STSCI 5080
Probability Models and Inference
Lecture 20: Maximal Likelihood Estimation

November 8, 2018

Setting

- Let $\{f_\theta \mid \theta \in \Theta\}$ be a class of pmfs/pdfs where $\Theta \subset \mathbb{R}^k$, and suppose that
$$X_1, \ldots, X_n \sim f_\theta \text{ i.i.d.}$$
for some $\theta \in \Theta$.
- The likelihood function is
$$L_n(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$
- The log likelihood function is
$$\ell_n(\theta) = \log L_n(\theta).$$
- The MLE is a maximizer of the log likelihood function:
$$\ell_n(\widehat{\theta}) = \max_{\theta \in \Theta} \ell_n(\theta).$$

In the one-dimensional case ($k = 1$), the MLE is obtained by solving the first order condition (FOC) w.r.t. θ:

$$\ell_n'(\theta) = 0.$$

Example 20.1

Example

Let $X \sim Bin(n, p)$ for some $0 < p < 1$.

(a) Find the log likelihood function for p.

(b) Find the FOC for the MLE of p.

(c) Find the MLE.

The sample size is 1 in this example. The pmf of X is

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The likelihood function is

$$L(p) = \binom{n}{X} p^X (1-p)^{n-X}.$$

The log likelihood function is

$$\ell(p) = \log \binom{n}{X} + X \log p + (n - X) \log(1-p).$$

We note that

$$\ell'(p) = \frac{X}{p} - \frac{n - X}{1 - p}.$$

So the FOC is

$$\frac{X}{p} - \frac{n - X}{1 - p} = 0.$$

Solving the FOC w.r.t. p, we obtain the MLE:

$$\widehat{p} = \frac{X}{n}.$$

Functions of MLE

Definition

Let $\widehat{\theta}$ be the MLE of θ. Then the MLE of $g(\theta)$ is $g(\widehat{\theta})$.

Example 20.2

Example

Let $X_1, \dots, X_n \sim Po(\lambda)$ i.i.d. for some λ. The MLE is

$$\widehat{\lambda} = \overline{X}.$$

We want to estimate

$$\theta = g(\lambda) = P_\lambda(X_1 = 0) = e^{-\lambda}.$$

Then the MLE of θ is

$$\widehat{\theta} = g(\widehat{\lambda}) = e^{-\overline{X}}.$$

Example: Number of misprints

- The publisher checks the numbers of misprints of copies of a book, and they fit a Poisson distribution.
- Suppose that the total number of misprints is 4 among 52 copies.
- Find the MLE of the probability that there is no misprint in a randomly chosen copy.

- Let X_i denote the number of misprints of the i-th copy, and we have

$$X_1, \ldots, X_n \sim Po(\lambda) \text{ i.i.d.}$$

  The MLE of λ is

$$\widehat{\lambda} = \overline{X} = \frac{4}{52} = \frac{1}{13}.$$

- The MLE of $\theta = P_\lambda(X_1 = 0) = e^{-\lambda}$ is

$$\widehat{\theta} = e^{-\overline{X}} = e^{-1/13} \approx 0.93.$$

Example: Number of penalty shootouts

- You are a big fan of a soccer team in the English premier league.
- There are 7 fouls committed by your team that led to penalty shootouts among 38 games.
- Find the MLE of the probability that your team commits no such fouls in a randomly chosen game.

Example: Number of penalty shootouts

- You are a big fan of a soccer team in the English premier league.
- There are 7 fouls committed by your team that led to penalty shootouts among 38 games.
- Find the MLE of the probability that your team commits no such fouls in a randomly chosen game.
- Answer:

$$e^{-7/38} \approx 0.816.$$

List of MLEs

- $Po(\lambda)$: $\widehat{\lambda} = \overline{X}$.
- $N(\mu, \sigma_0^2)$ (where σ_0^2 is known): $\widehat{\mu} = \overline{X}$.
- $N(0, \sigma^2)$: $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} X_i^2$.
- $Ex(\lambda)$: $\widehat{\lambda} = 1/\overline{X}$.
- $Bin(n, p)$ (where $X \sim Bin(n, p)$): $\widehat{p} = X/n$.

Convergence in probability and in distribution

- Let Y_n and Y be random variables with cdfs F_n and F, respectively.
- Y_n converges in probability to Y, denoted as $Y_n \xrightarrow{P} Y$, if

$$\lim_{n\to\infty} P(|Y_n - Y| > \varepsilon) = 0$$

for any $\varepsilon > 0$.

- Y_n converges in distribution to Y, denoted as $Y_n \xrightarrow{d} Y$, if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for any continuity point of F. If $Y \sim N(0, \sigma^2)$ e.g., we also write

$$Y_n \xrightarrow{d} N(0, \sigma^2).$$

Asymptotic properties of MLE

Definition

Suppose $k = 1$ (i.e., θ is one-dim.). An estimator $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ is consistent for θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta$$

as $n \to \infty$ whatever the value of θ is.
The estimator $\widehat{\theta}_n$ is asymptotically normal if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

as $n \to \infty$, where $\sigma^2(\theta) > 0$.

Consistency

- Consistency is the minimum requirement for any reasonable estimator.
- Given $X_1, \ldots, X_n \sim Po(\lambda)$ i.i.d., the MLE of λ is $\widehat{\lambda}_n = \overline{X}_n$. But you know that λ is between 0 and 100, and $\widehat{\lambda}_n$ looks too small to you. So you come up with a new estimator

$$\widetilde{\lambda}_n = 10\sqrt{\widehat{\lambda}_n}.$$

Is $\widetilde{\lambda}_n$ consistent?

Consistency

- Consistency is the minimum requirement for any reasonable estimator.

- Given $X_1, \ldots, X_n \sim Po(\lambda)$ i.i.d., the MLE of λ is $\widehat{\lambda}_n = \overline{X}_n$. But you know that λ is between 0 and 100, and $\widehat{\lambda}_n$ looks too small to you. So you come up with a new estimator

$$\widetilde{\lambda}_n = 10\sqrt{\widehat{\lambda}_n}.$$

  Is $\widetilde{\lambda}_n$ consistent?

- Answer: No. You should not use $\widetilde{\lambda}_n$.

Case I: MLE = sample mean

- Consider the case where the MLE coincides with the sample mean:
$$\widehat{\theta}_n = \overline{X}_n.$$
In addition, suppose that $E_\theta(X_1) = \theta$ and $\mathrm{Var}_\theta(X_1) = \sigma^2(\theta)$. E.g. $N(\mu, \sigma_0^2)$ and $Po(\lambda)$.

- By LLN and CLT, the MLE is consistent and asymptotically normal:
$$\widehat{\theta}_n \xrightarrow{P} \theta,$$
$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

Example 20.2

Example

Suppose that

$$X_1, \ldots, X_n \sim Po(\lambda) \text{ i.i.d.}$$

for some $\lambda > 0$. The MLE is

$$\widehat{\lambda}_n = \overline{X}_n.$$

Because $E_\lambda(X_1) = \lambda$ and $\mathrm{Var}_\lambda(X_1) = \lambda$, we have

$$\widehat{\lambda}_n \xrightarrow{P} \lambda,$$

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda).$$

Case II: Exponential distribution

- Suppose that
$$X_1, \ldots, X_n \sim Ex(\lambda) \text{ i.i.d.}$$
for some $\lambda > 0$.
- The MLE is
$$\widehat{\lambda}_n = \frac{1}{\overline{X}_n}.$$

- Since
$$E_\lambda(X_1) = 1/\lambda \quad \text{and} \quad \mathrm{Var}_\lambda(X_1) = 1/\lambda^2,$$
the LLN and CLT imply that
$$\overline{X}_n \xrightarrow{P} \frac{1}{\lambda} \quad \text{and} \quad \sqrt{n}(\overline{X}_n - 1/\lambda) \xrightarrow{d} N(0, 1/\lambda^2).$$
- By the continuous mapping theorem and delta method,
$$\widehat{\lambda}_n \xrightarrow{P} \lambda,$$
$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda^2).$$

Delta method

Theorem

If $\sqrt{n}(Y_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ *and* $g(x)$ *is differentiable at* $x = \mu$*, then*

$$\sqrt{n}\{g(Y_n) - g(\mu)\} \xrightarrow{d} N(0, \{g'(\mu)\}^2\sigma^2).$$

Case III: General case (not included in Final)

Theorem

In general, the MLE $\widehat{\theta}_n$ is consistent and asymptotically normal under suitable regularity conditions:

$$\widehat{\theta}_n \xrightarrow{P} \theta,$$

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, 1/I(\theta)),$$

where $I(\theta)$ is the Fisher information:

$$I(\theta) = E_\theta\left[-\frac{\partial^2 \log f_\theta(X_1)}{\partial \theta^2}\right].$$

Optimality of MLE

- The previous theorem suggests that

$$E_\theta(\widehat{\theta}_n) \approx \theta \quad \text{(MLE is approximately unbiased)}$$

$$\mathrm{Var}_\theta(\sqrt{n}(\widehat{\theta}_n - \theta)) \approx \frac{1}{I(\theta)}, \text{ i.e., } \mathrm{Var}_\theta(\widehat{\theta}_n) \approx \frac{1}{nI(\theta)}.$$

- Let $\widetilde{\theta}_n$ be another estimator for θ such that

$$\widetilde{\theta}_n \xrightarrow{P} \theta,$$

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} N(0, \tau^2(\theta)).$$

  Under regularity conditions,

$$\tau^2(\theta) \geq \frac{1}{I(\theta)}$$

  for any $\theta \in \Theta$.

Recap

Optimality of MLE

Under regularity conditions, the MLE is approximately unbiased and achieves the minimum variance among "reasonable" estimators.

See Chapter 8.5.2 in Rice for the detail.