Maximum Likelihood Estimation (MLE)

It is the most important and widespread method of estimation (in parametric models) Many estimators such as the sample mean are MLE. ML estimation is very useful in practice and tends to give more efficient estimates (having the smallest possible value of the MSE) than other methods.

Let $(X_1, \dots, X_n)$ be an i.i.d. (independent and identically distributed) sample and let $\theta = (\theta_1, \dots, \theta_p)^T$ be a vector of parameters of the sample distribution.

Denote by $f(x;\theta)$ the PDF of X_i (if X_i is a continuous random variable) or the PMF of X_i (if X_i is a discrete r.v.).

As the variables $X_1, \dots, X_n$ are independent, we can write the PDF of $(X_1, \dots, X_n)$ as

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) = f_{X_1}(x_1;\theta) \cdots f_{X_n}(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Definition

The function $L(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$ viewed as a function of θ with X_i fixed at the observed variable is called the <u>likelihood function</u> of the sample $(X_1, \dots, X_n)$.

The <u>maximum likelihood estimator</u> (MLE) is the value of θ that maximizes the likelihood function.

We denote the MLE by $\hat{\theta}_{ML}$

$L(\theta)$ tells us the likelihood of the sample that was actually observed. $\hat{\theta}_{ML}$ is the value of θ where the likelihood of the data is largest.

Examples:

1. $(X_1, \dots, X_n) \sim \text{Poisson}(\theta)$, $\theta > 0$.

$$f(x, \theta) = P(X_i = x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, 2, 3, \dots$$

$$L(\theta) = e^{-n\theta} \theta^{\sum_{i=1}^n X_i} \prod_{i=1}^n \frac{1}{X_i!}.$$

2. $(X_1, \dots, X_n) \sim \text{Exponential}(\theta)$, $\theta > 0$

$$f(x, \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), \quad \theta > 0$$

$$L(\theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^n X_i\right)$$

It is mathematically easier to maximize $\log(L(\theta))$.
Since $\theta \to \log(L(\theta))$ is increasing, maximizing the log-likelihood $\log L(\theta)$ is equivalent to maximizing $L(\theta)$.

1. $$\log L(\theta) = -n\theta + \sum_{i=1}^n X_i \log\theta + \underbrace{c(X_1, \dots, X_n)}_{\text{constant}}.$$

$$(\log L)'(\theta) = -n + \frac{\sum_{i=1}^n X_i}{\theta}$$

$$(\log L)'(\hat{\theta}) = 0 \implies \hat{\theta} = \frac{\sum X_i}{n} = \overline{X}$$

Likelihood equation

$\hat{\theta}$ is indeed a global maxima of $\log L$ since

$$(\log L)''(\theta) = -\frac{\sum X_i}{\theta^2} < 0 \quad \forall \theta$$

Therefore the MLE of θ is $\hat{\theta} = \overline{X}$

2. $$\log L(\theta) = -n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^n X_i$$

$$(\log L)'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum X_i$$

$$(\log L)'(\hat{\theta}) = 0 \implies \hat{\theta} = \overline{X}$$

$$(\log L)''(\hat{\theta}) = \frac{n}{\hat{\theta}^2} - \frac{2\sum X_i}{\hat{\theta}^3} = \frac{n}{(\overline{X})^2} - \frac{2n}{(\overline{X})^2} = -\frac{n}{(\overline{X})^2} < 0$$

Hence $\hat{\theta}_{ML} = \overline{X}$

Remarks :

- As $\hat{\theta}_{ML}$ is calculated from a random sample, $\hat{\theta}_{ML}$ is a random variable. It has a probability law, an expectation and a standard error.

Ex 1 $\hat{\theta}_{ML} = \overline{X}$. As $E(X_i) = Var(X_i) = \theta$, $E(\hat{\theta}_{ML}) = \theta$ and $Var(\hat{\theta}_{ML}) = \frac{\theta}{n}$

$s.e.(\hat{\theta}_{ML}) = \frac{\sqrt{\theta}}{n}$ can be estimated by $\widehat{s.e.}(\hat{\theta}_{ML}) = \frac{\sqrt{\hat{\theta}_{ML}}}{n} = \frac{\sqrt{\overline{X}}}{n}$

- It can be shown that, under certain conditions of the model, MLE have good properties

→ MLE are consistent

→ MLE are "equivariant" : if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the ME of $g(\theta)$

→ MLE are efficient, at least for large samples.

(as $n \to +\infty$, $Bias(\hat{\theta}_{ML}) \to 0$ and $\hat{\theta}_{ML}$ has the smallest variance)

Computation of the MLE

a/ To find the MLE, we can solve the likelihood equation $(\log \mathcal{L})'(\hat{\theta}) = 0$. The solution is a maxima if $(\log \mathcal{L})''(\hat{\theta}) < 0$.

If $\log \mathcal{L}$ is not derivable, you have to study the variations of the function. (example : $(X_1, \dots, X_n) \sim Uniform(0, \theta)$)

b/ In a multiparameter model, θ is a vector. The calculations can directly be extended.

Let $G(\theta) = \frac{\partial \log \mathcal{L}}{\partial \theta}(\theta) = \left(\frac{\partial \log \mathcal{L}}{\partial \theta_1}(\theta), \dots, \frac{\partial \log \mathcal{L}}{\partial \theta_p}(\theta) \right)^p$ be the gradient of the log-likelihood.

Let $H(\theta) = \frac{\partial^2 \log L}{\partial \theta^2}(\theta)$ be the hessian matrix $(p \times p)$.

with coefficient $H_{ij} = \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}(\theta)$.

To compute $\hat{\theta}_{ML}$, solve the likelihood equations $G(\hat{\theta}) = 0$.

You can verify that the solution is a maxima if $H(\hat{\theta})$ is negative definite. Example: $(X_1, \dots, X_n) \sim \mathcal{N}(\mu, \sigma^2)$; $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$

c/ Apart from simple sample models, it is not possible to find solutions of the likelihood equations with analytical expressions

Example: $(X_1, \dots, X_n) \sim \text{Gamma}(\alpha, \lambda)$, $\theta = \begin{pmatrix} \alpha \\ \lambda \end{pmatrix}$

$$f(x;\theta) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0$$

$$\log L(\theta) = n\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i - n \log \Gamma(\alpha)$$

Likelihood equations:

$$\begin{cases} \frac{\partial \log L}{\partial \alpha}(\theta) = n \log \lambda + \sum \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \\ \frac{\partial \log L}{\partial \lambda}(\theta) = \frac{n\alpha}{\lambda} - \sum X_i = 0 \end{cases}$$

From the second equation we find $\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}}$ and we can substitute $\hat{\lambda}$ into the first equation:

$$n \log \hat{\alpha} - n \log \bar{X} + \sum \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

But this equation cannot be solved in closed form and an iterative method for finding the roots has to be used.

In problems involving several parameters, we must solve systems of nonlinear equations to find the MLE's -

⟶ we use <u>iterative optimization algorithms</u>

They are implemented in python or R.

When using these algorithms, you need

- <u>to start</u> the iterative procedure. For example, you can use the method of moments estimator as an <u>initial value</u> -
- to check that the <u>algorithm has converged</u>.
- to check that the solution $\hat{\theta}$ is reasonable/correct:

  Is $G(\hat{\theta}) = 0$? $H(\hat{\theta}) < 0$? (check the outputs)

  Is $\hat{\theta}$ a global maximum? (change the initial value and restart the algo.)