Section 8.2. Asymptotic normality

We assume that $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, where the X_i's are i.i.d. with common density $p(x; \theta_0) \in \mathcal{P} = \{p(x; \theta) : \theta \in \Theta\}$. We assume that θ_0 is *identified* in the sense that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $p(x; \theta) \neq p(x; \theta_0)$ with respect to the dominating measure μ.

In order to prove asymptotic normality, we will need certain regularity conditions. Some of these were encountered in the proof of consistency, but we will need some additional assumptions.

Regularity Conditions

i. θ_0 lies in the interior of Θ, which is assumed to be a compact subset of R^k.

ii. $\log p(x;\theta)$ is continuous at each $\theta \in \Theta$ for all $x \in \mathcal{X}$ (a.e. will suffice).

iii. $|\log p(x;\theta)| \leq d(x)$ for all $\theta \in \Theta$ and $E_{\theta_0}[d(X)] < \infty$.

iv. $p(x;\theta)$ is twice continuously differentiable and $p(x;\theta) > 0$ in a neighborhood, $\mathcal{N}$, of θ_0.

v. $\|\frac{\partial p(x;\theta)}{\partial \theta}\| \leq e(x)$ for all $\theta \in \mathcal{N}$ and $\int e(x) d\mu(x) < \infty$.

vi. Defining the score vector

$$\psi(x;\theta) = (\partial \log p(x;\theta)/\partial\theta_1, \ldots, \partial \log p(x;\theta)/\partial\theta_k)'$$

then we assume that $I(\theta_0) = E_{\theta_0}[\psi(X;\theta_0)\psi(X;\theta_0)']$ exists and is non-singular.

vii. $\|\frac{\partial^2 \log p(x;\theta)}{\partial\theta\partial\theta'}\| \leq f(x)$ for all $\theta \in \mathcal{N}$ and $E_{\theta_0}[f(X)] < \infty$.

viii. $\|\frac{\partial^2 p(x;\theta)}{\partial\theta\partial\theta'}\| \leq g(x)$ for all $\theta \in \mathcal{N}$ and $\int g(x)d\mu(x) < \infty$.

Theorem 8.6: If these 8 regularity conditions hold, then

$$\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) \stackrel{D(\theta_0)}{\rightarrow} N(0, I^{-1}(\theta_0))$$

Proof: Note that conditions i. - iii. guarantee that the MLE is consistent. Since θ_0 is assumed to lie in the interior of Θ, we know that with sufficiently large probability that the MLE will lie in $\mathcal{N}$ and cannot be on the boundary. This implies that the maximum is also a local maximum, which implies that $\partial Q(\hat{\theta}(\boldsymbol{X}_n); \boldsymbol{X}_n)/\partial\theta = 0$ or $\frac{1}{n}\sum_{i=1}^n \psi(X_i; \hat{\theta}(\boldsymbol{X}_n)) = 0$. That is, the MLE is the solution to the score equations.

By the mean value theorem, applied to each element of the score vector, we have that

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i; \hat{\theta}(\boldsymbol{X}_n)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i; \theta_0) + \{-J_n^*(\boldsymbol{X}_n)\}\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0)$$

Note that $J_n^*(\boldsymbol{X}_n)$ is a $k \times k$ random matrix where the jth row of the matrix is the jth row of J_n evaluated at $\theta_{jn}^*(\boldsymbol{X}_n)$ where $\theta_{jn}^*(\boldsymbol{X}_n)$ is an intermediate value between $\hat{\theta}(\boldsymbol{X}_n)$ and θ_0. $\theta_{jn}^*(\boldsymbol{X}_n)$ may be different from row to row but it will be consistent for θ_0.

We will establish two facts:

F1: $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i; \theta_0) \stackrel{D(\theta_0)}{\rightarrow} N(0, I(\theta_0))$

F2: $J_n^*(\boldsymbol{X}_n) \stackrel{P(\theta_0)}{\rightarrow} I(\theta_0)$

By assumption vi., we know that $I(\theta_0)$ is non-singular. The inversion of a non-singular matrix is a continuous function in θ. Since $J_n^*(\boldsymbol{X}_n) \stackrel{P}{\rightarrow} I(\theta_0)$, we know that $\{J_n^*(\boldsymbol{X}_n)\}^{-1} \stackrel{P}{\rightarrow} I(\theta_0)^{-1}$. This also means that with sufficiently large probability, as n gets large, $J_n^*(\boldsymbol{X}_n)$ is invertible.

Therefore, we know that

$$\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) = \{J_n^*(\boldsymbol{X}_n)\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i; \theta_0)$$

We then use the Slutsky's theorem to conclude that

$$\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) \xrightarrow{D} N(0, I(\theta_0)^{-1})$$

Establishing F1

The random vectors $\psi(X_1;\theta_0),\ldots\psi(X_n;\theta_0)$ are i.i.d. We need to show that they have mean zero. Then, $I(\theta_0)$ will be the covariance matrix of $\psi(X;\theta_0)$ and an application of the multivariate central limit theorem for i.i.d. random vectors gives the desired result.

We will show something stronger, namely $E_\theta[\psi(X;\theta)] = 0$ for all $\theta \in \mathcal{N}$. Condition v. guarantees that we can interchange integration and differentiation. Consider the case where $k = 1$. We know that $1 = \int p(x;\theta)d\mu(x)$ for all $\theta \in \mathcal{N}$. This implies that $0 = \frac{d}{d\theta}\int p(x;\theta)d\mu(x)$. Let's show that $\frac{d}{d\theta}\int p(x;\theta)d\mu(x) = \int \frac{d}{d\theta}p(x;\theta)d\mu(x)$. Choose a sequence $\theta_n \in \mathcal{N}$ such that $\theta_n \to \theta$. Then, by definition of a derivative, we know that

$$\frac{dp(x;\theta)}{d\theta} = \lim_{n\to\infty}\left\{\frac{p(x;\theta_n) - p(x;\theta)}{\theta_n - \theta}\right\} \text{ for all } x \in \mathcal{X}$$

By the mean value theorem, we know that

$$p(x;\theta_n) = p(x;\theta) + \frac{dp(x;\theta_n^*)}{d\theta}(\theta_n - \theta)$$

where θ_n^* lies between θ and θ_n so that $\theta_n^* \in \mathcal{N}$. This implies that

$$\left|\frac{p(x;\theta_n) - p(x;\theta)}{\theta_n - \theta}\right| = \left|\frac{dp(x;\theta_n^*)}{d\theta}\right| \leq e(x)$$

Since $e(x)$ is integrable, we can employ the dominated convergence theorem. This says that

$$\begin{aligned} 0 = \frac{d}{d\theta}\int p(x;\theta)d\mu(x) &= \lim_{n\to\infty}\int \{\frac{p(x;\theta_n) - p(x;\theta)}{\theta_n - \theta}\}d\mu(x) \\ &= \int \lim_{n\to\infty}\{\frac{p(x;\theta_n) - p(x;\theta)}{\theta_n - \theta}\}d\mu(x) \\ &= \int \frac{dp(x;\theta)}{d\theta}d\mu(x) \end{aligned}$$

This can be generalized to partial derivatives which can then be used to formally show that $E_\theta[\psi(X;\theta)] = 0$ for $\theta \in \mathcal{N}$. We know that $\int p(x;\theta)d\mu(x) = 1$. This implies that $\frac{\partial}{\partial\theta_j}\int p(x;\theta)d\mu(x) = 0$. By dominated convergence, we can interchange differentiation and integration so that $\int \frac{\partial p(x;\theta)}{\partial\theta_j} d\mu(x) = 0$. Then, we know that

$$\int \frac{\partial p(x;\theta)/\partial\theta_j}{p(x;\theta)} p(x;\theta)d\mu(x) = 0$$

We can divide by $p(x;\theta)$ since it is greater than zero for all $\theta \in \mathcal{N}$. This implies that $E_\theta[\psi_j(X;\theta)] = 0$.

Establishing F2

First, we shall study the large sample behavior of the matrix of second partial derivatives of the log-likelihood. Define

$$J_n(\theta) = [-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log p(X_i;\theta)}{\partial\theta\partial\theta'}]$$

This is a $k \times k$ random matrix.

We want to demonstrate that $E_{\theta_0}[J_n(\theta_0)] = I(\theta_0)$. We have already established that $\int \frac{\partial}{\partial \theta_j} p(x;\theta) d\mu(x) = 0$ for all $j = 1, \ldots, k$. This implies that

$$\frac{\partial}{\partial \theta_{j'}} \int \frac{\partial}{\partial \theta_j} p(x;\theta) d\mu(x) = 0 \text{ for all } j, j' = 1, \ldots, k.$$

Since the norm of the matrix of second partial derivatives of $p(x;\theta)$ is bounded by an integrable function $g(x)$ (see condition viii), we know that we can interchange integration and differentiation. This implies that

$$\int \frac{\partial^2 p(x;\theta)}{\partial \theta_{j'} \partial \theta_j} d\mu(x) = 0 \text{ for all } j, j' = 1, \ldots, k.$$

We note, however, that

$$\frac{\partial^2 \log p(x;\theta)}{\partial \theta_{j'} \partial \theta_j} = \frac{\partial^2 p(x;\theta)}{\partial \theta_{j'} \partial \theta_j} / p(x;\theta) - \psi_{j'}(x;\theta) \psi_j(x;\theta)$$

This implies that

$$\frac{\partial^2 p(x;\theta)}{\partial\theta_{j'}\partial\theta_j} = p(x;\theta)[\frac{\partial^2 \log p(x;\theta)}{\partial\theta_{j'}\partial\theta_j} + \psi_{j'}(x;\theta)\psi_j(x;\theta)]$$

So,

$$\begin{aligned} 0 = \int \frac{\partial^2 p(x;\theta)}{\partial\theta_{j'}\partial\theta_j} d\mu(x) &= \int p(x;\theta)\frac{\partial^2 \log p(x;\theta)}{\partial\theta_{j'}\partial\theta_j} d\mu(x) + \\ & \int p(x;\theta)\psi_{j'}(x;\theta)\psi_j(x;\theta) d\mu(x) \end{aligned}$$

This implies that $I_{j'j}(\theta) = -E_\theta[\frac{\partial^2 \log p(X;\theta)}{\partial\theta_{j'}\partial\theta_j}]$. In matrix form, this says that $I(\theta) = E_\theta[-\frac{\partial^2 \log p(X;\theta)}{\partial\theta\partial\theta'}]$. So, $E_{\theta_0}[J_n(\theta_0)] = I(\theta_0)$.

By the weak law of large numbers, we know that

$$J_n(\theta) \xrightarrow{P} E_{\theta_0}[-\frac{\partial^2 \log p(X;\theta)}{\partial\theta\partial\theta'}]$$

Define $I_0^*(\theta) = E_{\theta_0}[-\frac{\partial^2 \log p(X;\theta)}{\partial\theta\partial\theta'}]$. Note that $I_0^*(\theta_0) = I(\theta_0)$.

The WLLN enables us to show that $J_n(\theta) \xrightarrow{P} I_0^*(\theta)$ for all $\theta \in \mathcal{N}$. This is pointwise convergence. However, as we will see shortly, this will not suffice to prove our desired results. We will need something stronger. We will have to show that $J_n(\theta)$ converges uniformly in probability to $I_0^*(\theta)$ for $\theta \in \mathcal{N}$. The proof we follow the exact same way as it did when we proved uniform convergence in probability for consistency.

We know that $\frac{\partial^2 \log p(x;\theta)}{\partial\theta\partial\theta'}$ is continuous in $\theta \in \mathcal{N}$. This follows by assumption iv. By condition vii., we know that $\|\frac{\partial^2 \log p(x;\theta)}{\partial\theta\partial\theta'}\| \leq f(x)$ for $\theta \in \mathcal{N}$ and $E_{\theta_0}[f(X)] < \infty$. By the same proof of Lemma 8.3, we know that a. $I_0^*(\theta)$ is continuous for $\theta \in \mathcal{N}$ and b. $\sup_{\theta \in \mathcal{N}} \|J_n(\theta) - I_0^*(\theta)\| \stackrel{P}{\rightarrow} 0$.

This result is important because it allows us to prove the following lemma:

Lemma 8.7 If $\theta_n^*(\boldsymbol{X}_n)$ is a consistent estimator for θ_0, then $J_n(\theta_n^*(\boldsymbol{X}_n)) \stackrel{P}{\rightarrow} I(\theta_0)$.

Proof: By the triangle inequality, we know that

$$\|J_n(\theta^*(\boldsymbol{X}_n)) - I(\theta_0)\| \leq \|J_n(\theta^*(\boldsymbol{X}_n)) - I_0^*(\theta^*(\boldsymbol{X}_n))\| + \|I_0^*(\theta^*(\boldsymbol{X}_n)) - I(\theta_0)\|$$

So, it suffices to prove that 1. $\|J_n(\theta^*(\boldsymbol{X}_n)) - I_0^*(\theta^*(\boldsymbol{X}_n))\| \xrightarrow{P} 0$ and 2. $\|I_0^*(\theta^*(\boldsymbol{X}_n)) - I(\theta_0)\| \xrightarrow{P} 0$.

Since $I_0^*(\theta)$ is a continuous function of $\theta \in \mathcal{N}$, we know that $I_0^*(\theta^*(\boldsymbol{X}_n)) \xrightarrow{P} I(\theta_0)$. This is because a continuous function of a consistent estimator converges in probability to the function of the estimand. So, 2 true.

In order to prove 1, we first note that with arbitrarily large probability and n sufficiently large, $\theta_n^*(\boldsymbol{X}_n) \in \mathcal{N}$. When this happens, we have that

$$\|J_n(\theta_n^*(\boldsymbol{X}_n)) - I_0^*(\theta_n^*(\boldsymbol{X}_n))\| \leq \sup_{\theta \in \mathcal{N}} \|J_n(\theta) - I_0^*(\theta)\| \xrightarrow{P} 0$$

So, we know that $J_n(\theta^*_{jn}(\boldsymbol{X}_n)) \overset{P(\theta_0)}{\longrightarrow} I(\theta_0)$ for all j. This implies that $J^*_n(\boldsymbol{X}_n) \overset{P(\theta_0)}{\longrightarrow} I(\theta_0)$.