STAT 200B 2019 Week04

Soyeon Ahn

1 properties of the MLE

Some properties of the MLE that we will explore are:

1. Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.
2. Consistency: $\hat{\theta}_n \xrightarrow{P} \theta_*$, where θ_* is the true value of the parameter. (The estimator converges in probability to the value being estimated)
3. Asymptotic normality: $(\hat{\theta}_n - \theta_*)/se(\hat{\theta}_n) \xrightarrow{D} N(0,1)$. (As the sample size increases, the distribution of the MLE tends to the Gaussian distribution with mean and covariance matrix equal to the inverse of the Fisher information matrix)
4. Asymptotic efficiency: The MLE has the smallest asymptotic variance among asymptotically normal estimators. (The estimator achieves the Cramér-Rao lower bound, no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE)

Conditions for the last three can be somewhat technical, so we'll start with the case that $\theta \in \Theta \subseteq \mathbb{R}$ and will focus more on intuition than on details.

1.1 Equivariance

Equivariance: Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be the MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.

Proof: Suppose that g is one-to-one. Then it possesses an inverse g^{-1}, and we can define the induced likelihood $\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau))$. But for any τ,

$$\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau)) \leq \mathcal{L}(\hat{\theta}) = \mathcal{L}^*(g(\hat{\theta}))$$

so $\hat{\tau} = g(\hat{\theta})$ maximizes $\mathcal{L}^*$.

The general case is only slightly more complicated; we define

$$\mathcal{L}^*(\tau) = \sup_{\theta : g(\theta) = \tau} \mathcal{L}(\theta)$$

1.2 Consistency

The following conditions are sufficient for consistency of the MLE:

1. $X_1, \ldots, X_n$ are iid with density $f(x;\theta)$.
2. Identifiability, i.e. if $\theta \neq \theta'$, then $f(x;\theta) \neq f(x;\theta')$.
3. The densities $f(x;\theta)$ have common support, i.e. $\{x : f(x;\theta) > 0\}$ is the same for all θ. (The support does not depend on θ)
4. The parameter space Θ contains an open set ω of which the true parameter value θ_* is an interior point. (The true parameter value θ_* lies in a compact set Θ)
5. The function $f(x;\theta)$ is differentiable with respect to θ in ω.

We expect $\hat{\theta}_n$ to converge to θ_*, true value of θ. Showing consistency requires that the convergence is uniform in θ. We also need to show that $E_{\theta_*}[\log f(X_1;\theta)]$ is maximized at $\theta = \theta_*$.

Note that $\ell_n(\theta)$ converges to $l(\theta)$ in probability.

$$\begin{aligned} \ell_n(\theta) &= \sum_{i=1}^{n} \log f(X_i;\theta) \\ &\propto \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \\ &\xrightarrow{P} E_{\theta_*}[\log f(X_1;\theta)] \text{ for any fixed } \theta \text{ by WLLN} \end{aligned}$$

where θ_* denotes the true value of θ.

Note that MLE minimizes Kullback-Leibler divergence (θ, θ_*). We will show $M(\theta_*) - M(\hat{\theta}_n)$ converges to 0 with probability.

The regularity conditions ensure uniform convergence in probability of a normalized form of the log-likelihood to its expected value.

Kullback-Leibler distance between f and g, measure the difference between two probability distributions. We assume that the model is identifiable (i.e., $\theta \neq \psi$ implies that $D(\theta, \psi) > 0$)

$$D(f,g) = \int f(x) \log\Big(\frac{f(x)}{g(x)}\Big) dx$$

Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$$

This follows since $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_*))$ and $\ell_n(\theta_*))$ is a constant. By law of large number, $M_n(\theta)$ converges to

$$\begin{aligned} E_{\theta_*}\Big(\log \frac{f(X_i;\theta)}{f(x_i;\theta_*)}\Big) &= \int \log\Big(\frac{f(x_i;\theta)}{f(X_i;\theta_*)}\Big) f(x_i;\theta_*)dx \\ &= \int -\log\Big(\frac{f(x_i;\theta_*)}{f(X_i;\theta)}\Big) f(x_i;\theta_*)dx \\ &= -D(\theta_*,\theta) \end{aligned}$$

Define $\Theta_\epsilon = \{\theta \in \Theta : |\theta - \theta*| \geq \epsilon\}$ for arbitrary $\epsilon > 0$. Note that Θ_ϵ is compact as intersection of Θ with a closed set. The function ℓ is continuous on Θ_ϵ, therefore it attains its maximum on it, so there exists θ_ϵ such that

$$\ell(\theta_\epsilon) = \sup_{\theta\in\Theta_\epsilon} = c(\epsilon) < \ell(\theta_*)$$

since θ_* is the unique maximum of ℓ on Θ. Therefore, there exists $\delta(\epsilon) > 0$ such that $c(\epsilon) + \delta(\epsilon) < \ell(\theta_*) - \delta(\epsilon)$

$$\begin{aligned} \sup_{\theta\in\Theta_\epsilon} \ell_n(\theta) &= \sup_{\theta\in\Theta_\epsilon} \ell_n(\theta) - \ell(\theta) + \ell(\theta) \\ &\leq \sup_{\theta\in\Theta_\epsilon} \ell(\theta) + \sup_{\theta\in\Theta} |\ell_n(\theta) - \ell(\theta)| \end{aligned}$$

On the sequence of events

$$A_n(\epsilon) = \{\sup_{\theta\in\Theta} |\ell_n(\theta) - \ell(\theta)| < \delta(\epsilon)\},$$

the following sequence of inequalities hold

$$\sup_{\theta\in\Theta_\epsilon} \ell_n(\theta) \leq c(\epsilon) + \delta(\epsilon) < \ell(\theta_*) - \delta(\epsilon),$$

the last inequality holding by definiton of $\delta(\epsilon)$. On $A_n(\epsilon)$, we can have θ_* such that $\ell(\theta_*) - \ell_n(\theta_*) \leq \delta(\epsilon)$. As a result, we have

$$\sup_{\theta\in\Theta_\epsilon} \ell_n(\theta) \leq \ell_n(\theta_*)$$

On $A_n(\epsilon)$, $\hat{\theta}_n$ cannot be in Θ_ϵ, as it would lead to contradiction $\ell_n(\hat{\theta}_n) < \ell_n(\theta_*)$ since $\hat{\theta}$ is the maximizer of the likelihood, by the definition. As a consequence, we have that $A_n(\epsilon) \subset \{|\hat{\theta}_n - \theta_*| < \epsilon\}$. Since $P[A_n(\epsilon)] \xrightarrow{P} 1$ as $n \to \infty$ by the uniform LLN, we have $\hat{\theta}_n$ to converge to θ_* in probability.

Example unique MLE in Normal, $X_i \sim N(\theta, 1)$ (Casella and Berger Example 7.2.5, continuation of Week03 note)

$\bar{x}$ is the MLE of $theta$ (The textbook Example 9.11). We need to verify that $\bar{x}$ is a global maximum of the likelihood function. First, note that $\hat{\theta}$ is a unique solution to $\ell_n(\theta)' \propto \sum(x_i - \theta) = 0$, i.e. the value that sets the first derivative of the (log) likelihood function to be zero. Second, verify that

$$\ell_n(\theta)'' < 0$$

Thus, $\bar{x}$ is the only extreme point in the interior and it is a maximum. To finally verify that $\bar{x}$ is a global maximum, we should check the boundaries, $\pm\infty$ as well. By taking limits, the likelihood is 0 at $\pm\infty$.

(Or, we established that $\bar{x}$ is a unique interior extremum and is a maximum, there can be no maximum at $\pm\infty$. If there were, then there would have to be an interior minimum, which contradicts uniqueness.)

Example the unique MLE in Bernoulli (continuation of Week03 note)
Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p). The probability function is $f(x; p) = p^x(1-p)^{1-x}$. We found that

$$\ell_n(p) = n\bar{x}\log(p) + n(1-\bar{x})\log(1-p)$$

If $0 < \sum x_i < n$, differentiating the log likelihood and setting the result equal to 0 give the solution, $\bar{x}$, as we did in the last time.

If $\sum x_i = 0$ or n, then

$$\ell_n(p) = \begin{cases} n\log(1-p) & \text{if } \sum x_i = 0 \\ n\log(p) & \text{if } \sum x_i = n \end{cases}$$

In either case $\ell_n(p)$ is a monotone function of p, and it is again easy to verify that $\hat{p} = \bar{x}$ in each case.
In this derivation, we have assumed that the parameter space is $0 \leq p \leq 1$. The values $p = 0$ and 1 must be in the parameter space in order for $\hat{p} = \bar{x}$ to be MLE for $\sum x_i = 0$ and n.

Example MLE in half Normal
Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, where θ is nonnegative. If there is no restriction on θ, $\hat{\theta} = \bar{X}$. If $\sum X_i$ is negative, however, this estimator will be outside the range of the parameter.
If $\sum X_i$ is negative, it is easy to check that likelihood function is decreasing in θ and is maximized at $\hat{\theta} = 0$. Thus,

$$\hat{\theta} = \begin{cases} \bar{X} & \text{if } \sum X_i \geq 0 \\ 0 & \text{if } \sum X_i < 0 \end{cases}$$

Again, when finding A MLE, the maximization takes place only over the range of parameter values.

Example MLE in Uniform Example 4.31. in Jun Shao

Example MLE in Gamma Example 4.3. in Jun Shao

1.3 Asymptotic normality and efficiency

1.3.1 Exponential family model

One class of distributions that satisfies the conditions is known as the exponential family. For $\Theta \subseteq \mathbb{R}$, these have densities that can be written as

$$f(x;\theta) \quad = \quad h(x)\exp^{\eta(\theta)T(x)-B(\theta)}$$

We can rewrite an exponential family as

$$\tilde{f}(x;\eta) \quad = \quad h(x)\exp^{\eta(\theta)T(x)-A(\eta)} \text{ [the canonical form for the family]}$$

where $\eta = \eta(\theta)$ [natural parameter] with the new parameter space $\{\eta(\theta) : \theta \in \Theta\}$ and

$$A(\eta) = \log \int h(x)\exp^{\eta T(x)}\, dx$$

Example $Binomial(n,p)$ with n known is exponential family.

$$f_\theta(x) = \exp\Big\{x\log\frac{\theta}{1-\theta} + n\log(1-\theta)\Big\}\binom{n}{x} I_{\{0,1,\ldots,x\}}(x)$$

$T(x) = x$, $\eta(\theta) = \log\frac{\theta}{1-\theta}$, $B(\theta) = -n\log(1-\theta)$, $h(x) = \binom{n}{x} I_{\{0,1,\ldots,x\}}(x)$.
If we let $\eta = \log\frac{\theta}{1-\theta}$,

$$\tilde{f}_\eta(x) = \exp\Big\{x\eta - n\log(1+\exp^\eta)\Big\}\binom{n}{x} I_{\{0,1,\ldots,x\}}(x)$$

is a natural exponential family of full rank.

Example $Exponential(\lambda)$ is exponential family.

$$\begin{aligned} f_\theta(x) &= \lambda\exp^{-\lambda x} \\ &= \exp\Big\{-\lambda x + \log(\lambda)\Big\} \end{aligned}$$

$T(x) = x$, $\eta(\theta) = -\lambda$, $B(\theta) = -\log\lambda$, $A(\eta) = -log(-\eta)$

Example $Unif(0,\theta)$ is not exponential family. Let $X^n = (X_1,\ldots,X_n)$ be iid with $Unif(0,\theta)$.

$$f_{x^n,\theta}(x) = \frac{1}{\theta^n} I(x_n \geq \theta)$$

where $x_n = max\{x_1, \ldots, x_n\}$.

$T(x) = max\{X_1, \ldots, X_n\}$ and $T(X^n) \neq \sum_i T(X_i)$.

Theorem

Let X have density in an exponential family,

$$\tilde{f}(x;\eta) \quad = \quad h(x) \exp^{\eta(\theta)T(x) - A(\eta)} \text{ [the canonical form for the family]}$$

Then,

$$E(T(X)) = A^{'}(\eta)$$
$$Var(T(X)) = A^{''}(\eta)$$

1.3.2 Score function and Fisher information

Define the score function $s(X;\theta) = \frac{\partial}{\partial\theta} \log f(X;\theta)$.

Then the Fisher information (based on n observations) is

$$\begin{aligned}
I_n(\theta) &= V_\theta \left(\frac{\partial}{\partial\theta} \ell_n(\theta) \right) \\
&= V_\theta \left(\sum_{i=1}^{n} s(X_i;\theta) \right) \text{ ()definition of the Fisher information)} \\
&= \sum_{i=1}^{n} V_\theta(s(X_i;\theta)) \quad \text{(if } X_1, \ldots, X_n \text{ are independent)} \\
&= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1, \ldots, X_n \text{ are identically distributed)} \\
&= nI_1(\theta) \quad \text{(when } n = 1\text{)} \\
&\equiv nI(\theta)
\end{aligned}$$

1.3.3 Asymptotic normality and efficiency

In addition, under a condition satisfied for exponential family models, we can calculate

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta) \right]$$

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Calculate $I_n(\lambda)$.

The "observed" Fisher information

$$I_n^{obs}(\theta) = \frac{-\partial^2}{\partial\theta^2} \sum_{i=1}^{n} \log f(X_i;\theta)$$

measures the curvature of the log-likelihood function. In particular $I_n^{obs}(\hat{\theta})$ measures the curvature at the MLE. The more peaked $\ell_n(\theta)$ is around $\hat{\theta}$, the more "information" the likelihood gives us. $I(\theta)$ measures the average value of this quantity.

Under two additional conditions (also satisfied by iid observations under exponential family models), we have

- Asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, 1/I(\theta))$
- Asymptotic efficiency: If $\tilde{\theta}_n$ is some other estimator s.t. $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0, v(\theta))$, then $v(\theta) \geq 1/I(\theta)$ for all θ.

Asymptotic normality still holds replacing $I(\theta)$ by $I(\hat{\theta})$, that is,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{1/I(\hat{\theta}_n)}} \xrightarrow{D} N(0, 1)$$

We can use this to construct approximate $1 - \alpha$ confidence intervals for θ.

Note that the efficiency relies on the Cramer-Rao lower bound. The Cramer-Rao lower bound is a theoretical minimum variance that any estimator can obtain

Example Under each of the following models, find the MLE for θ and calculate an approximate 95% confidence interval using the limiting normal distribution.

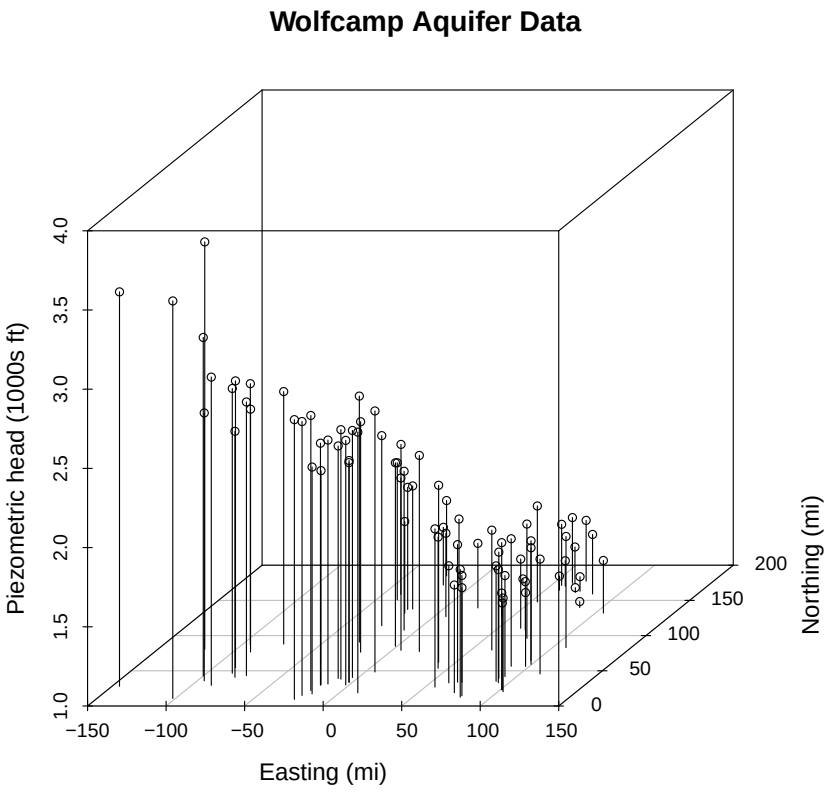
$X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta)$

$X_1, \ldots, X_n \overset{iid}{\sim} Binomial(m, \theta)$ for known m

$X_1, \ldots, X_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$ for known σ^2

1.4 Examples: multivariate normal distribution

A motivating example for more complicated likelihood problems: Given measurements of hydraulic head from an aquifer, create a predicted surface.



The Wolfcamp Aquifer lies below Deaf Smith County, Texas, once under consideration by DOE as a nuclear waste repository site.

Creating a smooth surface from the measurements would allow us to predict the path of potential contaminants.

- gradient, score vector $[\frac{\partial}{\partial\theta}\log f(X;\theta)]$
- Hessian matrix $[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X;\theta)]$

Aside: Multivariate normal distribution

The multivariate normal distribution for a vector $Z = (Z_1, Z_2, \ldots, Z_n)'$ with mean vector μ and covariance matrix Σ has pdf

$$f(z;\mu,\Sigma) = (2\pi)^{-1/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\}$$

Let μ_i denote the i^{th} element of μ, and Σ_{ij} the element of Σ in the i^{th} row and j^{th} column. Then

- $Z_i \sim N(\mu_i, \Sigma_{ii})$
- $Cov(Z_i, Z_j) = \Sigma_{ij}$

In the aquifer example, we could fit a universal kriging model, which is just a multivariate normal model where μ and Σ have special structure.

In this example, we could take

$$\mu_i = E[Z_i] = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

where (x_i, y_i) is the location of observation i.

The observations are clearly not independent, so Σ is not diagonal. One model would be to have correlation decay with distance, such as

$$\Sigma_{ij} = Cov(Z_i, Z_j) = \sigma^2 \exp\{-d_{ij}/\rho\}$$

where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

There is no closed form expression for the MLE of $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \rho)$.

When $\theta = (\theta_1, \ldots, \theta_k)$, we define the Fisher information matrix as follows.

The Hessian matrix is the matrix of second partial derivatives of the log-likelihood, with

$$H_{jj} = \frac{\partial^2}{\partial \theta_j^2} \ell_n(\theta); \quad H_{jk} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_n(\theta)$$

The Fisher information matrix is

$$I_n(\theta) = - \begin{bmatrix} E_\theta(H_{11}) & \cdots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & \cdots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & \cdots & E_\theta(H_{kk}) \end{bmatrix}$$

Let $\hat{\theta}$ be the (vector valued) MLE, and let $J_n(\theta) = I_n(\theta)^{-1}$. Then under appropriate regularity conditions and for large n,

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, J_n(\theta))$$

We can use the marginal densities ($\hat{\theta}_{n,i} \overset{D}{\approx} N(\theta_i, J_{n,ii}(\theta))$ to construct 95% confidence intervals for the individual parameters.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLEs for μ and σ are $\hat{\mu}_n = \bar{X}_n$ and $\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$. In addition...

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

$$J_n(\mu, \sigma) = I_n(\mu, \sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}$$

Using the fact that both $\hat{\mu}_n$ and $\hat{\sigma}_n$ are consistent, we can plug in to get

$$\hat{X}_n \pm 2\sqrt{\frac{\hat{\sigma}^2}{n}} \text{ and } \hat{\sigma}_n \pm 2\sqrt{\frac{\hat{\sigma}^2}{2n}}$$

as approximate 95% confidence intervals for μ and σ.

2 Multiparameter delta method

Suppose $\tau = g(\theta_1, \ldots, \theta_k)$ is a differentiable function. Let $\nabla g = (\frac{\partial}{\partial \theta_1} g(\theta) \cdots \frac{\partial}{\partial \theta_k} g(\theta))'$ be the gradient of g and suppose that ∇g evaluated at $\hat{\theta}_n$ is not zero. Then

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau}_n)} \xrightarrow{D} N(0, 1)$$

where

$$\widehat{se}(\hat{\tau}_n) = \sqrt{(\hat{\nabla} g)' J_n(\hat{\theta}_n)(\hat{\nabla} g)}$$

and $\hat{\nabla} g$ is ∇g evaluated at $\hat{\theta}_n$.

Example: Continuing the last example, let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for τ and its limiting normal distribution.

In many cases, it's not possible to find a closed-form expression for the MLE in multiparameter models. This is true even for some common distributions like the Gamma and Beta distributions.

However, numerical optimization is a highly developed field that comes to our rescue in applied problems (that is, when we have actual values for $X_1, \ldots, X_n$).

Most of these algorithms are written for minimization, so we need to

- Write a function for the negative log-likelihood
- Minimize it numerically
- Examine the behavior of the negative log-likelihood at the minimum
- Optionally, get a numerical approximation of the Hessian and compute the observed Fisher information matrix

In these problems, we often replace the Fisher information matrix with the observed Fisher information matrix. In many cases confidence intervals constructed using the observed Fisher information actually perform better than those using the Fisher information, so this is not a big issue.

See `betaexample.R` and `aquifer.R` for examples of this process in a toy dataset and for the aquifer example.