11 Maximum Likihood Estimation (MLE)

MLE is one of the few statistical estimation methods which are independent of loss function. Because it is not derived from loss function, MLE is often inadmissible. And MLE might be biased. But the derivation of MLE is very intuitive, and asymptotic behavior of MLE is appealing (under some regularity conditions), hence the method of likelihood function is one of the most widely used statistical inference tools.

Motivating example:

Assume we observe a random variable from Bernoulli distribution with $p = 0.2$ or $p = 0.9$.

If the observed value is $X = 0$, it is more plausible that it comes from $Ber(0.2)$ than $Ber(0.9)$ since $P(X = 0; p = 0.2) = 0.8 > P(X = 0; p = 0.9) = 0.1$.

This idea can be extended to discrete distributions $P_\theta : \theta \in \Theta \subset \mathbb{R}^k$. If $X = x$ is observed, θ_1 is more plausible than θ_2, iff

$$P(X = x; \theta_1) > P(X = x; \theta_2).$$

Then we can estimate θ by maximizing $P(X = x; \theta)$ over $\theta \in \Theta$, if such an maximizer exists.

More generally, considering $P(X = x; \theta)$ as p.d.f with respect to counting measure, we may extend the idea to continuous distributions and compare p.d.f, $f(x; \theta)$, by changing the dominating measure to Lebesgue measure.

Definition

1. **Likelihood function**: a function of θ, $L(\theta; x) = f(x; \theta)$, where $f(x; \theta)$ is the joint density of observing $X = x$ given θ.

2. **MLE**: the estimator of θ, $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x)$.

Remarks:

1. In many cases, we observe a collection of realizations of iid random variables, $X_i \overset{\text{iid}}{\sim} f(x; \theta), i = 1, \cdots, n$. Then $L(\theta; x) = \prod_{i=1}^n f(x_i; \theta)$.

2. Sometimes, $\hat{\theta}$ may not exist in Θ but exists in $\bar{\Theta}$, the closure of the parameter space. In the textbook, the MLE is defined in $\bar{\Theta}$. Furthermore, even exists, $\hat{\theta}$ may not be unique. (Example: $U(\theta - 1/2, \theta + 1/2)$.)

3. If $\hat{\theta}$ is the MLE for θ, the $g(\hat{\theta})$ is the MLE for $g(\theta)$, where g is a measurable function from Θ to $\mathbb{R}^p, p \leq k$.

11.1 Derivation of MLE

Note logorithm is a monotone increasing function. Let $l(\theta; x) = \log L(\theta; x)$. Then

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x) = \arg\max_{\theta \in \Theta} l(\theta; x) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(x_i; \theta).$$

This is very helpful for distributions in exponential families. A common way to find MLE is by solving **likelihood equation**,

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

Examples:

1. **Binomial.** $X_i \overset{\text{iid}}{\sim} Berp, p \in (0,1), i = 1, \cdots, n$. $\hat{p} = \bar{X}$. Check the second derivative. Check the boundary $\bar{x} = 0$ and $\bar{x} = 1$.

2. **Exponential family.** $X_i \overset{\text{iid}}{\sim} f(x; \eta)$, which follows a natural exponential family of full rank with natural parameter $\eta \in \mathcal{E} \in \mathbb{R}^p$. The p.d.f. is

$$f(x_i; \eta) = h^*(x_i) \exp\{\eta^T T(x_i) - A(\eta)\},$$

and so the likelihood function is

$$L(\eta) \propto \exp\{\eta^T \sum_i T(x_i) - nA(\eta)\}.$$

Hence the likelihood equation is equivalent to

$$\frac{\partial A(\eta)}{\partial \eta} = \sum_{i=1}^{n} T(X_i)/n.$$

Recall: in exponential families with natural parameter η being an interior point of the natural parameter space, we have derived

$$E(T(X)) = \frac{\partial A(\eta)}{\partial \eta}, \quad Var(T(X)) = \frac{\partial^2 A(\eta)}{\partial \eta^2} > 0.$$

Since $\frac{\partial^2 l(\eta)}{\partial \eta^2} = -\frac{\partial^2 A(\eta)}{\partial \eta^2} < 0$, the solution of the likelihood equation,

$$\hat{\eta} = \big(\frac{\partial A(\eta)}{\partial \eta}\big)^{-1} (\sum_{i=1}^{n} T(X_i)/n),$$

is the unique MLE. Furthermore, if $\mu = E(X) = \mu(\eta)$, then the MLE for μ is $\hat{\mu} = \mu^{-1}(\hat{\eta})$.

3. Normal. More specifically, consider $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0, i = 1, \cdots, n$. Then

$$\eta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}), \quad T(X_i) = (x_i, x_i^2), \quad A(\eta) = \frac{1}{2}\left[-\frac{\eta_1^2}{2\eta_2} - \log(-\eta_2)\right].$$

So the likelihood equation is:

$$\begin{cases} \frac{\partial A(\eta)}{\partial \eta_1} = -\frac{1}{2}\frac{\eta_1}{\eta_2} = \sum_i x_i/n \\ \frac{\partial A(\eta)}{\partial \eta_2} = \frac{1}{2}[-\frac{\eta_1^2}{2\eta_2^2} + \frac{1}{\eta_2}] = \sum_i x_i^2/n. \end{cases}$$

Note that L is bounded, $\mathcal{E}$ is open, and $L \to 0$ as $\|\eta\| \to \infty$. So L is uniquely maximized at the solution of above equations. Hence the MLE for η is

$$\begin{cases} \hat{\eta_1} = \frac{\bar{x}}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} \\ \hat{\eta_2} = -\frac{1}{2}\frac{1}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} \end{cases}$$

and for parameterization μ, σ^2, the MLE is

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \bar{x})^2. \end{cases}$$

3. **Direct Maximization**: Sometimes the likelihood function is not a continuous function of parameter. In this case, the direct maximization (without using derivative) works better. Let $X_i \overset{\text{iid}}{\sim} U(0, \theta)$. Then

$$L(\theta) = \theta^{-n}\mathbf{1}(x_{(n)} \le \theta),$$
$$\therefore \hat{\theta} = x_{(n)}.$$

4. **Non-unique MLE**: $X_i \overset{\text{iid}}{\sim} U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Then

$$L(\theta) = \mathbf{1}(x_{(n)} \le \theta + \frac{1}{2})\mathbf{1}(x_{(1)} \ge \theta - \frac{1}{2}) = \mathbf{1}(x_{(n)} - \frac{1}{2}\theta \le x_{(1)} + \frac{1}{2}),$$

and any estimator satisfying $x_{(n)} - \frac{1}{2} \le \hat{\theta} \le x_{(1)} + \frac{1}{2}$ can be MLE.

5. **Non-analytical MLE**: In applications, one often cannot find the analytical form of MLE and need to evaluate it numerically.
Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} Gamma(\alpha, \gamma)$, then $\theta = (\alpha, \gamma)$

$$l(\theta) = -n\alpha \log\gamma - n\log\Gamma(\alpha) + (\alpha - 1)\sum \log x_i - \frac{1}{\gamma}\sum x_i,$$

and the likelihood equation becomes

$$\begin{cases} 0 & = -n\log\gamma - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum \log x_i \\ 0 & = -\frac{n\alpha}{\gamma} + \frac{1}{\gamma^2}\sum x_i \qquad \Leftrightarrow \gamma = \bar{x}/\alpha \end{cases}$$

Using the solution to the second one $\gamma = \bar{x}/\alpha$, we have

$$\log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \frac{1}{n}\sum \log x_i - \log\bar{x} = 0,$$

which has no explicit solution, but can be proved to have a unique solution (MLE).

11.2 MLE for Missing Data and EM Algorithm

Let $X = (X_1, \cdots, X_n)$, $Y = (Y_1, \cdots, Y_n)$ be covariates and repsponse variables. Denote $f(Y|X;\theta)$ as the joint pdf of Y given X, where θ is the parameter of the (regression) model. Then without missing data, the likelihood function of θ is

$$L(\theta; y|x) = f(y|x;\theta).$$

If missing some Y_i, what is the likelihood function for θ? Let $Y = (Y_o, Y_m)$, where Y_o is the observed data and Y_m is the missing data. Furthermore, denote $A = (A_1, \cdots, A_n)$ as the set of indicators of observing Y, where $A_i = 1$ if $Y_i \in Y_o$ and 0 otherwise. Let $f(A|Y, X;\phi)$ be the density of A given Y and X, where ϕ is the unknown parameter for missing scheme.

Missing Completely At Random (MCAR)
We say Y is missing completely at random (MCAR) if

$$f(A|Y, X;\phi) = f(A|Y_o, Y_m, X;\phi) = f(A|X;\phi).$$

Missing At Random (MAR)
We say Y is missing at random (MAR) if

$$f(A|Y, X;\phi) = f(A|Y_o, Y_m, X;\phi) = f(A|Y_o, X;\phi).$$

Then the likelihood function with MAR is

$$\begin{aligned} L(\theta, \phi; Y_o, A, X) &= f(Y_o, A|X;\theta,\phi) \\ &= \int f(Y_o, Y_m, A|X;\theta,\phi) dY_m \\ &= \int f(A|Y_o, Y_m, X;\phi) f(Y_o, Y_m|X;\theta) dY_m \\ &= f(A|Y_o, X;\phi) \int f(Y_o, Y_m|X;\theta) dY_m \quad (\text{ MAR }) \\ &= f(A|Y_o, X;\phi) f(Y_o|X;\theta) \int (Y_m|X;\theta) dY_m \quad (\text{ iid }) \\ &= f(A|Y_o, X;\phi) f(Y_o|X;\theta) \end{aligned}$$

Hence the MLE for θ is $\hat{\theta} = \arg\max_\theta f(Y_o|X;\theta)$, which is samse as the MLE ignoring missing data, Y_m.

So far, we only consider univariate case. If Y_i is a vector, one may not miss Y_i completely, but only partially. Let's start with $p = 2$, and $Y_i = (Y_{i1}, Y_{i2})$. Assume missing second components of the last few Y_i,

$$Y_o = ((Y_{11}, Y_{12})^T, \cdots, (Y_{n_2 1}, Y_{n_2 2})^T), \quad Y_m = (Y_{(n_2+1)2}, \cdots, Y_{n2}).$$

Then the likelihood function with MAR is

$$\begin{aligned} L(\theta, \phi; Y_o, A, X) &= f(A|Y_o, X; \phi) f(A|Y_o, X; \phi) f(Y_o|X; \theta) \int (Y_m|X; \theta) dY_m \\ &= f(A|Y_o, X; \phi) \prod_{i=1}^{n_2} f(Y_{i1}, Y_{i2}|X; \theta) \int \prod_{i=n_2+1}^{n} (Y_{i1}, Y_{i2}|X; \theta) dY_{i2} \\ &= f(A|Y_o, X; \phi) \prod_{i=1}^{n_2} f(Y_{i1}, Y_{i2}|X; \theta) \prod_{i=n_2+1}^{n} (Y_{i1}|X; \theta^*), \end{aligned}$$

where θ^* in the last term may only depend on part of θ or is a function of θ. To reparameterize, we replace $f(Y_{i1}, Y_{i2}|X; \theta)$ by $f(Y_{i1}|X; \theta^*) f(Y_{i2}|Y_{i1}, X; \varphi)$, then

$$L(\theta, \phi; Y_o, A, X) = f(A|Y_o, X; \phi) \prod_{i=1}^{n} (Y_{i1}|X; \theta^*) \prod_{i=1}^{n_2} f(Y_{i2}|Y_{i1}, X; \varphi).$$

where θ and (θ^*, φ) is 1-1.

Example: Bivariate Normal

For simplicity, let's assume no covariates X. Let $Y_i = (Y_{i1}, Y_{i2})^T$ be iid bivariate normal random variables with mean (μ_1, μ_2), variance (σ_1^2, σ_2^2) and correlation ρ.
Let $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Note the marginal distribution of Y_{i1} is

$$Y_{i1} \sim N(\mu_1, \sigma_1^2),$$

and the conditional distribution of Y_{i2} given Y_{i1} is

$$Y_{i2}|Y_{i1} \sim N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(Y_{i1} - \mu_1), \quad \sigma_2^2(1 - \rho^2) \right).$$

Hence $\theta^* = (\mu_1, \sigma_1^2)$ and $\varphi = (\alpha, \beta, \tau^2)$, where

$$\alpha = \mu_2 - \frac{\rho\sigma_2}{\sigma_1}\mu_1, \quad \beta = \frac{\rho\sigma_2}{\sigma_1}, \quad \tau^2 = \sigma_2^2(1 - \rho^2).$$

Then the likelihood function is $L(\theta, \phi) =$

$$f(A|Y_o; \phi) \left[\frac{1}{(2\pi\sigma_1^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (Y_{i1} - \mu_1)^2 \right\} \right] \left[\frac{1}{(2\pi\tau_1^2)^{n_2/2}} \exp\left\{ -\frac{1}{2\tau^2} \sum_{i=1}^{n_2} (Y_{i2} - \alpha - \beta Y_{i1})^2 \right\} \right]$$

So, letting $Y_{\cdot 1} = \sum_{i=1}^{n} Y_{i1}/n$, $\tilde{Y}_{\cdot 1} = \sum_{i=1}^{n_2} Y_{i1}/n_2$, and $Y_{\cdot 2} = \sum_{i=1}^{n_2} Y_{i2}/n_2$, we have

$$\hat{\mu}_1 = Y_{\cdot 1}, \quad \hat{\sigma}_1^2 = \sum_{i=1}^{n} (Y_{i1} - Y_{\cdot 1})^2/n$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n_2}(Y_{i2} - Y_{\cdot 2})(Y_{i1} - \tilde{Y}_{\cdot 1})}{\sum_{i=1}^{n_2}(Y_{i1} - \tilde{Y}_{\cdot 1})^2}, \quad \hat{\alpha} = Y_{\cdot 2} - \hat{\beta}\tilde{Y}_{\cdot 1}, \quad \hat{\tau}^2 = \sum_{i=1}^{n_2} (Y_{i2} - \hat{\alpha} - \hat{\beta} Y_{i1})^2/n_2.$$

Re-parameterize back to μ_2, σ_2^2 and ρ, we have the MLE for θ.
Compare it with the MLE without missing data. (For example, $\hat{\mu}_2 = Y_{\cdot 2} + \hat{\beta}(\mu_1 - \tilde{Y}_{\cdot 1})$.)

In general, we may consider monotone missing data, which is the most common type of missing in longitudinal data/panel data. Let $Y_i = (Y_{i1}, \cdots, Y_{im})^T$. For the t-th component of Y_i, suppose that $\{Y_{1t}, \cdots, Y_{n_t t}\}$ are observed and $\{Y_{(n_t=1)t}, \cdots, Y_{nt}\}$ are missing, and $n \geq n_1 \geq n_2 \geq \cdots \geq n_m \geq 2$. Then the likelihood function is

$$L(\theta, \phi; Y_o, A, X) = f(A|Y_o, X; \phi) \prod_{t=1}^{m} \prod_{i=1}^{n_t} f(Y_{it}|Y_{i1}, \cdots, Y_{i(t-1)}, X; \theta_t^*),$$

where $(\theta_1^*, \cdots, \theta_m^*)$ is a 1-1 function of θ. The MLE for θ then can be obtained iteratively by first finding the MLE of θ_t^* in the linear regression with Y_{it} and $(X, Y_{i1}, \cdots, Y_{i(t-1)})$, and then deriving from $(\theta_1^*, \cdots, \theta_m^*)$.

If missing is not monotone, the maximization can be very difficulty or impossible. The EM algorithm can be used to partially solve the problem (see Little and Rubin, 2002).

The EM algorithm contains two steps: Expectation step and Maximization step, and iterations between two steps. Let $L(\theta; Y_o, Y_m, X)$ be the likelihood function using the complete data, Y_o and Y_m, referred to as the complete likelihood. Let $\theta^{(k)}$ be the estimate of θ at the (k)-th iteration of the EM algorithm. The E step at the k-th iteration calculates the conditional expectation of the complete likelihood given observed data and parameter estimation in the previous iteration,

$$Q(\theta|\theta^{(k-1)}) = E_{\theta^{(k-1)}}[\log L(\theta; Y_o, Y_m, X)] = \int \log L(\theta; Y_o, Y_m, X) f(Y_m|Y_o, X; \theta^{(k-1)}) dY_m.$$

The M-step at the k-th iteration maximize above conditional expectation and find $\theta^{(k)}$ s.t.

$$Q(\theta^{(k)}|\theta^{(k-1)}) = \max_\theta Q(\theta|\theta^{(k-1)}).$$

Now we show why the EM-algorithm works. Note

$$\begin{aligned}\log L(\theta; Y_o, X) &= \int [\log L(\theta; Y_o, X)] f(Y_m|Y_o, X; \theta) dY_m \\ &= \int \left[\log \frac{f(Y_o, Y_m|X; \theta)}{f(Y_m|Y_o, X; \theta)}\right] f(Y_m|Y_o, X; \theta) dY_m \\ &= \int [\log f(Y_o, Y_m|X; \theta)] f(Y_m|Y_o, X; \theta) dY_m - \int [\log f(Y_m|Y_o, X; \theta)] f(Y_m|Y_o, X; \theta) dY_m \\ &= Q(\theta|\theta) - H(\theta|\theta),\end{aligned}$$

where $H(\theta_1|\theta) = \int [\log f(Y_m|Y_o, X; \theta_1)] f(Y_m|Y_o, X; \theta) dY_m$, and $H(\theta_1|\theta) \leq H(\theta|\theta)$ by Jensen's inequality. Hence at the k-th iteration,

$$\begin{aligned}&\log L(\theta^{(k)}; Y_o, X) - \log L(\theta^{(k-1)}; Y_o, X) \\ =&Q(\theta^{(k)}|\theta^{(k-1)}) - H(\theta^{(k)}|\theta^{(k-1)}) - Q(\theta^{(k-1)}|\theta^{(k-1)}) + H(\theta^{(k-1)}|\theta^{(k-1)}) \\ \geq&Q(\theta^{(k)}|\theta^{(k-1)}) - Q(\theta^{(k-1)}|\theta^{(k-1)}) \\ \geq&0\end{aligned}$$

with equality holds iff $Q(\theta^{(k)}|\theta^{(k-1)}) - Q(\theta^{(k-1)}|\theta^{(k-1)})$. This means the change from $\theta^{(k-1)}$ to $\theta^{(k)}$ increases likelihood. Hence EM algorithm produces a sequence of $\theta^{(k)}, k = 1, 2, \cdots$ Under certain conditions, this sequence converges and the limit is considered as the EM estimator of θ.

11.3 Related Reading

1. Sh Chapter 4.4,