Maximum Likelihood Estimation

Ercan Karadas

New York University
Department of Economics

Spring, 2018

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

Definitions

- Problem: we have a random variable X of interest, but its pdf $f(x;\theta)$ is unknown because it depends on an *unknown* parameter $\theta \in \Omega$
- Goal: estimate and make inference for θ
- For now we assume θ is a scalar and r.v. is continuous but the results extend to the cases where θ is a vector and the r.v. is discrete
- For information we have a random sample (iid) on X:

$$\mathbf{X} = \{X_1, \ldots, X_n\}$$

- The information in the sample and the parameter θ are involved in the joint distribution of the random sample

$$\prod_{i=1}^{n} f(x_i;\theta)$$

- We want to view this as a function of θ, so we write it as

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

  This is called the **likelihood function** of the random sample
- It is usually more convenient to work with the log of this function

$$\ell(\theta) = \log L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

  which is called the **log-likelihood function** of the random sample

Example: Exponential Distribution

- Suppose the common pdf of the random sample $\{X_1, \ldots, X_n\}$ is exponential with parameter θ, i.e. $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$
- The log-likelihood function of the random sample

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

- Then the critical question is: find the parameter value $\hat{\theta}$ that maximize the likelihood function, given the observations
- The first order condition of the log-likelihood function wrt θ

$$\frac{\partial \ell(\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i$$

- Setting this to 0 and solving for θ we obtain $\hat{\theta} = \bar{x}$
- Check that the second derivative of the log-likelihood function wrt θ evaluated at the sample is strictly negative, so that $\hat{\theta} = \bar{x}$ is indeed a maximum
- Therefore, the maximum likelihood estimator of θ is

$$\hat{\theta} = \bar{x}$$

- Note that in this example, $\hat{\theta}$ is an unbiased estimator of θ since $E(X) = \theta$ we have $E(\bar{x}) = \theta$

Example: Bernoulli Distribution

- Let X be Bernoulli with the parameter $\theta \in [0, 1]$ denoting the probability of success. The pmf of X is

$$p(x; \theta) = \begin{cases} \theta^x (1-\theta)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

- Suppose we have a random sample $\{x_1, \ldots, x_n\}$ on X. Then the log-likelihood function of the random sample

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta) = \log \theta \sum_{i=1}^{n} x_i + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-\theta)$$

- The first order condition of the log-likelihood function wrt θ

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta}$$

(Check the second order condition!)

- Setting this to 0 and solving for θ we obtain the maximum likelihood estimator of θ as

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \equiv \hat{p}$$

that is, the mle $\hat{p}$ is the proportion of successes in the n trials

- $\hat{\theta}$ is an unbiased estimator of θ since $E(X) = \theta$ we have $E(\hat{\theta}) = \theta$

Example: Bernoulli Distribution with a Prior Information on θ

- Consider the previous Bernoulli example, where we found that the mle is $\hat{p}$, the proportion of sample successes.
- Now suppose that we know in advance that, instead of $\theta \in [0, 1]$, θ is restricted to $\theta \in [0, 1/2]$.
- Let's find mle of θ in this case.
- Just note that if the observations in the sample were such that $\hat{p} > 1/2$, then $\hat{p}$ would not be a satisfactory estimate.
- Therefore, in this case we obtain the mle of θ as

$$\hat{\theta} = \min\{\hat{p}, \frac{1}{2}\}$$

Example: Normal Distribution

- Let X have a $N(\mu, \sigma^2)$ distribution with the pdf given by

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty$$

  and suppose we have a random sample $\{x_1, \ldots, x_n\}$ on X
- In this example, θ is a vector $\boldsymbol{\theta} = (\mu, \sigma)$
- The log-likelihood function of the random sample

$$\ell(\theta) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

- The first order conditions of the log-likelihood function wrt θ

$$\frac{\partial \ell(\theta)}{\partial \mu} = -\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right)$$

$$\frac{\partial \ell(\theta)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Setting these to 0 and solving for (μ, σ) simultaneously we obtain the mle as

$$\hat{\mu} = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- $\hat{\mu}$ is an unbiased estimator of μ since $E(X) = \mu$,
- But $\hat{\sigma}$ is a biased estimator of σ with the bias

$$E(\hat{\sigma}^2 - \sigma^2) = -\frac{\sigma^2}{n}$$

which converges to 0 as $n \to \infty$

Example: Uniform Distribution

- Let X have $U[0, \theta]$ with the pdf given by
$$f(x; \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta$$
and suppose we have a random sample $\{x_1, \ldots, x_n\}$ on X
- The likelihood function of the random sample
$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \max_{x_i} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$
- This function is strictly decreasing in θ, so it is maximized for the smallest value that θ can assume, i.e.
$$\hat{\theta} = \max\{x_1, \ldots, x_n\}$$
- Note that in this example, differentiation doesn't work

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

- After some motivating examples now we look at the theory of MLE
- Let θ_0 denote the *true value* of θ
- First we will show that the maximum of $L(\theta)$ asymptotically separates the true model at θ_0 from models at $\theta \neq \theta_0$
- To prove this theorem, we assume certain assumptions, usually called regularity conditions:
  - **A1.**The pdfs are distinct for different θs, i.e.
    $$\theta \neq \theta' \implies f(x;\theta) \neq f(x;\theta')$$
    (the parameter identifies the pdf)
  - **A2.** The pdfs have common support for all θ
    (the support of X_i does not depend on θ)
  - **A3.** The point θ_0 is an interior point in Ω

Theorem 1

Let θ_0 be the true parameter. Under the assumptions **A1** *and* **A2**

$$\lim_{n\to\infty} P_{\theta_0}\left[L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})\right] = 1, \quad \text{for all } \theta \neq \theta_0$$

Proof.

- *By taking logs, the inequality $L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})$ is equivalent to*

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(x_i;\theta)}{f(x_i;\theta_0)}\right] < 0$$

- *Since the summands are iid with finite expectation*

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(x_i;\theta)}{f(x_i;\theta_0)}\right] \xrightarrow{p} E_{\theta_0}\left\{\log\left[\frac{f(x_1;\theta)}{f(x_1;\theta_0)}\right]\right\} < \log E_{\theta_0}\left\{\frac{f(x_1;\theta)}{f(x_1;\theta_0)}\right\}$$

  where for the last inequality we applied Jensen's inequality to $\log x$

- *But*

$$E_{\theta_0}\left\{\frac{f(x_1;\theta)}{f(x_1;\theta_0)}\right\} = \int \frac{f(x_1;\theta)}{f(x_1;\theta_0)} f(x_1;\theta_0)dx_1 = 1$$

- *Because* $\log 1 = 0$*, the theorem follows.*

- This theorem says that asymptotically the likelihood function is maximized at the true value θ_0
- So in considering estimates of θ_0, it seems natural to consider the value of θ which maximizes the likelihood function
- We say that $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is a **maximum likelihood estimator(mle)** of θ if

$$\hat{\theta} \in \arg\max\, L(\theta; \mathbf{X})$$

- To determine the mle, as we did in the examples, we often take the log of the likelihood and determine its critical value; that is, letting $\ell(\theta) = \log L(\theta)$, the mle solves the equation

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

Example: Laplace (or double exponential) Distribution

- Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{1}{2} e^{-|x-\theta|}, \qquad -\infty < x < \infty, -\infty < \theta < \infty$$

- The log-likelihood function of the random sample

$$\ell(\theta) = -n \log 2 - \sum_{i=1}^{n} |x_i - \theta|$$

- The first order condition of the log-likelihood function wrt θ

$$\ell'(\theta) = \sum_{i=1}^{n} \text{sgn}(x_i - \theta)$$

  where $\text{sgn}(t) = 1, 0, -1$ depending on whether $t > 0, t = 0$, or $t < 0$, respectively

- Note that setting this to 0 means that half of the terms in the sum should be nonpositive and half nonnegative
- Therefore, the solution for θ is

$$\hat{\theta} = \text{median}\{x_1, \ldots, x_n\} \equiv Q_2$$

  that is, $\hat{\theta} = Q_2$ is the mle of θ for the Laplace distribution.

Theorem 2

Let $X_1, \ldots, X_n$ be iid with the pdf $f(x;\theta)$ and suppose $\hat{\theta}$ is the mle of θ. For a given function g, let $\nu = g(\theta)$ be a parameter of interest. Then the mle of ν is $g(\hat{\theta})$.

- Actually we have already used this theorem when we found the mle of σ^2 in the example with normal pdf
- Consider the Bernoulli example, where we showed that the mle of θ is the proportion of successes: $\hat{\theta} = \hat{p}$
- Now if we need an estimate for $\sqrt{\theta(1-\theta)}$, by using this theorem we can say that the mle of this quantity is $\sqrt{\hat{p}(1-\hat{p})}$.

Theorem 3

Assume that $X_1, \ldots, X_n$ be iid with the pdf $f(x;\theta)$ that is differentiable with respect to θ; let θ_0 be the true parameter and the regularity conditions **A1-A3** *are satisfied. Then the likelihood equation,*

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

or equivalently

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

has a solution $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{p} \theta_0$.

- This theorem says that maximum likelihood estimators, under regularity conditions, are consistent estimators
- Note that this theorem is vague in that it discusses solutions of the equation
- If, however, we know that the mle is the unique solution of the equation $\ell'(\theta) = 0$ then it is consistent.

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

- In this section, we establish a remarkable inequality called the **Rao-Cramer** lower bound, which gives a lower bound on the variance of any unbiased estimate
- We then show that, under regularity conditions, the variances of the maximum likelihood estimates achieve this lower bound asymptotically.
- For this we need additional regularity conditions
  - **A4.**The pdf $f(x;\theta)$ is twice differentiable as a function of θ.
  - **A5.** The integral $\int f(x;\theta)dx$ can be differentiated twice under the integral sign as a function of θ.
- These two regularity conditions, together with **A2-A3**, mean that the parameter θ does not appear in the endpoints of the interval in which $f(x;\theta) > 0$ and that we can interchange integration and differentiation with respect to θ
- Before showing the Rao-Cramer lower bound, let's make some preparations and define a random variable

$$S \equiv \frac{\partial \log f(x;\theta)}{\partial \theta}$$

which is called **score function**

Theorem 4

For the score function S defined above we have

$$E(S) = 0$$

$$V(S) = -E\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right]$$

Proof.

- *To see $E(S) = 0$, just differentiate $1 = \int f(x;\theta)dx$ wrt θ and observe that it can be manipulated to obtain*

$$0 = \int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx = E\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]$$

- *To obtain the second part of the theorem, differentiate this again*

$$0 = \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx + \int \left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2 f(x;\theta)dx$$

- *Now, since $V(S) = E(S^2) - [E(S)]^2$, using the results above proves the second part.*

□

Fisher Information (Matrix)

- The variance of S, $V(S)$, is called **Fisher Information (Matrix)** and denoted by $I(\theta)$, that is

$$I(\theta) = V\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right] = E\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right]$$

- And we have shown that Fisher Information (Matrix) $I(\theta)$ is equal to

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

- Note that Fisher Information (Matrix) $I(\theta)$ defined for a sample of size 1. What about a sample of size n?
- The likelihood $L(\theta)$ is the pdf of the random sample, and the random sample variable whose variance is the information in the sample is given by

$$\frac{\partial \log L(\theta, \mathbf{X})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta}$$

- The summands are iid with common variance $I(\theta)$. Hence the information in the sample is

$$V\left(\frac{\partial \log L(\theta, \mathbf{X})}{\partial \theta}\right) = n \times I(\theta)$$

Thus the information in a random sample of size n is n times the information in a sample of size 1.

Example: Fisher Information for a Bernoulli R.V.

- Let X be a Bernoulli r.v. thus

$$\log f(x;\theta) = x \log \theta + (1-x)\log(1-\theta)$$
$$\frac{\partial \log f(x;\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$
$$\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

- Now we can compute $I(\theta)$

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right] \\ &= -E\left[-\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}\right] \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

- Note that $I(\theta)$ is larger for θ values close to zero or one, the smallest for $1/2$

Theorem 5 (Rao-Cramer Lower Bound)

Let $X_1, \ldots, X_n$ be iid with common pdf $f(x;\theta)$ and assume the regularity conditions **A1**-**A5** *are satisfied. Let $Y = u(X_1, \ldots, X_n)$ be a statistic with mean $E(Y) = E[u(X_1, \ldots, X_n)] = k(\theta)$. Then*

$$V(Y) \geq \frac{[k'(\theta)]^2}{n \times I(\theta)} \qquad \textit{(R-C Ineq)}$$

As a special case of this theorem if Y is an unbiased estimator of θ, so that $k(\theta) = \theta$, then the Rao-Cramer inequality becomes

$$V(Y) \geq \frac{1}{n \times I(\theta)}$$

Efficient Estimator: The statistic Y is called an efficient estimator of θ if and only if the variance of Y attains the Rao-Cramer lower bound.

Proof

- Write the mean of Y as
$$k(\theta) = \int \dots \int u(x_1, \dots, x_n) f(x_1; \theta) \cdots f(x_1; \theta) dx_1 \cdots dx_n$$
- Differentiating with respect to θ, we obtain
$$k'(\theta) = \int \dots \int u(x_1, \dots, x_n) \left[\sum_{i=1}^{n} \frac{1}{f(x_i; \theta)} \frac{\partial f(x_i; \theta)}{\partial \theta} \right] f(x_1; \theta) \cdots f(x_1; \theta) dx_1 \cdots dx_n$$
- Note that the terms inside the summation are equal to $\frac{\partial \log f(x_i;\theta)}{\partial \theta}$. Let's define the expression in the parantesis as Z. We know that $E(Z) = 0$ and $V(Z) = nI(\theta)$
- Also this equation can be expressed in terms of expectations as $k'(\theta) = E(YZ)$. Hence we have
$$k'(\theta) = E(YZ) = E(Y)E(Z) + \rho\sigma_Y \sqrt{n \times I(\theta)}$$
where ρ is the correlation coefficient between Y and Z. Using $E(Z) = 0$ this simplifies to
$$\rho = \frac{k'(\theta)}{\sigma_Y \sqrt{n \times I(\theta)}}$$
- Because $\rho^2 \leq 1$ we obtain the result.

Examples that attains Rao-Cramer Lower Bound

- **Example 1.** Consider again the Bernoulli example, for which we obtained, $\hat{\theta} = \hat{p}$.
  - In the previous example, we showed that $I(\theta) = \frac{1}{\theta(1-\theta)}$, and therefore we can compute R-C lower bound as $\frac{1}{n \times I(\theta)} = \frac{\theta(1-\theta)}{n}$
  - On the other hand, direct computation shows that the variance of $\hat{p}$ is $\theta(1-\theta)/n$
  - By comparing these two, we can say that the variance of the mle attains the Rao-Cramer lower bound, and therefore, it is an efficient estimator of θ.
- **Example 2.** Let $X_1, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter $\theta > 0$. It's easy to show that $\hat{\theta} = \bar{x}$ is an mle of θ. Let's see that it's also an efficient estimator:
  - Using the pmf of a Poisson distribution we have

$$\log f(x;\theta) = x \log \theta - \theta - \log x!$$

$$\frac{\partial \log f(x;\theta)}{\partial \theta} = \frac{x}{\theta} - 1 = \frac{x-\theta}{\theta}$$

$$I(\theta) = E\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right] = \frac{E(x-\theta)^2}{\theta^2} = \frac{\sigma^2}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

  - Therefore, we compute R-C lower bound as $\frac{1}{n \times I(\theta)} = \frac{\theta}{n}$
  - But $\frac{\theta}{n}$ is the variance of $\bar{x}$. Hence $\bar{x}$ is an efficient estimator of θ

Theorem 6 (Asymptotic Distribution of MLE)

Let $X_1, \ldots, X_n$ be iid with common pdf $f(x; \theta_0)$ and assume the regularity conditions **A1**-**A5** *are satisfied. Suppose further that the Fisher information satisfies $0 < I(\theta) < \infty$. Then any consistent sequence of solutions of the mle equations satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

or equivalently

$$\hat{\theta}_n \xrightarrow{d} N\left(\theta_0, \frac{1}{n \times I(\theta_0)}\right)$$

Proof

- Expanding the function $\ell'(\theta)$ into a Taylor series of order 2 about θ_0 and evaluating it at $\hat{\theta}_n$, we get

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\ell'''(\theta_n^*)$$

  where θ_n^* is between θ_0 and $\hat{\theta}_n$. Since $\ell'(\hat{\theta}_n) = 0$, rearranging gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0) - (2n)^{-1}(\hat{\theta}_n - \theta_0)\ell'''(\theta_n^*)}$$

- Bt CLT

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \frac{1}{\sqrt{n}}\sum \frac{\partial \log f(x_i;\theta_0)}{\partial\theta} \xrightarrow{d} N(0, I(\theta_0))$$

- Also by LLN

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum \frac{\partial^2 \log f(x_i;\theta)}{\partial\theta^2} \xrightarrow{p} I(\theta_0)$$

- Combining these two completes the proof, provided that the second term in the denominator goes to zero in probability.
- That is we also need to show that $(2n)^{-1}(\hat{\theta}_n - \theta_0)\ell'''(\theta_n^*) \xrightarrow{p} 0$. But since $\hat{\theta}_n \xrightarrow{p} \theta_0$, the result follows as long as $n^{-1}\ell'''(\theta_n^*)$ is bounded (in probability). Showing this last piece is little tedious and not constructive so I skip that part.

Corollary 7 (Some Consequencies and Applications of Theorem 6)

Suppose that the assumptions of Theorem 6 hold.

1. *Suppose $g(x)$ is a continuous function of x which is differentiable at θ_0 such that $g'(\theta_0) \neq 0$. Then,*

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{[g'(\theta_0)]^2}{I(\theta_0)}\right)$$

2. *The following asymptotic representation of $\hat{\theta}$ holds*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{I(\theta_0)} \frac{1}{\sqrt{n}} \sum \frac{\partial \log f(x_i; \theta_0)}{\partial \theta} + R_n$$

   where $R_n \xrightarrow{p} 0$.

3. *For a given $\alpha \in (0, 1)$, the following interval is an approximate $(1 - \alpha)100\%$ confidence interval for θ*

$$\left(\hat{\theta}_n - z_{\alpha/2} \frac{1}{\sqrt{n \times I(\hat{\theta}_n)}}, \quad \hat{\theta}_n + z_{\alpha/2} \frac{1}{\sqrt{n \times I(\hat{\theta}_n)}}\right)$$

   Therefore, Theorem 6 is also a practical result for it gives us a way of doing inference.

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

Computation of MLE by Newton's Method

- So far in all of our examples we were able to find closed form solutions to the equation

$$\ell'(\theta) = 0 \qquad (\star)$$

  and therefore we were able to write explicit formulas for the mle $\hat{\theta}$.
- However, sometimes, we can verify the existence of the mle, but the solution of the equation $(\star)$ cannot be obtained in closed form. In such situations, numerical methods are used.
- One such numerical method is Newton's method that we discussed at the programming part in lecture notes for R.
- In Newton's method we start with an initial guess and then the method takes us to another point, and this becomes the initial point in the second round, and so on.

Computation of MLE by Newton's Method

- Suppose $\hat{\theta}^{(0)}$ is an initial guess at the solution.
- Then we know that in Newton's method the next point $\hat{\theta}^{(1)}$ is given by

$$\hat{\theta}^{(1)} = \hat{\theta}^{(0)} - \frac{\ell'(\hat{\theta}^{(0)})}{\ell''(\hat{\theta}^{(0)})}$$

- We then substitute $\hat{\theta}^{(1)}$ for $\hat{\theta}^{(0)}$ and repeat the process.
- This method is easy to apply and fast. But there is no guarantee that it will converge to the solution; for some initial guesses it won't so giving a good initial guess is important.
- The estimate $\hat{\theta}^{(1)}$ is called **one-step estimator**.
- There is a remarkable fact: when the initial guess $\hat{\theta}^{(0)}$ is a consistent estimate of θ, then this estimator has the same asymptotic distribution as the mle $\hat{\theta}$.
- In order to see this just substitute the expressions for $\ell'(\hat{\theta}^{(0)})$ and $\ell''(\hat{\theta}^{(0)})$ from the the proof of Theorem 6 into the expression for $\hat{\theta}^{(1)}$.

Example: Numerical Computation of MLE for Logistic Distribution

- Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{e^{-(x-\theta)}}{\left(1+e^{-(x-\theta)}\right)^2}, \quad -\infty < x, \theta < \infty$$

- The corresponding log-likelihood function is

$$\ell(\theta) = n\theta - n\bar{x} - 2\sum \log\left(1+e^{-(x_i-\theta)}\right)$$

- Then the first partial derivative:

$$\ell'(\theta) = n - 2\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1+e^{-(x_i-\theta)}},$$

- Setting this equation to 0 and rearranging terms results in the equation

$$\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1+e^{-(x_i-\theta)}} = \frac{n}{2}$$

- We can't solve for the mle $\hat{\theta}$ explicitly in this equation
- However, we will show next that there is a $\hat{\theta}$ that solves the equation.

- Let's denote the left-hand-side of the last equation by $g(\theta)$. Then observe the following

$$\lim_{\theta \to -\infty} g(\theta) = 0$$

$$\lim_{\theta \to \infty} g(\theta) = n$$

$$g'(\theta) = \sum_{i=1}^{n} \frac{e^{-(x_i - \theta)}}{\left(1 + e^{-(x_i - \theta)}\right)^2} > 0$$

- Thus $g(\theta)$ is strictly increasing ranging from 0 to n, and therefore has to take the value $n/2$ for some θ and this value must be unique.
- So, we find that the equation has a unique solution.
- In order to show that the unique solution is a maximum we still need to show that $\ell''(\theta) < 0$. But

$$\ell''(\theta) = -2 \sum_{i=1}^{n} \frac{e^{-(x_i - \theta)}}{\left(1 + e^{-(x_i - \theta)}\right)^2} < 0$$

  Thus the mle exists and is unique.
- Now, having shown that the mle exists and is unique, we can use Newton's method to obtain the solution.
- Note that we have already computed all the derivatives, $\ell'(\theta)$ and $\ell''(\theta)$, needed in Newton's method.
- All we need is an initial guess. Let it be the sample mean $\bar{x}$.

Sample Code for Finding MLE by Newton's Method

```
#NewtonMleLogistic obtains the maximum likelihood estimate for
#the logistic density function by using Newton's method.

NewtonMleLogistic = function(x,numstp=100,eps=.0001){
n = length(x)
theta0=mean(x)
numfin = numstp
small = 1.0*10^(-8)
ic = 0
istop = 0
while(istop == 0){
  ic = ic + 1
  expx = exp(-(x - theta0))
  lprime = n-2*sum(expx/(1+expx))
  ldprime = -2*sum(expx/(1+expx)^2)
  theta1 = theta0 - (lprime/ldprime)
  check = abs(theta0-theta1)/abs(theta0 + small)
  if(check < eps){istop=1}
  theta0 = theta1
}
list(theta1=theta1,check=check,realnumstps=ic)
}
```

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

- In this section, we discuss the case where $\boldsymbol{\theta}$ is a vector of p parameters.
- We have already seen one example of this when we computed mle for the normal distribution, where $\boldsymbol{\theta} = (\mu, \sigma)$.
- In regard to consistency and efficiency of mle of $\boldsymbol{\theta}$, there are analogs to the theorems in the previous sections in which θ is a scalar, and we present their results but, for the most part, without proofs.
- To fix the notation: let $X_1, \ldots, X_n$ be iid with common pdf $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Omega \subset \mathbf{R}^p$ and as before

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \qquad \text{Likelihood function}$$

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}) \qquad \text{Log-likelihood function}$$

- We need some additional regularity conditions:
  - **A6.** There exists an open subset $\Omega_0 \subset \Omega$ such that $\boldsymbol{\theta}_0 \subset \Omega_0$ and all third partial derivatives of $f(x; \boldsymbol{\theta})$ exist for all $\boldsymbol{\theta}_0 \subset \Omega_0$, where $\boldsymbol{\theta}_0$ is the true parameter vector.
  - **A7.** The following equations are true (essentially, we can interchange expectation and differentiation):

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \log f(x; \boldsymbol{\theta})}{\partial \theta_j}\right] = 0, \qquad j = 1, \ldots, p$$

$$I_{jk}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f(x; \theta)}{\partial \theta_j \partial \theta_k}\right], \qquad j, k = 1, \ldots, p$$

  - **A8.** For all $\boldsymbol{\theta} \in \Omega_0$, $I(\boldsymbol{\theta})$ is positive definite

Theorem 1 and Theorem 2 Remain Valid

- Note that the proof of Theorem 1 does not depend on whether the parameter is a scalar or a vector. Therefore, with probability going to 1, $L(\boldsymbol{\theta})$ is maximized at the true value of $\boldsymbol{\theta}$.
- Hence, as an estimate of $\boldsymbol{\theta}$ we consider the value which maximizes $L(\boldsymbol{\theta})$ or equivalently solves the vector equation

$$\frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta}) = 0, \quad j = 1, \ldots, p$$

- If it exists, this value is called the **maximum likelihood estimator** (mle) and we denote it by $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} \in \arg\max\, L(\boldsymbol{\theta}; \mathbf{X})$$

- Often we are interested in a function of $\boldsymbol{\theta}$, say, the parameter $\nu = g(\boldsymbol{\theta})$. Theorem 2 remains true for $\boldsymbol{\theta}$ as a vector, so $\hat{\nu} = g(\hat{\boldsymbol{\theta}})$ is the mle of ν.

Example: General Laplace Distribution

- Let $X_1, \ldots, X_n$ be iid with density

$$f(x; \boldsymbol{\theta}) = \frac{1}{2b} e^{-\frac{|x-a|}{b}}, \qquad -\infty < x < \infty,$$

  where $\boldsymbol{\theta} = (a, b)$ and the parameter space is $\Omega = \{(a, b) | -\infty < a < \infty, b > 0\}$.
- The log-likelihood function of the random sample

$$\ell(\boldsymbol{\theta}) = -n \log 2 - n \log b - \sum_{i=1}^{n} \frac{|x_i - a|}{b}$$

- The first order condition of the log-likelihood function wrt a and b

$$\frac{\partial \ell(a, b)}{\partial a} = \frac{1}{b} \sum_{i=1}^{n} \operatorname{sgn}(x_i - a)$$

$$\frac{\partial \ell(a, b)}{\partial b} = -\frac{n}{b} + \frac{1}{b^2} \sum_{i=1}^{n} |x_i - a|$$

- Setting these to 0 gives mle

$$\hat{a} = \text{median}\{x_1, \ldots, x_n\} \equiv Q_2$$

$$\hat{b} = \frac{1}{n} \sum_{i=1}^{n} |x_i - Q_2|$$

Fisher Information Matrix

- Recall that the Fisher information in the scalar case was the variance of the random variable

$$S \equiv \frac{\partial \log f(x;\theta)}{\partial \theta}$$

- The analog in the multiparameter case is the variance-covariance matrix of the gradient of $\log f(x;\boldsymbol{\theta})$, that is, the variance- covariance matrix of the random vector given by

$$\nabla \log f(x,\boldsymbol{\theta}) = \left(\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_p} \right)'$$

- For convenience again define a random variable

$$S_j \equiv \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_j}$$

- With this notation, **Fisher information matrix** defined as the $p \times p$ matrix

$$I(\boldsymbol{\theta}) = [I_{jk}],$$

where (j,k) entry of $I(\boldsymbol{\theta})$ is given by

$$I_{j,k} = \text{cov}(S_j, S_k)$$

Theorem 8

For the random variable S_j defined above we have

$$E(S_j) = 0$$

$$I_{jk} = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(x;\boldsymbol{\theta})\right]$$

Proof.

- *For the first part, the proof is the same as in the scalar case. For the second part when you take the partial derivative second time take the partial derivative with respect to θ_k and do similar rearrangements to derive the result (fill the details!).*

□

- Information for a random sample follows in the same way as the scalar case. The pdf of the sample is the likelihood function $L(\boldsymbol{\theta}; \mathbf{X})$.
- Because $\log L$ is a sum, this results in the random vector

$$\nabla \log L(\boldsymbol{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \nabla \log f(x_i; \boldsymbol{\theta})$$

- The summands are iid with common variance $I(\boldsymbol{\theta})$. Hence the information in the sample is

$$\operatorname{cov}\left(\nabla \log L(\boldsymbol{\theta}, \mathbf{X})\right) = n \times I(\boldsymbol{\theta})$$

  As in the scalar case, the information in a random sample of size n is n times the information in a sample of size 1.
- The diagonal entries of $I(\boldsymbol{\theta})$ are

$$I_{jj} = V\left[\frac{\partial}{\partial \theta_j} \log f(x; \boldsymbol{\theta})\right] = -E\left[\frac{\partial^2}{\partial \theta_j^2} \log f(x; \boldsymbol{\theta})\right]$$

  This is similar to the case when θ is a scalar, except now $I_{jj}(\boldsymbol{\theta})$ is a function of the vector $\boldsymbol{\theta}$.

- Recall in the scalar case that $(n \times I(\theta))^{-1}$ was the Rao-Cramer lower bound for an unbiased estimate of θ. There is an analog to this in the multiparameter case.
- In particular, if $Y_j = u_j(X_1, \ldots, X_n)$ is an unbiased estimate of θ_j, then it can be shown that
$$V(Y_j) \geq \frac{1}{n \times I_{jj}(\boldsymbol{\theta})}$$
- As in the scalar case, we shall call an unbiased estimate efficient if its variance attains this lower bound.

Theorem 9 (Asymptotic Distribution of MLE)

Let $X_1, \ldots, X_n$ be iid with common pdf $f(x; \boldsymbol{\theta}_0)$ and assume the regularity conditions are satisfied. Then

- *The likelihood equation,*

$$\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0,$$

  has a solution $\hat{\boldsymbol{\theta}}_n$ such that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$.

- *Any consistent sequence of solutions of the mle equations satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, I^{-1}(\boldsymbol{\theta}_0)\right)$$

  or equivalently

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{d} N\left(\boldsymbol{\theta}_0, \frac{1}{n} I^{-1}(\boldsymbol{\theta}_0)\right)$$

- *In particular, let $\hat{\boldsymbol{\theta}}_n$ be a sequence of consistent solutions of the likelihood equation. Then $\hat{\boldsymbol{\theta}}_n$ are asymptotically efficient estimates; that is, for $j = 1, \ldots, p$,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{j,n} - \boldsymbol{\theta}_{j,0}) \xrightarrow{d} N\left(0, I_{jj}^{-1}(\boldsymbol{\theta}_0)\right)$$

- Let g be a transformation $g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta}))'$ such that $1 \leq k \leq p$ and that the $k \times p$ matrix of partial derivatives

$$\mathbf{B} = \left[\frac{\partial g_i}{\partial \theta_j}\right], \quad i = 1, \ldots, k, \quad j = 1, \ldots, p,$$

has continuous elements and does not vanish in a neighborhood of $\boldsymbol{\theta}$. Then the mle of $\nu = g(\theta)$ is given by $\hat{\nu} = g(\hat{\theta})$ and

$$\sqrt{n}(\hat{\nu} - \nu) \xrightarrow{d} N\left(0, \mathbf{B}I^{-1}(\boldsymbol{\theta})\mathbf{B}'\right)$$

- Hence the information matrix for $\sqrt{n}(\hat{\nu} - \nu)$ is

$$I(\nu) = \left[\mathbf{B}I^{-1}(\boldsymbol{\theta})\mathbf{B}'\right]^{-1}$$

This result is very useful when we want to find the information matrix of a function of parameters. See the next example.

Example: Information Matrix for the Normal Distribution

- Recall the example with Normal distribution. But here we are going to do calculations for a sample of size 1, i.e. $n = 1$. Then we have already found that:

$$\begin{aligned}
\boldsymbol{\theta} &= (\mu, \sigma) \\
\ell(\boldsymbol{\theta}) &= -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu} &= \left(\frac{x-\mu}{\sigma^2}\right) \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma} &= -\frac{1}{\sigma} + \frac{1}{\sigma^3}(x-\mu)^2
\end{aligned}$$

- But to compute the information matrix we also need

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(x-\mu)^2 \\
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma \partial \mu} &= -\frac{2}{\sigma^3}(x-\mu)
\end{aligned}$$

- Upon taking the negative of the expectations of the second partial derivatives, the information matrix for a normal density is

$$I(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

- We may want the information matrix for $I(\mu, \sigma^2)$.
- This can be obtained by taking partial derivatives with respect to σ^2 instead of σ.
- But we will obtain it via a transformation. Consider the transformation $g(\mu, \sigma) = \left(\mu, \sigma^2\right)$
- The corresponding 2×2 matrix of partial derivatives is

$$\mathbf{B} = \left[\frac{\partial g_i}{\partial \theta_j}\right] = \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma \end{bmatrix}$$

- Hence the information matrix for $\nu = g(\mu, \sigma) = \left(\mu, \sigma^2\right)$ is

$$I(\nu) = \left[\mathbf{B} I^{-1}(\mu, \sigma) \mathbf{B}'\right]^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

Table of Contents

Definitions and Examples

Consistency of MLE

Efficiency of MLE

Numerical Computation of MLE

Estimation in Multiparameter Case

Likelihood Ratio Test

- To motivate the test, consider Theorem 1, which says that if θ_0 is the true value of θ, then, asymptotically, $L(\theta_0)$ is the max. value of $L(\theta)$.
- Now, consider the ratio of two likelihood functions, namely,

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$$

- Consider the two-sided hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

  where θ_0 is a specified value.
- Note that $\Lambda \leq 1$, but if H_0 is true, Λ should be large (close to 1), while if H_1 is true, Λ should be smaller.
- This leads to the intuitive decision rule for, for a specified significance level α,

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } \Lambda \leq c,$$

  where c is such that $\alpha = P_{\theta_0}[\Lambda \leq c]$.
- We call it the **likelihood ratio test** (LRT)

LR Test for the Mean of a Normal Distribution

- Consider a random sample $X_1, \ldots, X_n$ from a $N(\theta, \sigma^2)$ distribution and σ^2 is known.
- Consider the two-sided hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

  where θ_0 is a specified value.
- The likelihood function is

$$\begin{aligned} L(\theta) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-(2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\theta)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-(2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \exp\left\{-(2\sigma^2)^{-1}n(\bar{x}-\theta)^2\right\} \end{aligned}$$

- We know that for this example the mle $\hat{\theta} = \bar{x}$ and thus

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} = \exp\left\{-(2\sigma^2)^{-1}n(\bar{x}-\theta_0)^2\right\}$$

- Then $\Lambda \leq c$ is equivalent to $-2\log\Lambda \geq -2\log c$.

LR Test for the Mean of a Normal Distribution

- On the other hand,
$$-2\log\Lambda = \left(\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right)^2$$
which has a $\chi^2(1)$ distribution under H_0.
- Thus, the likelihood ratio test with significance level α states that we reject H_0 and accept H_1 when
$$-2\log\Lambda = \left(\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right)^2 \geq \chi^2_\alpha(1)$$

Theorem 10

Assume the same regularity conditions as for Theorem 5. Under the null hypothesis $H_0 : \theta = \theta_0$,
$$-2\log\Lambda \xrightarrow{d} \chi^2(1)$$