n'\right)\left(\frac{Y}{\sigma}\right) \sim \chi^2(n-1),$$

we have $s_n^2 = \frac{\chi^2(n-1)}{n-1}\gamma$ and thus

$$var\left(s_n^2\right) = \frac{\gamma^2}{(n-1)^2}2(n-1) = \frac{2\gamma^2}{n-1} > \frac{2\gamma^2}{n}$$

does not satisfy Cramér-Rao efficient.

## 1.5 Asymptotic Normality

Under regularity conditions, $\hat{\theta} \xrightarrow{p} \theta_0$, and asymptotically normal:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathscr{H}_0^{-1}\mathscr{I}_0\mathscr{H}_0^{-1}\right)$$

When the information equality holds, the asymptotic variance is simplified as $\mathscr{I}_0^{-1}\mathscr{I}_0\mathscr{I}_0^{-1} = \mathscr{I}_0^{-1}$, and thus

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathscr{I}_0^{-1}\right).$$

This is a nice result, because not only $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ is asymptotically unbiased (the limiting normal distribution is centered at 0), but also it asymptotically achieves the Cramér-Rao efficiency bound with $\mathscr{I}_0^{-1}$.

This beautiful result holds if and only if the model is correctly specified. The next section discusses the asymptotic behavior of the MLE estimator when the model is misspecified.

## 1.6 Misspecified Model

Kullback-Leibler information criterion (KLIC) is defined as

$$KLIC\left(f, g\right) = \int f\left(z\right) \log \frac{f\left(z\right)}{g\left(z\right)} dz.$$

It is a scalar measure of the difference between two probability measures. Because $-\log\left(\cdot\right)$ is a convex function, we can use the Jensen's inequality to show $KLIC\left(f, g\right) \geq 0$. Obviously $KLIC\left(f, f\right) = 0$.

We narrow down our discussion to parametric models. Suppose the data $z$ is drawn from a parameter model $f$; this is the data generating process (DGP). The researcher specifies a family of models $g\left(z;\theta\right)$ and a parameter space $\Theta$; thus such a specification covers a model space $G\left(\Theta\right) = \left\{g\left(z;\theta\right) : \theta \in \Theta\right\}$. If $f \in G\left(\Theta\right)$, the model is **correctly specified**; otherwise it is misspecified. Notice that

$$\begin{aligned} KLIC\left(f, g\left(z;\theta\right)\right) &= \int f\left(z\right) \log f\left(z\right) dz - \int f\left(z\right) \log g\left(z;\theta\right) dz \\ &= E\left[\log f\left(z\right)\right] - E\left[\log g\left(z;\theta\right)\right] \\ &= E\left[\log f\left(z\right)\right] - \ell\left(\theta\right). \end{aligned}$$

Notice that the first term is irrelevant to the choice of $\theta$. If the model is correctly specified, then the true value

$$\theta_0 = \arg\max_{\theta\in\Theta} \ell\left(\theta\right)$$

and $KLIC\left(f, g\left(z;\theta_0\right)\right) = 0$.

If the model is misspecified, $\min_{\theta\in\Theta} KLIC\left(f, g\left(z;\theta\right)\right) > 0$. In this case, we can define

$$\theta^* = \arg\max_{\theta\in\Theta} \ell\left(\theta\right)$$

as the **pseudo-true** parameter value. In other words, $\theta^*$ is the minimizer of $KLIC\left(f, g\left(z;\theta\right)\right)$ in the parameter space $\Theta$. This provides meaningful interpretation of MLE even if the model is misspecified: although in reality the researcher may not know the true family of distributions where

the data is drawn, she can run MLE to look for the value $\theta^*$ that minimizes the gap between $f$ and $g(z;\theta)$ in terms of KLIC.

Under standard assumption, the MLE estimator $\widehat{\theta} \xrightarrow{p} \theta^*$ despite model misspecification, and the asymptotic distribution is

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N\left(0, \mathscr{H}_*^{-1}\mathscr{I}_*\mathscr{H}_*^{-1}\right),$$

where $\mathscr{I}_*$ and $\mathscr{H}_*$ are evaluated at the pseudo-true value $\theta^*$. Attention should be paid to the sandwich-form asymptotic variance, since the information equality does not hold under misspecification and as a result, $\mathscr{I}_* \neq \mathscr{H}_*$.