Biostat 602 Winter 2017

Lecture Set 7

Point Estimation

Maximum Likelihood Estimation

Reading: CB 7.2

Maximum Likelihood Estimation

Recap

$X_1, \cdots, X_n$ *i.i.d.* $f_X(x|\theta)$. The joint distribution of $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

Given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by $L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$ is called the **likelihood function**.

For a given sample point $\mathbf{x} = (x_1, \cdots, x_n)$, let $\hat{\theta}(\mathbf{x})$ be the value such that $L(\theta|\mathbf{x})$ attains its maximum. More formally,

$$L(\hat{\theta}(\mathbf{x})|\mathbf{x}) \geq L(\theta|\mathbf{x}) \quad, \forall \theta \in \Omega, \quad \text{where} \quad \hat{\theta}(\mathbf{x}) \in \Omega.$$

$\hat{\theta}(\mathbf{x})$ is called the *maximum likelihood estimate* of θ based on data $\mathbf{x}$, and

$\hat{\theta}(\mathbf{X})$ is the *maximum likelihood estimator (MLE)* of θ.

Strategies for finding MLE of θ

There are two situations.

If the function is differentiable with respect to θ

1. Find candidates that makes first order derivative to be zero
2. Check second-order derivative to check local maximum.
   - For one-dimensional parameter, $\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$ implies local maximum.
   - For two-dimensional parameter, we need to show
     (a) $\partial^2 L(\theta_1, \theta_2)/\partial \theta_1^2 < 0$ or $\partial^2 L(\theta_1, \theta_2)/\partial \theta_2^2 < 0$.
     (b) Determinant of second-order derivative is positive
3. Check whether boundary gives global maximum.
   - Or clearly justify that boundaries cannot be global maximum.

If the function is NOT differentiable with respect to θ

- Use numerical methods, or
- Directly maximize using inequalities or properties of the function.

In general, one is content with MLEs that are local maximum.

Example 1 – Normal MLEs, both parameters unknown

Let $X_1, \cdots, X_n$ be $i.i.d$ observations from $\mathcal{N}(\mu, \sigma^2)$. Find MLE of (μ, σ^2).

Two possible approaches

- Use second-order partial derivatives and their Hessian to show global maximum
- Find a workaround to avoid complex calculations.

Common step : Calculate first-order derivatives

Likelihood Function

$$\begin{aligned}
L(\mu, \sigma^2|\mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\
l(\mu, \sigma^2|\mathbf{x}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}
\end{aligned}$$

Partial derivative with respect to μ

$$\begin{aligned}
L(\mu, \sigma^2|\mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\
l(\mu, \sigma^2|\mathbf{x}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \\
\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}
\end{aligned}$$

partial derivative with respect to σ^2

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\end{aligned}$$

Checking second-order partial derivatives

With respect to μ

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$$

With respect to σ^2

$$\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

With respect to both parameters

$$\frac{\partial^2 l}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)$$

Calculate Hessian

$$\begin{aligned} & \left| \begin{array}{cc} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l}{\partial (\sigma^2)^2} \end{array} \right|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma^2}} \\ = & \left| \begin{array}{cc} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{array} \right|_{\mu=\overline{x}, \sigma^2=\hat{\sigma^2}} \\ = & \left. \frac{1}{\sigma^6}\left[-\frac{n^2}{2} + \frac{n}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu)\right)^2\right] \right|_{\mu=\overline{x}, \sigma^2=\hat{\sigma^2}} \\ = & \frac{1}{\hat{\sigma^6}}\left[-\frac{n^2}{2} + \frac{n}{\hat{\sigma^2}}(n\hat{\sigma^2}) - \frac{1}{\hat{\sigma^2}}\left(\sum_{i=1}^{n}(x_i - \overline{x})\right)^2\right] = \frac{1}{\hat{\sigma^6}}\frac{n^2}{2} > 0 \end{aligned}$$

Thus, the conditions for local (interior) maximum is indeed found. Because this is a unique interior maximum, it is also a global maximum. Therefore, $(\hat{\mu}, \hat{\sigma^2}) = (\overline{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2)$ is an MLE.

A simpler workaround

First, fix one parameter, say σ^2.

$$l(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}$$

If

$$\mu \neq \overline{x}, \quad \text{then} \quad \sum_{i=1}^n (x_i - \mu)^2 > \sum_{i=1}^n (x_i - \overline{x})^2$$

so $\hat{\mu} = \overline{x}$ must hold to maximize the log-likelihood.

Second, reduce the problem into one-parameter maximization
Given $\hat{\mu} = \overline{x}$, the log-likelihood is maximized at $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$, because

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^4}(\sigma^2 - \hat{\sigma}^2)$$

is always positive when $\sigma^2 < \hat{\sigma}^2$ and always negative when $\sigma^2 > \hat{\sigma}^2$. Hence l as a function of σ^2 increases upto $\hat{\sigma}^2$ and then decreases.

Therefore, $(\hat{\mu}, \hat{\sigma^2}) = (\overline{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2)$ is an MLE.

Example 2 – Ranged Normal with Known Variance

Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\mu, 1)$ where $\underline{\mu \geq 0}$. Find MLE of μ.

Solution:

$$\begin{aligned} L(\mu|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2}\right] = (2\pi)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right] \\ l(\mu|\mathbf{x}) &= \log L(\mu, \mathbf{x}) = C - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} \\ \frac{\partial l}{\partial \mu} &= \frac{2\sum_{i=1}^{n}(x_i-\mu)}{2} = 0, \qquad \frac{\partial^2 l}{\partial \mu^2} < 0 \\ \hat{\mu} &= \sum_{i=1}^{n} x_i/n = \overline{x} \end{aligned}$$

Question: ARE WE DONE?

The MLE parameter must be within the parameter space.

We need to check whether $\hat{\mu}$ is within the parameter space $[0, \infty)$.

- If $\overline{x} \geq 0$, $\hat{\mu} = \overline{x}$ falls into the parameter space.
- If $\overline{x} < 0$, $\hat{\mu} = \overline{x}$ does NOT fall into the parameter space.

When $\overline{x} < 0$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}(x_i - \mu) = n(\overline{x} - \mu) < 0$$

for $\mu \geq 0$. Therefore, $l(\mu|\mathbf{x})$ is a decreasing function of μ. So $\hat{\mu} = 0$ when $\overline{x} < 0$.

Therefore, MLE is

$$\hat{\mu}(\mathbf{X}) = \max(\overline{X}, 0)$$

Example 3 – Binomial MLE, unknown number of trials

Let $X_1, \cdots, X_n$ be random sample from $Binomial(k, p)$ population, where p is known and k is unknown. Find the MLE of k.

Likelihood Function

$$L(k|\mathbf{x}, p) = \begin{cases} \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i} & (k \geq \max_i x_i) \\ 0 & (k < \max_i x_i) \end{cases}$$

The likelihood function is not differentiable with respect to k because k is an integer.

So how can we find MLE?

Idea: Instead of differentiating, take a ratio

We want to find k such that

$$\frac{L(k|\mathbf{x}, p)}{L(k-1|\mathbf{x}, p)} \geq 1 \qquad \text{and} \qquad \frac{L(k+1|\mathbf{x}, p)}{L(k|\mathbf{x}, p)} < 1$$

$$\begin{aligned} \frac{L(k, \mathbf{x}, p)}{L(k-1, \mathbf{x}, p)} &= \frac{\prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}}{\prod_{i=1}^n \binom{k-1}{x_i} p^{x_i}(1-p)^{k-1-x_i}} \\ &= \frac{\prod_{i=1}^n \frac{k!}{x_i!(k-x_i)!} p^{x_i}(1-p)^{k-x_i}}{\prod_{i=1}^n \frac{(k-1)!}{x_i!(k-1-x_i)!} p^{x_i}(1-p)^{k-1-x_i}} \\ &= \prod_{i=1}^n \frac{k(1-p)}{k-x_i} = \frac{k^n(1-p)^n}{\prod_{i=1}^n (k-x_i)} \end{aligned}$$

Finding MLE

Find maximum k such that $\frac{L(k|\mathbf{x},p)}{L(k-1|\mathbf{x},p)} \geq 1$

$$\begin{aligned} k^n(1-p)^n &\geq \prod_{i=1}^{n}(k-x_i) \\ (1-p)^n &\geq \prod_{i=1}^{n}\left(1-\frac{x_i}{k}\right) \qquad (k \geq \max_i x_i) \end{aligned} \tag{1}$$

- The right-hand side is an increasing function of k
- The right-hand side is $0 < (1-p)^n$, when $k = \max_i x_i$.
- The right-hand side will converge to $1 > (1-p)^n$ as $k \to \infty$.
- Thus, there is a unique maximum for the likelihood function.
- $\hat{k}_{MLE}$ can be numerically solved as the maximum k satisfying (1).

Example 4 Let $X_1, \cdots, X_n$ be a random sample from a pdf

$$f_X(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad 0 < \theta < \infty.$$

(a) Find method of moments estimator for θ.

(b) Find the MLE of θ.

Example 5 – Two-parameter Exponential

Let $X_1, \cdots, X_n$ be *i.i.d.* observations from a location-scale family of an exponential distribution with pdf

$$f_X(x|\theta) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right), \qquad x \geq \mu, \sigma > 0$$

(a) Find MLEs of μ and σ.

(b) Find MLE of $S(t) = \Pr(X > t)$ for a fixed t.

Invariance

MLE is invariant under monotonic transfotmation.

Question: If $\hat{\theta}$ is the MLE of θ, what is the MLE of $\tau(\theta)$?

Example 6: Let $X_1, \cdots, X_n$ be a random sample from $Bernoulli(p)$ where $0 < p < 1$.

1. What is the MLE of p?
2. What is the MLE of odds, defined by $\eta = p/(1-p)$?

MLE of p

$$\begin{aligned}
L(p|\mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i} \\
l(p|\mathbf{x}) &= \log p \sum_{i=1}^{n} x_i + \log(1-p)(n - \sum_{i=1}^{n} x_i) \\
\frac{\partial l}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\hat{p} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}$$

MLE of $\eta = \frac{p}{1-p}$

- $\eta = p/(1-p) = \tau(p)$
- $p = \eta/(1+\eta) = \tau^{-1}(\eta)$

$$
\begin{aligned}
L^*(\eta|\mathbf{x}) &= p^{\sum x_i}(1-p)^{n-\sum x_i} \\
&= \frac{p}{1-p}^{\sum x_i}(1-p)^n = \frac{\eta^{\sum x_i}}{(1+\eta)^n} \\
l^*(\eta|\mathbf{x}) &= \sum_{i=1}^{n} x_i \log \eta - n \log(1+\eta) \\
\frac{\partial l^*}{\partial \eta} &= \frac{\sum_{i=1}^n x_i}{\eta} - \frac{n}{1+\eta} = 0 \\
\hat{\eta} &= \frac{\sum_{i=1}^n x_i/n}{1-\sum_{i=1}^n x_i/n} = \frac{\overline{x}}{1-\overline{x}} = \tau(\hat{p})
\end{aligned}
$$

Another way to get MLE of $\eta = \frac{p}{1-p}$

$$
L^*(\eta|\mathbf{x}) = \frac{\eta^{\sum x_i}}{(1+\eta)^n}
$$

- From MLE of $\hat{p}$, we know $L^*(\eta|\mathbf{x})$ is maximized when $p = \eta/(1+\eta) = \hat{p}$.
- Equivalently, $L^*(\eta|\mathbf{x})$ is maximized when $\eta = \hat{p}/(1-\hat{p}) = \tau(\hat{p})$, because τ is a one-to-one function.
- Therefore $\hat{\eta} = \tau(\hat{p})$.

Result: Denote the MLE of θ by $\hat{\theta}$. If $\tau(\theta)$ is an one-to-one function of θ, then MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Proof: The likelihood function in terms of $\tau(\theta) = \eta$ is

$$\begin{aligned} L^*(\tau(\theta)|\mathbf{x}) &= \prod_{i=1}^n f_X(x_i|\theta) = \prod_{i=1}^n f(x_i|\tau^{-1}(\eta)) \\ &= L(\tau^{-1}(\eta)|\mathbf{x}) \end{aligned}$$

We know this function is maximized when $\tau^{-1}(\eta) = \hat{\theta}$, or equivalently, when $\eta = \tau(\hat{\theta})$. Therefore, MLE of $\eta = \tau(\theta)$ is $\hat{\eta} = \tau(\hat{\theta})$.

Induced Likelihood Function

- Let $L(\theta|\mathbf{x})$ be the likelihood function for a given data $x_1, \cdots, x_n$,
- and let $\eta = \tau(\theta)$ be a (possibly not a one-to-one) function of θ.

We define the *induced likelihood function* L^* by

$$L^*(\eta|\mathbf{x}) = \sup_{\theta \in \tau^{-1}(\eta)} L(\theta|\mathbf{x})$$

where $\tau^{-1}(\eta) = \{\theta : \tau(\theta) = \eta, \ \theta \in \Omega\}$.

- The value of η that maximize $L^*(\eta|\mathbf{x})$ is called the MLE of $\eta = \tau(\theta)$.

Theorem 7.2.10: If θ is the MLE of $\hat{\theta}$, then the MLE of $\eta = \tau(\theta)$ is $\tau(\hat{\theta})$, where $\tau(\theta)$ is any function of θ.

Proof - Using Induced Likelihood Function

$$
\begin{aligned}
L^*(\hat{\eta}|\mathbf{x}) &= \sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} \sup_{\theta \in \tau^{-1}(\eta)} L(\theta|\mathbf{x}) \\
&= \sup_{\theta} L(\theta|\mathbf{x}) = L(\hat{\theta}|\mathbf{x}) \\
L(\hat{\theta}|\mathbf{x}) &= \sup_{\theta \in \tau^{-1}(\tau(\hat{\theta}))} L(\theta|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]
\end{aligned}
$$

Hence, $L^*(\hat{\eta}|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]$ and $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

Properties of MLE

1. Optimal in some sense : We will study this later
2. By definition, MLE will always fall into the range of the parameter space.
3. Not always easy to obtain; may be hard to find the global maximum.
4. Heavily depends on the underlying distributional assumptions (i.e. not robust).