Chapter 7

Asymptotic Properties of MLE

7.1 Examples of MLE

Normal, Logistic, Probit, Poisson

7.2 Consistency

We specify a parametric distribution (pdf) $f(x;\theta)$ and a parameter space Θ. Define $Q(\theta) = E[\log f(x;\theta)]$, and $\theta_0 = \arg\max_{\theta\in\Theta} Q(\theta)$ maximizes the expected log-likelihood. Given a sample of n observations, we compute the average sample log-likelihood $\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x;\theta)$. The MLE estimator is $\widehat{\theta} = \arg\max_{\theta\in\Theta} \ell_n(\theta)$.

We say that *correctly specified* if the data $(x_1, \ldots, x_n)$ is generated from the pdf $f(x;\theta)$ for some $\theta \in \Theta$. Otherwise if the data is not generated

from any member in the class of distributions $\mathcal{M}^* := \{\theta \in \Theta : f(x;\theta)\}$, we say it is *misspecified*. When the model is misspecified, strictly speaking the log-likelihood function $\ell_n(\theta)$ should be called quasi log-likelihood and the MLE estimator $\widehat{\theta}$ should be called the *quasi MLE*.

We will discuss under what condition $\widehat{\theta} \xrightarrow{p} \theta_0$, that is, the maximizer of the sample log-likelihood converges in probability to the maximizer of the expected log-likelihood in population. Notice that unlike OLS, most MLE estimators do not admit a closed-form. They are defined as a maximizer and solved by numerical optimization.

The first requirement for the consistency of MLE is that θ_0 uniquely defined. Suppose $\theta_0 \in \text{int}(\Theta)$ lies in the interior of Θ. Let $N(\theta_0, \varepsilon) = \{\theta \in \Theta : |\theta - \theta_0| < \varepsilon\}$ is a neighborhood around θ_0 with radius ε for some $\varepsilon > 0$.

Definition 7.1 (Identification)**.** The value θ_0 is identified if for any $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $Q(\theta_0) > \sup_{\theta \in \Theta \setminus N(\theta_0, \varepsilon)} Q(\theta) + \delta$.

We know under suitable condition, LLN implies $\ell_n(\theta) \xrightarrow{p} Q(\theta)$ for each $\theta \in \Theta$. This is a pointwise result, meaning θ is taken as fixed as $n \to \infty$. However, $\widehat{\theta}$ is random in finite-sample, which makes $\ell_n(\widehat{\theta})$ a complicated function of the data in particular when $\widehat{\theta}$ has no closed-form solution. We therefore need to strengthen the pointwise LLN.

Definition 7.2 (ULLN)**.** We say a *uniform law of large numbers* (ULLN) for

$Q(\theta)$ holds on Θ if

$$P\left\{\sup_{\theta\in\Theta}|\ell_n(\theta)-Q(\theta)|\geq\varepsilon\right\}\to 0 \tag{7.1}$$

for all $\varepsilon>0$ as $n\to\infty$.

ULLN can be established under pointwise LLN plus some regularity conditions, for example when Θ is a compact set, and $\log f(x;\cdot)$ is continuous in θ almost everywhere on the support of x.

Theorem 7.1. *If θ_0 is identified and ULLN* (7.1) *hold, then $\widehat{\theta}\xrightarrow{p}\theta_0$.*

Proof. According to the definition of consistency, we can check

$$\begin{aligned}
&P\left\{\left|\widehat{\theta}-\theta_0\right|>\varepsilon\right\}\leq P\left\{Q(\theta_0)-Q(\widehat{\theta})>\delta\right\}\\
&=P\left\{Q(\theta_0)-\ell_n(\theta_0)+\ell_n(\theta_0)-\ell_n(\widehat{\theta})+\ell_n\left(\widehat{\theta}\right)-Q(\widehat{\theta})>\delta\right\}\\
&\leq P\left\{|Q(\theta_0)-\ell_n(\theta_0)|+\ell_n(\theta_0)-\ell_n(\widehat{\theta})+\left|\ell_n\left(\widehat{\theta}\right)-Q(\widehat{\theta})\right|>\delta\right\}\\
&\leq P\left\{|Q(\theta_0)-\ell_n(\theta_0)|+\left|\ell_n(\widehat{\theta})-Q(\widehat{\theta})\right|\geq\delta\right\}\\
&\leq P\left\{2\sup_{\theta\in\Theta}|\ell_n(\theta)-Q(\theta)|\geq\delta\right\}=P\left\{\sup_{\theta\in\Theta}|\ell_n(\theta)-Q(\theta)|\geq\frac{\delta}{2}\right\}\to 0.
\end{aligned}$$

The first line holds because of identification, the third line by the triangle inequality, the fourth line by the definition of MLE that $\ell_n(\widehat{\theta})\geq\ell_n(\theta_0)$, and the last line by ULLN. □

Identification is a necessary condition for consistent estimation. Although $\widehat{\theta}$ has no closed-form solution in general, we establish consistency

via ULLN over all point $\theta \in \Theta$ under consideration.

7.3 Asymptotic Normality

The next step is to derive the asymptotic distribution of the MLE estimator.

Theorem 7.2. *Under suitable regularity conditions, the MLE estimator*

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \left(E\left[\frac{\partial^2 \log f\left(x;\theta_0\right)}{\partial\theta\partial\theta'}\right]\right)^{-1} \operatorname{var}\left[\frac{\partial \log f\left(x;\theta_0\right)}{\partial\theta}\right]\left(E\left[\frac{\partial^2 \log f\left(x;\theta_0\right)}{\partial\theta\partial\theta'}\right]\right)^{-1}\right).$$

Remark 7.1. The "suitable regularity conditions" will be spelled out later. Indeed, those conditions can be observed in the proof.

Proof. That $\widehat{\theta}$ is a maximizer entails $\frac{\partial}{\partial\theta}\ell_n\left(\widehat{\theta}\right)=0$. Take a Taylor expansion of $\frac{\partial}{\partial\theta}\ell_n\left(\widehat{\theta}\right)$ around $\frac{\partial}{\partial\theta}\ell_n\left(\theta_0\right)$:

$$0-\frac{\partial}{\partial\theta}\ell_n\left(\theta_0\right)=\frac{\partial}{\partial\theta}\ell_n\left(\widehat{\theta}\right)-\frac{\partial}{\partial\theta}\ell_n\left(\theta_0\right)=\frac{\partial}{\partial\theta\partial\theta'}\ell_n\left(\dot{\theta}\right)\left(\widehat{\theta}-\theta_0\right)$$

where $\dot{\theta}$ is some point on the line segment connecting $\widehat{\theta}$ and θ_0. Rearrange the above equation and multiply both side by $\sqrt{n}$:

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)=-\left(\frac{\partial}{\partial\theta\partial\theta'}\ell_n\left(\dot{\theta}\right)\right)^{-1}\sqrt{n}\frac{\partial}{\partial\theta}\ell_n\left(\theta_0\right). \tag{7.2}$$

When $Q\left(\theta\right)$ is differentiable at θ_0, we have $\frac{\partial}{\partial\theta}Q\left(\theta_0\right)=0$ by the first condition of optimality of θ_0 for $Q\left(\theta\right)$. Notice that $E\left[\frac{\partial}{\partial\theta}\log f\left(x;\theta_0\right)\right]=$

$\frac{\partial}{\partial\theta}Q(\theta_0) = 0$ if differentiation and integration are interchangeable. By CLT, the second factor in (7.2) follows

$$\sqrt{n}\frac{\partial}{\partial\theta}\ell_n(\theta_0) \xrightarrow{d} N\left(0, \operatorname{var}\left[\frac{\partial}{\partial\theta}\log f(x;\theta_0)\right]\right).$$

Suppose the second factor in (7.2) follows $\frac{\partial}{\partial\theta\partial\theta'}\ell_n(\dot{\theta}) \xrightarrow{p} E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta_0)\right]$ (sufficient if we assume $E\left[\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_l}\log f(x;\theta_0)\right]$ is continuous in θ for all $i,j,l \leq K$. Thus we have the conclusion by Slutsky's theorem. □

When the model is misspecified, the asymptotic variance takes a complicated sandwich form. When the parametric model is correctly specified, then the asymptotic variance can be further simplified, thanks to the following important result of information matrix equality.

7.4 Information Matrix Equality

When the model is correctly specified, θ_0 is the *true* parameter value. The variance $\mathcal{I}(\theta_0) := \operatorname{var}_{f(x;\theta_0)}\left[\frac{\partial}{\partial\theta}\log f(x;\theta_0)\right]$ is called the *(Fisher) information matrix*, and $\mathcal{H}(\theta_0) := E_{f(x;\theta_0)}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta_0)\right]$ is called the *expected Hessian matrix*. Here we emphasize the true underlying distribution $f(x;\theta_0)$ by writing it as the subscript of the mathematical expectations.

Fact 7.1. *Under suitable regularity conditions, we have* $\mathcal{I}(\theta_0) = -\mathcal{H}(\theta_0)$

Proof. Because $f(x;\theta_0)$ a pdf, $\int f(x;\theta_0)\,dx = 1$. Take partial derivative

with respect to θ,

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial \theta} f\left(x;\theta_0\right) dx = \int \frac{\partial f\left(x;\theta_0\right)/\partial\theta}{f\left(x;\theta_0\right)} f\left(x;\theta_0\right) dx \\ &= \int \left[\frac{\partial}{\partial \theta} \log f\left(x;\theta_0\right)\right] f\left(x;\theta_0\right) dx = E_{f(x;\theta_0)}\left[\frac{\partial}{\partial \theta} \log f\left(x;\theta_0\right)\right] \end{aligned} \tag{7.3}$$

where the third equality holds as by the chain rule

$$\frac{\partial}{\partial \theta} \log f\left(\theta_0\right) = \frac{\partial f\left(x;\theta_0\right)/\partial\theta}{f\left(x;\theta_0\right)}. \tag{7.4}$$

Take a second partial derivative of (7.3) with respective to θ, according to the chain rule:

$$\begin{aligned} 0 &= \int \left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(x;\theta_0\right)\right] f\left(x;\theta_0\right) dx + \int \left[\frac{\partial}{\partial\theta} \log f\left(x;\theta_0\right)\right] \frac{\partial}{\partial\theta'} f\left(x;\theta_0\right) dx \\ &= \int \left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(x;\theta_0\right)\right] f\left(x;\theta_0\right) dx + \int \frac{\partial}{\partial\theta} \log f\left(x;\theta_0\right) \frac{\partial f\left(x;\theta_0\right)/\partial\theta}{f\left(x;\theta\right)} f\left(x;\theta_0\right) dx \\ &= \int \left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(x;\theta_0\right)\right] f\left(x;\theta_0\right) dx + \int \left[\frac{\partial}{\partial\theta} \log f\left(x;\theta_0\right) \frac{\partial}{\partial\theta'} \log f\left(x;\theta_0\right)\right] f\left(x;\theta_0\right) dx \\ &= E_{f(x;\theta_0)}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(x;\theta_0\right)\right] + E_{f(x;\theta_0)}\left[\frac{\partial}{\partial\theta} \log f\left(x;\theta_0\right) \frac{\partial}{\partial\theta'} \log f\left(x;\theta_0\right)\right] \\ &= \mathcal{H}\left(\theta_0\right) + \mathcal{I}\left(\theta_0\right). \end{aligned}$$

The second equality follows by (7.4). The last equality by (7.3) as the zero mean ensures the variance of $\frac{\partial}{\partial\theta} \log f\left(x;\theta_0\right)$ is equal to the expectation of its out-product. □

Notice that a correct specification is essential for the information matrix

equality. If the true data generating distribution is $g \notin \mathcal{M}^*$, then (7.3) breaks down because

$$0 = \int \frac{\partial}{\partial \theta} f(x;\theta_0) = \int \left[g^{-1} \frac{\partial}{\partial \theta} f(x;\theta_0) \right] g = E_g \left[g^{-1} \frac{\partial}{\partial \theta} f(x;\theta_0) \right]$$

but $g^{-1} \frac{\partial}{\partial \theta} f(x;\theta_0) \neq (f(x;\theta_0))^{-1} \frac{\partial}{\partial \theta} f(x;\theta_0) = \frac{\partial}{\partial \theta} \log f(\theta_0)$. The asymptotic variance in Theorem 7.2,

$$\left(E_g \left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(x;\theta_0) \right] \right)^{-1} \operatorname{var}_g \left[\frac{\partial}{\partial \theta} \log f(x;\theta_0) \right] \left(E_g \left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(x;\theta_0) \right] \right)^{-1},$$

written explicitly in $E_g[\cdot]$, is still valid.

When the parametric model $\mathcal{M}^*$ is correctly specified, then we can replace $E_g \left[\frac{\partial^2 \ell_n}{\partial \theta \partial \theta'} (\theta_0) \right]$ by $\mathcal{H}(\theta_0)$ and replace $\operatorname{var}_g \left[\frac{\partial \ell_n}{\partial \theta} (\theta_0) \right]$ by $\mathcal{I}(\theta_0)$, we simplify the asymptotic variance as

$$(\mathcal{H}(\theta_0))^{-1} \mathcal{I}(\theta_0) (\mathcal{H}(\theta_0))^{-1} = (-\mathcal{I}(\theta_0))^{-1} \mathcal{I}(\theta_0) (-\mathcal{I}(\theta_0))^{-1} = (\mathcal{I}(\theta_0))^{-1}$$

by the information matrix equality Fact 7.1.

Corollary 7.1. *If the model is correctly specified, under the conditions for Theorem 7.3 and Fact 7.1 the MLE estimator*

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

This is the classical asymptotic normality result of MLE.

7.5 Cramer-Rao Lower Bound

7.6 Summary

Further reading: White (1996), Newey and McFadden (1994).

Zhentao Shi. Oct 29, 2020.

Bibliography

Newey, K. and D. McFadden (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245. 7.6

White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge university press. 7.6