S&DS 242/542: Theory of Statistics

Lecture 14: Consistency and asymptotic normality of the MLE

Recap: Consistency and asymptotic normality

Given data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ from a parametric model, an estimator $\hat{\theta}$ for θ is *consistent* if

$$\hat{\theta} \to \theta$$

in probability as $n \to \infty$.

It is *asymptotically normal* if, furthermore, for some asymptotic variance $v(\theta)$,

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, v(\theta))$$

in distribution as $n \to \infty$.

We showed last lecture why a method of moments estimator $\hat{\theta}$ for a parameter $\theta \in \mathbb{R}$ is usually asymptotically normal, and used the delta method to derive its asymptotic variance.

Consistency and asymptotic normality of the MLE

Theorem

Let $f(x \mid \theta)$ be a parametric model, with a single parameter $\theta \in \mathbb{R}$. Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, and let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be the MLE. Under regularity conditions for $f(x \mid \theta)$, as $n \to \infty$,[1]

(a) *$\hat{\theta}$ is consistent.*

(b) *$\hat{\theta}$ is asymptotically normal, and $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \frac{1}{I(\theta)})$.*

The function $I(\theta)$ in this asymptotic variance has the two equivalent forms

$$I(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta)\right]$$

where $\mathbb{E}_\theta$ and Var_θ denote expectation/variance over $X \sim f(x \mid \theta)$.

---

[1] In this course, we won't discuss the exact conditions, which are technical. Three of the conditions are that θ is not on the boundary of the parameter space, $\theta \mapsto \log f(x \mid \theta)$ is twice differentiable, and $I(\theta)$ is non-zero.

Fisher Information

The function

$$I(\theta) = \text{Var}_\theta \left[\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta) \right]$$

is called the **Fisher information**. The quantity $\frac{\partial}{\partial\theta} \log f(X \mid \theta)$ is called the **score**.

The first expression for $I(\theta)$ states that the Fisher information is the variance of the score. We will see that the score has mean zero under the true parameter θ:

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right] = 0$$

So this first expression may also be written as

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right)^2 \right].$$

Example of the Poisson model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$. The MLE is $\hat{\lambda} = \bar{X}$.

$$f(x|\lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \Rightarrow \log f(x|\lambda) = -\lambda + x \log \lambda - \log(x!)$$

The score is: $\frac{\partial}{\partial \lambda} \log f(x|\lambda) = -1 + \frac{x}{\lambda}$

$$\mathbb{E}_\lambda\left[\frac{\partial}{\partial \lambda} \log f(X|\lambda)\right] = \mathbb{E}_\lambda\left[-1 + \frac{X}{\lambda}\right] = -1 + \frac{\mathbb{E}_\lambda[X]}{\lambda} = -1 + \frac{\lambda}{\lambda} = 0$$

$$I(\lambda) = \mathrm{Var}_\lambda\left[\frac{\partial}{\partial \lambda} \log f(X|\lambda)\right] = \mathrm{Var}_\lambda\left[-1 + \frac{X}{\lambda}\right] = \frac{1}{\lambda^2} \mathrm{Var}_\lambda[X] = \frac{1}{\lambda}.$$

Alternatively: $\frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) = -\frac{x}{\lambda^2}$

$$I(\lambda) = -\mathbb{E}_\lambda\left[\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda)\right] = \mathbb{E}_\lambda\left[\frac{X}{\lambda^2}\right] = \frac{1}{\lambda^2} \mathbb{E}_\lambda[X] = \frac{1}{\lambda}.$$

So $\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}\left(0, \frac{1}{I(\lambda)}\right) = \mathcal{N}(0, \lambda)$.

Example of the Pareto model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(\theta, 1)$. The MLE is $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i}$.

$f(x|\theta) = \frac{\theta}{x^{\theta+1}}$ (for $x>1$) $\Rightarrow \log f(x|\theta) = \log \theta - (\theta+1) \log x$.

$\frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{1}{\theta} - \log x$ ← This is the score.

Recall: $\mathbb{E}_\theta[\log X] = \frac{1}{\theta}$. So $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right] = \frac{1}{\theta} - \frac{1}{\theta} = 0$.

$\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) = -\frac{1}{\theta^2}$

$\Rightarrow I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)\right] = \frac{1}{\theta^2}$.

Example of the Pareto model

Alternatively: $I(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \mathrm{Var}_\theta[\log X]$.

$$\mathbb{E}_\theta\left[(\log X)^2\right] = \int_1^\infty (\log x)^2 \cdot \frac{\theta}{x^{\theta+1}}\,dx \quad (u = \log x,\ x = e^u)$$

$$= \int_0^\infty u^2 \cdot \frac{\theta}{e^{u(\theta+1)}} e^u\,du$$

$$= \int_0^\infty u^2 \cdot \underbrace{\theta e^{-\theta u}}_{\text{PDF of Exponential}(\theta)}\,du = \ldots = \frac{2}{\theta^2}$$

$$\Rightarrow \mathrm{Var}_\theta[\log X] = \mathbb{E}_\theta\left[(\log X)^2\right] - (\mathbb{E}_\theta \log X)^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

$$\Rightarrow I(\theta) = \frac{1}{\theta^2}$$

So $\sqrt{n}(\hat{\theta} - \theta) \rightarrow \mathcal{N}\left(0, \frac{1}{I(\theta)}\right) = \mathcal{N}(0, \theta^2)$.

Comparison with method of moments

Recall from last lecture that for the method of moments estimator, we instead had (when $\theta > 2$)

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}\left(0, \frac{\theta(\theta-1)^2}{\theta-2}\right)$$

For any $\theta > 2$, it holds that

$$\frac{\theta(\theta-1)^2}{\theta-2} > \theta^2.$$

So when n is large, the method-of-moments estimator has larger standard error than the MLE. This is a general phenomenon, which we will discuss next lecture.

Plug-in estimator for the Pareto mean

Suppose we are interested in estimating the mean $\frac{\theta}{\theta-1}$ of this Pareto distribution, instead of θ. A plug-in estimator based on the MLE $\hat{\theta}$ would be $\frac{\hat{\theta}}{\hat{\theta}-1}$. To compute its asymptotic variance:

Apply delta method w/ $g(x) = \frac{x}{x-1} = 1 + \frac{1}{x-1}$, $g'(x) = -\frac{1}{(x-1)^2}$

$$\sqrt{n}\Big(\frac{\hat{\theta}}{\hat{\theta}-1} - \underbrace{\frac{\theta}{\theta-1}}_{\substack{\text{True mean} \\ \text{of Pareto}(\theta,1)}}\Big) \to \mathcal{N}\big(0, \theta^2 \cdot g'(\theta)^2\big)$$

$$= \mathcal{N}\Big(0, \frac{\theta^2}{(\theta-1)^4}\Big)$$

Comparison to the sample mean

The plug-in estimate $\frac{\hat{\theta}}{\hat{\theta}-1}$ is not the only reasonable estimate for the mean of the Pareto distribution: What if we decided to simply use the sample mean $\bar{X}$?

For this estimate $\bar{X}$, the CLT shows

$$\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta - 1}\right) \to \mathcal{N}\left(0, \frac{\theta}{(\theta-1)^2(\theta-2)}\right)$$

where $\frac{\theta}{(\theta-1)^2(\theta-2)}$ is the variance of the Pareto distribution (again assuming $\theta > 2$).

It may be checked that this variance is greater than the variance $\frac{\theta^2}{(\theta-1)^4}$ for the plug-in estimate using the MLE.

Comparison to the sample mean

When n is large, the plug-in estimate $\frac{\hat{\theta}}{\hat{\theta}-1}$ using the MLE $\hat{\theta}$ is more accurate than the sample mean $\bar{X}$.

In the Pareto model, one intuition is that the distribution is heavy-tailed, and the sample mean $\bar{X}$ is heavily influenced by rare but large data values. In contrast, $\hat{\theta}$ estimate the shape of the Pareto distribution in a more robust way, and then estimates the mean from its relationship to the shape of the distribution.

A downside of this plug-in approach is that it is model-dependent: The estimate $\frac{\hat{\theta}}{\hat{\theta}-1}$ relies strongly on the correctness of the Pareto model, whereas $\bar{X}$ would be a reasonable estimate of the mean of the data distribution even if the Pareto model doesn't hold true.

Proof sketch: consistency of the MLE

To explain why the MLE $\hat{\theta}$ is consistent, recall that $\hat{\theta}$ is the value of θ which maximizes

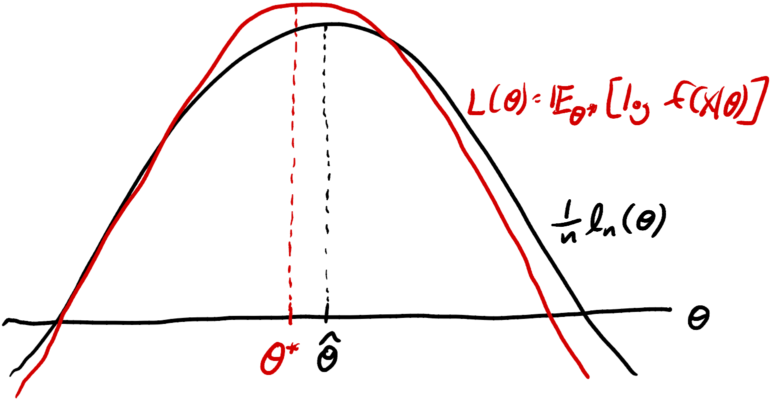
$$\frac{1}{n}\,\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta)$$

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta^*)$ with true parameter θ^*. Fixing *any* θ (not necessarily θ^*), the above is the sample average of n IID random variables, so the LLN implies

$$\frac{1}{n}\,\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta) \to \mathbb{E}_{\theta^*}[\log f(X \mid \theta)]$$

Here $\mathbb{E}_{\theta^*}[\log f(X \mid \theta)]$ is the expected value of each $\log f(X_i \mid \theta)$ when $X_i \sim f(x \mid \theta^*)$, where the log-likelihood is evaluated at an arbitrary parameter θ which may be different from θ^*.

Proof sketch: consistency of the MLE



Under suitable conditions, the value of θ maximizing $\frac{1}{n}\ell_n(\theta)$ (which is the MLE $\hat{\theta}$) converges in probability to the value of θ maximizing the limiting function $L(\theta) = \mathbb{E}_{\theta^*}[\log f(X \mid \theta)]$.

Proof sketch: consistency of the MLE

We claim that θ maximizing $L(\theta) = \mathbb{E}_{\theta^*}[\log f(X \mid \theta)]$ is exactly the true parameter θ^*:

$$L(\theta) - L(\theta^*) = \mathbb{E}_{\theta^*}\left[\log f(X|\theta) - \log f(X|\theta^*)\right]$$

$$= \mathbb{E}_{\theta^*}\left[\log \frac{f(X|\theta)}{f(X|\theta^*)}\right]$$

The function $\log x$ is concave. So by Jensen's inequality

$$L(\theta) - L(\theta^*) \leq \log \mathbb{E}_{\theta^*}\left[\frac{f(X|\theta)}{f(X|\theta^*)}\right]$$

$$= \log \int \frac{f(x|\theta)}{\cancel{f(x|\theta^*)}} \cdot \cancel{f(x|\theta^*)}\, dx$$

Proof sketch: consistency of the MLE

$$L(\theta) - L(\theta^*) \leq \log \int \underbrace{f(x|\theta)}_{\text{This is a PDF, so } \int f(x|\theta)dx = 1} dx = 0$$

So $L(\theta) - L(\theta^*) \leq 0$ for every θ, meaning that $L(\theta)$ is maximized at θ^*. This explains the consistency of $\hat{\theta}$.

Proof sketch: Definition of Fisher information

Next, let us check that the two definitions of $I(\theta)$ are the same, and that the score has mean zero:

Applying again $\int f(x|\theta)dx = 1$:

$$0 = \frac{\partial}{\partial\theta}\int f(x|\theta)dx$$

$$= \int \frac{\partial}{\partial\theta} f(x|\theta)dx$$

Note that $\frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}$

$$\Rightarrow 0 = \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]\cdot f(x|\theta)\,dx$$

$$= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]$$

Proof sketch: Definition of Fisher information

Differentiating again in θ:

$$0 = \frac{\partial}{\partial\theta} \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] \cdot f(x|\theta)\, dx$$

$$= \int \frac{\partial}{\partial\theta} \left\{ \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] \cdot f(x|\theta) \right\} dx$$

$$= \int \left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right] \cdot f(x|\theta)\, dx$$

$$+ \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] \cdot \underbrace{\frac{\partial}{\partial\theta} f(x|\theta)}_{= \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] \cdot f(x|\theta)} dx$$

$$= \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)\right] + \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)^2\right]$$

Proof sketch: Definition of Fisher information

$$\Rightarrow -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$$

$$= \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]$$

b/c score has mean 0.

These are our two definitions of Fisher information, $I(\theta)$.

Proof sketch: Asymptotic normality

Finally, let us show the convergence in distribution

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \tfrac{1}{I(\theta)})$$

when the true parameter is θ.

The MLE $\hat{\theta}$ maximizes $l_n(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$. So $0 = l_n'(\hat{\theta})$. Since $\hat{\theta} - \theta \to 0$ in probability as $n \to \infty$, for large n, $\hat{\theta}$ is close to θ. Taylor expand around θ:

$$0 = l_n'(\hat{\theta}) \approx l_n'(\theta) + l_n''(\theta) \cdot (\hat{\theta} - \theta)$$

$$\Rightarrow -l_n''(\theta) \cdot (\hat{\theta} - \theta) \approx l_n'(\theta)$$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta) \approx -\frac{\sqrt{n}}{l_n''(\theta)} \cdot l_n'(\theta) = \frac{l_n'(\theta)/\sqrt{n}}{-l_n''(\theta)/n}$$

Proof sketch: Asymptotic normality

$$\sqrt{n}(\hat{\theta}-\theta) \approx \frac{l_n'(\theta)/\sqrt{n}}{-l_n''(\theta)/n}$$

Recall $l_n(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$

For the denominator,

$$-\frac{l_n''(\theta)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right)$$

$$\rightarrow \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] \text{ by LLN}$$

$$= I(\theta) \quad \text{(in probability)}$$

Proof sketch: Asymptotic normality

$$\sqrt{n}(\hat{\theta}-\theta) \approx \frac{l_n'(\theta)/\sqrt{n}}{-l_n''(\theta)/n}$$

For the numerator:

$$\frac{l_n'(\theta)}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial\theta} \log f(X_i|\theta)}_{\text{score evaluated at } X_i}$$

$$\rightarrow \mathcal{N}\left(0, \operatorname{Var}_\theta\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]\right) \text{ by CLT}$$

(in distribution)

$$= \mathcal{N}(0, I(\theta))$$

By Slutsky's lemma: $\sqrt{n}(\hat{\theta}-\theta) \rightarrow \frac{1}{I(\theta)} \cdot \mathcal{N}(0, I(\theta)) = \mathcal{N}(0, \frac{1}{I(\theta)})$.