STAT 620: Asymptotic Statistics **Spring 2022**

Lecture: Feb 10

Lecturer: Xianyang Zhang

1 Asymptotic Normality

Our goal is to establish the asymptotic normality of MLE under some suitable assumptions.

1.1 Assumptions

- The Hessian is Lipchitz continuous:
$$||\nabla^2 l_{\theta_1}(x) - \nabla^2 l_{\theta_2}(x)||_{op} \le M(x)||\theta_1 - \theta_2||;$$
- $E_{\theta_0}[M(X)^2] < \infty$ where $X \sim P_{\theta_0}$;
- $\hat{\theta}_n = \text{argmax}_{\theta\in\Theta} P_n l_\theta$ is consistent, that is $\hat{\theta}_n \xrightarrow{P} \theta_0$;
- $E_{\theta_0}||\nabla l_{\theta_0}||^2 < \infty$.

1.2 Theorem

Under the above assumptions, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, [P_{\theta_0}\nabla^2 l_{\theta_0}]^{-1}[P_{\theta_0}\nabla l_{\theta_0}\nabla l_{\theta_0}^\top][P_{\theta_0}\nabla^2 l_{\theta_0}]^{-1}).$$

1.2.1 Proof

By the Taylor expansion for $\nabla l_\theta(x) : \mathbb{R}^d \to \mathbb{R}^d$, we have

$$\nabla l_\theta(x) = \nabla l_{\theta_0}(x) + \nabla^2 l_{\theta_0}(x)(\theta - \theta_0) + \gamma(x)(\theta - \theta_0)$$

where $\gamma(x) \in \mathbb{R}^{d\times d}$. Therefore, for $\hat{\theta}_n = \arg\max_{\theta\in\Theta} P_n l_\theta$,

$$\nabla l_{\hat{\theta}_n}(x) = \nabla l_{\theta_0}(x) + \nabla^2 l_{\theta_0}(x)(\hat{\theta}_n - \theta_0) + \hat{\gamma}(x)(\hat{\theta}_n - \theta_0).$$

Notice that $0 = P_n \nabla l_{\hat{\theta}_n}$. Then we get

$$0 = P_n\nabla l_{\hat{\theta}_n} = P_n \nabla l_{\theta_0} + P_n \nabla^2 l_{\theta_0}(\hat{\theta}_n - \theta_0) + P_n\hat{\gamma}(\hat{\theta}_n - \theta_0)$$

where $P_n\hat{\gamma} = \frac{1}{n}\sum_{i=1}^n \hat{\gamma}(X_i)$. By the first moment assumption and a lemma from previous lecture,

$$||P_n\hat{\gamma}||_{op} \xrightarrow{p} 0,$$

and thus

$$0 = P_n\nabla l_{\theta_0} + P_n\nabla^2 l_{\theta_0}(\hat{\theta} - \theta) - o_p(1)(\hat{\theta} - \theta).$$

From the second assumption,

$$-P_n\nabla l_{\theta_0} = [P_n\nabla^2 l_{\theta_0} + o_p(1)](\hat{\theta} - \theta).$$

Rearranging terms, we finally obtain

$$\sqrt{n}(\hat{\theta} - \theta) = -[P_n\nabla^2 l_{\theta_0} + o_p(1)]^{-1}\sqrt{n}(P_n\nabla l_{\theta_0} - P_{\theta_0}\nabla l_{\theta_0}).$$

The result follows from the central limit theorem, law of large numbers and Slutsky's theorem.

1.3 Corollary

Under the above assumptions, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1}).$$

1.3.1 Proof

From the previous theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, [P_{\theta_0}\nabla^2 l_{\theta_0}]^{-1}[P_{\theta_0}\nabla l_{\theta_0}\nabla l_{\theta_0}^\top][P_{\theta_0}\nabla^2 l_{\theta_0}]^{-1}).$$

Note that

$$P_{\theta_0}\nabla l_{\theta_0} = \mathbb{E}_{\theta_0}\nabla l_{\theta_0}(X) = \int \nabla l_{\theta_0} p_{\theta_0} d\mu = \int \frac{\nabla p_{\theta_0}}{p_{\theta_0}} p_{\theta_0} d\mu = \int \nabla p_{\theta_0} d\mu = \nabla \int p_{\theta_0} d\mu = 0,$$

and

$$\begin{aligned} P_{\theta_0}\nabla^2 l_{\theta_0} &= \int \frac{p_{\theta_0}\nabla^2 p_{\theta_0} - (\nabla p_{\theta_0})(\nabla p_{\theta_0})^\top}{p_{\theta_0}} d\mu \\ &= \int \nabla^2 p_{\theta_0} d\mu - \int \frac{(\nabla p_{\theta_0})(\nabla p_{\theta_0})^\top}{p_{\theta_0}} d\mu \\ &= -\int (\nabla l_{\theta_0})(\nabla l_{\theta_0})^\top p_{\theta_0} d\mu \\ &= -\mathbb{E}_{\theta_0}[(\nabla l_{\theta_0})(\nabla l_{\theta_0})^\top] \\ &= -\operatorname{cov}_{\theta_0}(\nabla l_{\theta_0}) = -I_{\theta_0}. \end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (-I_{\theta_0})^{-1} I_{\theta_0} (-I_{\theta_0})^{-1})$$

and hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1})$$

where I_{θ_0} is the Fisher information matrix.

2 Information Inequality

2.1 Lemma

Consider $\delta : \mathcal{X}^n \to \mathbb{R}$ and $\Psi : \mathcal{X}^n \to \mathbb{R}^d$. Suppose $\mathbb{E}\Psi = \mathbf{0}$. Define

$$\gamma = [\operatorname{cov}(\delta, \Psi_1), \operatorname{cov}(\delta, \Psi_2), \cdots, \operatorname{cov}(\delta, \Psi_d)]^\top \in \mathbb{R}^d$$

and $C = \operatorname{cov}(\Psi) \in \mathbb{R}^{d\times d}$. Then,

$$var(\delta) \geq \gamma^\top C^{-1} \gamma.$$

2.1.1 Proof

Fix $v \in \mathbb{R}^d$. Consider $\operatorname{cov}(\delta, v^\top\Psi)$. We have

$$\operatorname{cov}(\delta, v^\top\Psi) \leq \sqrt{var(\delta)\operatorname{var}(v^\top\Psi)} = \sqrt{var(\delta) v^\top C v}$$

which implies that

$$var(\delta) \geq \frac{[\mathrm{cov}(\delta, v^\top \Psi)]^2}{v^\top C v} = \frac{(v^\top \gamma)^2}{v^\top C v}.$$

Note that

$$\frac{(v^\top \gamma)^2}{v^\top C v} = \frac{(u^\top C^{-\frac{1}{2}} \gamma)^2}{u^\top u} \leq \frac{u^\top u \gamma^\top C^{-1} \gamma}{u^\top u} = \gamma^\top C^{-1} \gamma,$$

where $u = C^{\frac{1}{2}} v$ and the inequality is due to the Cauchy-Schwarz inequality. The inequality becomes equality when we pick $u = C^{-\frac{1}{2}} \gamma$. The result thus follows.

2.2 Theorem (Cramer-Rao)

Let $g(\theta) = \mathbb{E}_\theta \delta$ and $I_\theta = \mathbb{E}_\theta[\nabla l_\theta \nabla l_\theta^\top]$. Assume I_θ is non-singular and $g(\theta)$ is differentiable. Then,

$$\mathrm{var}_\theta(\delta) \geq \nabla g(\theta)^\top I_\theta^{-1} \nabla g(\theta).$$

2.2.1 Proof

We will apply the above lemma to prove this theorem. Let $\Psi = \nabla l_\theta$. Then $C = I_\theta$. We shall show that $\gamma = \nabla g(\theta)$. Toward this end, note that

$$\begin{aligned}
\mathrm{cov}(\delta, \nabla l_\theta) =& \mathbb{E}_\theta[\delta \nabla l_\theta] \\
=& \int \delta \nabla l_\theta p_\theta d\mu \\
=& \int \delta \frac{\nabla p_\theta}{p_\theta} p_\theta d\mu \\
=& \int \delta \nabla p_\theta d\mu \\
=& \nabla \int \delta p_\theta d\mu \\
=& \nabla g(\theta).
\end{aligned}$$

2.3 Theorem

If $\hat{\theta} : \mathcal{X}^n \to \Theta \in \mathbb{R}^d$ is unbiased, then

(1) $\mathbb{E}[||\hat{\theta} - \theta||^2] \geq \mathrm{tr}(I_\theta^{-1})$;

(2) $\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top] \geq I_\theta^{-1}$.

2.3.1 Note

For two positive semi-definite (psd) matrices A and B,

$$A \geq B \Leftrightarrow A - B \text{ is psd } \Leftrightarrow v^\top (A - B) v \geq 0 \text{ for all } v.$$

2.3.2 Proof

We only prove (2). Take $\delta = v^\top \hat{\theta}$ for $v \in \mathbb{R}^d$. Then

$$\mathbb{E}\delta = g(\theta) = v^\top \theta,$$

which implies that $\nabla g(\theta) = v$. Using the above result, we get

$$\mathrm{var}(\delta) = \mathbb{E}[(v^\top (\hat{\theta} - \theta))^2] \geq v^\top I_\theta^{-1} v.$$

2.4 Definition

An estimator T_n for θ is efficient for a family of models $\{P_\theta\}_{\theta\in\Theta}$ if

$$\sqrt{n}(T_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1}).$$

2.4.1 Example (Gaussian Mean)

Consider $\{N(\theta, 1)\}_{\theta\in\Theta}$. Let $T_n = \bar{X}_n$, where $X_1, \ldots, X_n \sim^{i.i.d} N(\theta_0, 1)$ and $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then,

$$\sqrt{n}(\bar{X}_n - \theta_0) \xrightarrow{d} N(0, 1).$$

Note that $I_{\theta_0} = 1$ in this case.