Mathematical Statistics2 Tutoring2

Seung Bong Jung

Seoul National University

October 4, 2021

Consistency of MLE

- For a set of observations $x_1, \dots, x_n$ of a random sample $X_1, \dots, X_n$ from pdf $f(\cdot;\theta)$, we have log-likelihood (function)

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

- Suppose that the likelihood is twice continuously differentiable and $\Theta = \prod_{i=1}^{k}(a_i, b_i)$. For a given x, if $\dot{l}(\hat{\theta};x) = 0$ and $\ddot{l}(\theta;x)$ is negative definite for all $\theta \in \Theta$, then $\hat{\theta}$ is the unique MLE.
- Suppose that the likelihood is twice continuously differentiable and $\Theta = \prod_{i=1}^{k}(a_i, b_i)$. For a given x, if $\lim_{\theta \to \partial\Theta} l(\theta) = -\infty$ and the second derivative $\ddot{l}(\theta;x)$ is negative definite for all $\theta \in \Theta$, then the solution of the likelihood equation exists and is the unique MLE.

Consistency of MLE

- Suppose that we observe a random sample from P_θ in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Assume that θ in Θ is identifiable and that P_θ have common support and $\Theta = \prod_{i=1}^{k}(a_i, b_i)$. Assume also that the likelihood is twice continuously differentiable, $\lim_{\theta \to \partial\Theta} l(\theta) = -\infty$, and $E_{\theta_0}(\log f(X_1;\theta))$ exists and is continuous with respect to θ. If $\dot{l}(\hat{\theta};x) = 0$ and $\ddot{l}(\theta;x)$ is negative definite for all $\theta \in \Theta$, the MLE of θ is consistent.
- **Remark)** Continuity of $E_{\theta_0}(\log f(X_1;\theta))$ for all $\theta \in \Theta$ and negative definiteness of $\ddot{l}(\theta;x)$ imply uniform convergence of $\frac{1}{n}l(\theta;x)$ to $E_{\theta_0}(\log f(X_1;\theta))$ in sense of probability. Pointwise convergence in probability is ensured by weak law of large numbers (WLLN).

Regularity conditions of MLE

Conditions for asymptotic normality of MLE

- **(R0)** The parameter θ is identifiable in Θ.
- **(R1)** The density $f(\cdot;\theta)$ have common support $\mathfrak{X}$.
- **(R2)** The parameter space is open in $\mathbb{R}^d$.
- **(R3)** The log-density pdf $\log f(x;\theta)$ is twice differentable as a function of θ for all $x \in \mathfrak{X}$.
- **(R4)** For any statistic $u(X_1, \ldots, X_n)$ with finite expectation, the integral

$$\mathsf{E}_\theta(u(X_1, \ldots, X_n)) = \int_{\mathfrak{X}^n} u(x_1, \ldots, x_n) \prod_{i=1}^n f(x_i;\theta) \prod_{i=1}^n dx_i$$

is twice differentiable under the integral sign.

Regularity conditions of MLE

Conditions for asymptotic normality of MLE

- **(R5)** The Fisher Information $I(\theta)$ exists and is invertible for all $\theta \in \Theta$.
- **(R6)** The likelihood equation $\dot{l}(\theta) = 0$ has the unique solution $\hat{\theta}$ and the solution is a consistent estimator of θ.
- **(R7)** For all $\theta \in \Theta$, there exists a function $M(\cdot)$ with $\mathsf{E}_\theta M(X_1) < \infty$ such that

$$\max_{\theta \in \Theta} \max_{1 \le h,i,j \le d} \left| \frac{\partial^3}{\partial \theta_h \partial \theta_j \partial \theta_j} \log f(X_1; \theta) \right| \le M(X_1), \mathsf{E}_\theta M(X_1) < \infty.$$

- Under regular conditions **(R0)**$-$**(R7)**,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathsf{d}} N(0, I(\theta)^{-1}).$$

Example: Logistic(θ, σ)

Let $X_1, \dots, X_n$ be a random sample from Logistic(θ, σ), where $\theta \in \mathbb{R}$ and $\sigma > 0$.

- Then, MLE $\hat{\eta}$ of $\eta = (\theta, \sigma)$ exists and is a consistent estimator of η.
- $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, V)$, where

$$V = \begin{pmatrix} 3\sigma^2 & 0 \\ 0 & \frac{9}{\pi^2+3}\sigma^2 \end{pmatrix}.$$

- One may use the fact that $\int_{\mathbb{R}} x^2 \frac{e^{2x}}{(1+e^x)^4} dx = \frac{\pi^2}{18} - \frac{1}{3}$ to find the asymptotic distribution of $\sqrt{n}(\hat{\eta} - \eta)$.

2020 Midterm Problem1

Suppose $X_1, \ldots, X_n$ are random sample from the distribution with p.d.f. f_θ, where

$$f_\theta(x) = \frac{1}{x\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(\log x - \theta)^2\right) I(0 < x < \infty).$$

(a) Find a MLE $\hat{\theta}$ of θ.
(b) Find Fisher's Information $I(\theta)$.

2020 Midterm Problem2

Suppose we observe data (X_i, Y_i), $i = 1, 2, \ldots, n$ where $n \geq 2$. Assume that $Y_i \sim \text{Exp}(\lambda_i)$ and Y_i's are mutually independent, where $\mu_i = E[Y_i] = \lambda_i^{-1} = \exp(\alpha + \beta X_i)$. Further assume that $X_i \neq X_j$ if $i \neq j$. Show that the MLE of (α, β) exists and is unique.

2020 Midterm Problem3

Consider the following model.

$$Y_i = \theta + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i = c\eta_{i-1} + \eta_i$, $i = 1, \ldots, n$, $\eta_0 = 0$, $\eta_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$, and $0 < c < 1$ is constant.

(a) Find a MLE $\hat{\theta}$ of θ.

(b) Show that $\hat{\theta}$ is consistent.

Exercises

2020 Midterm Problem5

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Poi(\lambda)$. Assume that we cannot observe X_i but can observe whether $X_i = 0$ or $X_i > 0$.
(a) Find a MLE $\hat{\lambda}$ of λ.
(b) Find the case when $\hat{\lambda}$ does not exist and the probability that $\hat{\lambda}$ does not exist assuming λ_0 is true parameter of λ.

Exercises

2018 Midterm Problem3

Let $X_1, \ldots, X_n$ be i.i.d. with p.d.f. f_{θ_1,θ_2} $(0 < \theta_1, \theta_2)$, where

$$f_{\theta_1,\theta_2}(x) = \begin{cases} (\theta_1 + \theta_2)^{-1} e^{-x/\theta_1} & \text{if } x \geq 0 \\ (\theta_1 + \theta_2)^{-1} e^{x/\theta_2} & \text{if } x < 0 \end{cases}.$$

(a) Find a MLE $\hat{\theta}$ of $\theta = (\theta_1, \theta_2)^t$.
(b) Find a 2-dimensional vector μ and 2×2 matrix Σ such that $\sqrt{n}(\hat{\theta} - \mu) \xrightarrow{\text{d}} N(0, \Sigma)$.

One-step Approximation of MLE

- **Motivation)** If initial choice $\hat{\theta}^{[0]}$ lies in a $n^{-1/2}-$ neighborhood of the true parameter θ_0 in probability, then one-step update in Newton-Raphson iteration is enough.

$$\hat{\theta}^{[1]} = \hat{\theta}^{[0]} - \ddot{l}(\hat{\theta}^{[0]})^{-1}\dot{l}(\hat{\theta}^{[0]}).$$

Theorem 1

Under (R0)-(R7) and assuming $\hat{\theta}^{[0]} = \theta_0 + O_p(n^{-1/2})$ *under* P_{θ_0},

$$\sqrt{n}(\hat{\theta}^{[1]} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

in $P_{\theta_0}-$*probability.*

Least Square Estimator

Consider the following model for

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

for $p \geq 1$.

- $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t, \mathbf{X} = (\mathbf{1}, \mathbf{x_1}, \ldots, \mathbf{x_p})$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$.
- Denote the column space of $\mathbf{X}$ by $\mathcal{C}_\mathbf{X}$(the linear space in $\mathbb{R}^n$ spanned by the columns of the matrix $\mathbf{X}$.
- Then $\mathbf{X}\hat{\boldsymbol{\beta}} = \Pi(\mathbf{Y}|\mathcal{C}_\mathbf{X})$
- Least squared estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$

Confidence Region for $\boldsymbol{\beta}$

- Recall that $\hat{\sigma}^2 = 1/(n-p-1)||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 = 1/n\text{SSE}$.
- Distribution of $\hat{\boldsymbol{\beta}}$: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$ so that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \mathbf{X}^t \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2(p+1).$$

- It can be shown that $\text{SSE}/\sigma^2 \sim \chi^2(n-p-1)$.
- Using the independence of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t(\mathbf{X}^t\mathbf{X})^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and SSE, one can see that $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\hat{\mathcal{I}} = \{\boldsymbol{\beta} : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t(\mathbf{X}^t\mathbf{X})^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1) \cdot \hat{\sigma}^2 \\ \cdot F_\alpha(p+1, n-p-1)\}.\end{aligned}$$

Least Square Estimator of Subvector β_2

Write $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)$ and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, so that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$.

- Orthogonal decomposition of the column space:

$$\mathcal{C}_{\mathbf{X}} = \mathcal{C}_{\mathbf{X}_1, \mathbf{X}_2 - \Pi_1 \mathbf{X}_2} = \mathcal{C}_{\mathbf{X}_1} \oplus \mathcal{C}_{\mathbf{X}_2 - \Pi_1 \mathbf{X}_2}$$

  where $\Pi_1 = \mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t$.

- Writing $\mathbf{X}_{2,\perp} = \mathbf{X}_2 - \Pi_1\mathbf{X}_2$, we get

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}_1[\hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{X}_2\hat{\boldsymbol{\beta}}_2] + \mathbf{X}_{2,\perp}\hat{\boldsymbol{\beta}}_2$$
$$\Pi(\mathbf{Y}|\mathcal{C}_{\mathbf{X}}) = \Pi_1\mathbf{Y} + \mathbf{X}_{2,\perp}(\mathbf{X}_{2,\perp}^t\mathbf{X}_{2,\perp})^{-1}\mathbf{X}_{2,\perp}^t\mathbf{Y}$$

- This gives

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t(\mathbf{Y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2), \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_{2,\perp}^t\mathbf{X}_{2,\perp})^{-1}\mathbf{X}_{2,\perp}^t\mathbf{Y}.$$

Confidence Region for $\boldsymbol{\beta}_2$

- Distribution of $\hat{\boldsymbol{\beta}}_2$: $(\mathbf{X}_{2,\perp}^t \mathbf{X}_{2,\perp})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \sim N_q(\mathbf{0}_q, \sigma^2 \mathbf{I}_q)$, so that $(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^t(\mathbf{X}_{2,\perp}^t \mathbf{X}_{2,\perp})(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \sim \sigma^2 \chi^2(q)$
- Distribution of SSE: SSE $\sim \sigma^2 \chi^2(n-p-1)$
- SSE and $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ are independent.
- Similarly, $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}$ is given by

$$\hat{\mathcal{I}} = \{\boldsymbol{\beta}_2 : (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^t \mathbf{X}_{2,\perp}^t \mathbf{X}_{2,\perp}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \leq q \cdot \hat{\sigma}^2 \cdot F_\alpha(q, n-p-1)\}.$$