Parametric Estimation

Maxim Panov

Skoltech

November, 2021

Outline

Outline

Problem Statement

Data: $X_1, \ldots, X_n \sim f_\circ$.

Model: we aim to model the data by $f(x; \theta)$ for some value of θ.

Definition

General form of a parametric model:

$$\mathfrak{F} = \left\{ f(x; \theta), \theta \in \Theta \subseteq \mathbb{R}^k \right\},$$

where Θ is the parameter space, $\theta = (\theta_1, \ldots, \theta_k)$ is parameter vector and $k \in \mathbb{N}$.

Problem: find an estimate of $T(\theta)$, where T is some function.

Usual assumption: There exists θ_* such that $f_\circ = f(x; \theta_*)$.

Problem Statement

Model: we aim to model the data by $f(x;\theta)$ for some value of θ.

Problem: find an estimate of $T(\theta)$, where T is some function.

Example

- Consider a random variable with the distribution $\mathcal{N}(\mu, \sigma^2)$.
- In this case $\theta = (\mu, \sigma)$.
- If the task is to estimate just μ, then $\mu = T(\theta)$.
- In this case σ is called a *nuisance* parameter.

Example

Example

- Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
- $\theta = (\mu, \sigma)$ is a vector from a parameter space $\Theta = \{(\mu, \sigma) \colon \mu \in \mathbb{R}, \sigma > 0\}$.
- Suppose that X_i is a certain integral characteristic of a blood sample.
- **Problem** is stated as follows: estimate τ is a fraction of the blood samples, for which this value exceeds 1.

$$\tau = \mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right) = \\ = 1 - \mathbb{P}\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right).$$

- $\tau = T(\mu, \sigma) = 1 - \Phi((1-\mu)/\sigma)$ is our parameter of interest and Z is a standard normal random variable.

One More Example

Example

- Let $X \sim Gamma(\alpha, \beta)$, that has the following density:

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad \text{where } \alpha, \beta, x > 0.$$

- Here $\theta = (\alpha, \beta)$ is a parameter vector and $\Gamma(\alpha)$ is Gamma function:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

- Gamma distribution may be used to model life expectancy.
- If the task is to estimate the average life expectancy then

$$T(\alpha, \beta) = \mathbb{E}_\theta(X) = \alpha\beta.$$

Method of Moments

- Let $\theta = (\theta_1, \ldots, \theta_k)$ be a parameter vector.
- For $1 \leq j \leq k$ define j-th (non-centered) moment as follows:

$$\alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x) = \int x^j f(x;\theta)dx,$$

- If we are given the values $\alpha_1, \ldots, \alpha_k$ then we can consider the system of equations:

$$\begin{aligned} \alpha_1(\theta) &= \alpha_1, \\ \alpha_2(\theta) &= \alpha_2, \\ &\ldots \\ \alpha_k(\theta) &= \alpha_k. \end{aligned}$$

- If the system above has solution, then it can be used as an estimate of θ_*.

Method of Moments

Idea: use δ-method by computing the corresponding j-th sample moment:

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

and plugging it into the equations.

Definition

$\widehat{\theta}_n$ is a method of moments estimate of $\theta = (\theta_1, \ldots, \theta_k)$ if

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1, \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2, \\ &\ldots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned}$$

Example

Example

- Let $X_1, \ldots, X_n \sim Bernoulli(p)$.
- Find an estimate of parameter p.

Example

Example

- Let $X_1, \ldots, X_n \sim Bernoulli(p)$.
- Find an estimate of parameter p.
- $\alpha_1 = \mathbb{E}_\theta(X) = p$ and $\widehat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$,
- from which we get $\widehat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

One More Example

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Estimate parameters μ and σ.

One More Example

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Estimate parameters μ and σ.

$$\begin{aligned}
\alpha_1 &= \mathbb{E}_\theta(X_1) = \mu, \\
\alpha_2 &= \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2, \\
\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2.
\end{aligned}$$

Solving the system of equations we get that

$$\widehat{\mu}_n = \overline{X}_n \quad \text{and} \quad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Properties of Method of Moments

Theorem

If $\widehat{\theta}_n$ is a method of moments estimate of θ then (given certain assumptions about the distribution of the sample) the following properties hold:

1. *$\widehat{\theta}_n$ exists with probability 1;*
2. *$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_*$ when $n \to \infty$;*
3. *estimate is asymptotically normal:*

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where $\Sigma = g\mathbb{E}(YY^T)g^T$, $Y = (X, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ and $g_j = \partial\alpha_j^{-1}(\theta_*)/\partial\theta$.

Remark: the last property can be used to derive standard errors and confidence intervals.

Method of Moments: Comments

- not optimal;
- easy to use;
- estimates can be used as initial values for more “accurate” methods.

Deriving MLE

- If we know the true density, then we can estimate θ in the following way:

$$\theta^*(f_\circ) = \arg\min_\theta \int_{\mathbb{R}^n} f_\circ(z) \log \frac{f_\circ(z)}{f(z;\theta)}\, dz.$$

- It easy to see that

$$\theta^*(f_\circ) = \arg\max_\theta \int_{\mathbb{R}^n} f_\circ(z) \log[f(z;\theta)]\, dz.$$

- Let us note that the knowledge of $f_\circ(z)$ is needed.
- However, importantly, the functional

$$L[f_\circ, \theta] = \int_{\mathbb{R}^n} f_\circ(z) \log[f(z;\theta)]\, dz$$

is linear in $f_\circ(\cdot)$!

Deriving MLE

Idea: let's estimate the functional $L[f_\circ, \theta]$ based on X^n via δ-method.

- It leads to estimation of the *logarithm of likelihood*:

$$\bar{L}(\theta; X^n) = \int_{\mathbb{R}^n} \delta(z - X^n) \log\big[f(z;\theta)\big]\, dz = \log\big[f(X^n;\theta)\big].$$

- Maximum likelihood method:

$$\hat{\theta}(X^n) = \arg\max_{\theta}\{\log[f(X^n;\theta)]\}.$$

- Maximum likelihood method was suggested by Ronald Fisher when he was 22 years old.

Maximum Likelihood Estimation

For the linear model

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

we obviously have

$$\bar{L}(\mu; Y^n) = \sum_{i=1}^{n} \log\big[f(Y_i - \mu)\big].$$

Thus, MLE reads as

$$\hat{\mu}(Y^n) = \arg\max_{\mu} \left\{ \sum_{i=1}^{n} \log[f(Y_i - \mu)] \right\}.$$

Maximum Likelihood Estimation

As random variables Y_i are i.i.d., then

$$\mathbb{E}\big|\log[f(Y_1 - \tilde{\mu})]\big| < \infty,$$

for any $\tilde{\mu}$, then by Law of Large Numbers for fixed $\tilde{\mu}$ and $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} \log\big[f(Y_i - \tilde{\mu})\big] \to \mathbb{E}\log\big[f(Y_1 - \tilde{\mu})\big] = \int_{\mathbb{R}^1} f_\circ(x)\log[f(x - \tilde{\mu})]\,dx,$$

where $f_\circ(x)$ is the true density of Y_1.

That's why for large n:

$$\hat{\mu}(Y^n) \to \arg\max_{\tilde{\mu}} \mathbb{E}\log\big[f(Y_1 - \tilde{\mu})\big] = \arg\min_{\tilde{\mu}} \mathbb{E}\log\frac{f_\circ(Y_1)}{f(Y_1 - \tilde{\mu})}.$$

Maximum Likelihood Estimation

Definition

Consider an $i.i.d.$ estimate $X_1, \ldots, X_n \sim F$ and the distribution has a density $f(x;\theta)$. The *likelihood function* is defined as follows:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Log-likelihood is just taking the logarithm of the expression above:

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta).$$

- We will view likelihood as a function of the model's parameters $\mathcal{L}_n \colon \Theta \to [0, \infty)$.
- A maximum likelihood estimate (MLE):

$$\hat{\theta}_n \equiv \hat{\theta}(X^n) = \arg\max_{\theta} \ell_n(\theta) = \arg\max_{\theta} \log \mathcal{L}_n(\theta).$$

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. Find MLE of the parameter p.

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. Find MLE of the parameter p.

Probability mass function: $f(x;p) = p^x(1-p)^{1-x}$ where $x = 0, 1$.

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. Find MLE of the parameter p.

Probability mass function: $f(x; p) = p^x (1-p)^{1-x}$ where $x = 0, 1$.

Likelihood function:

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} = p^S (1-p)^{n-S},$$

where $S = \sum_{i=1}^{n} X_i$.

Example

Let $X_1, \dots, X_n \sim Bernoulli(p)$. Find MLE of the parameter p.

Probability mass function: $f(x; p) = p^x (1-p)^{1-x}$ where $x = 0, 1$.

Likelihood function:

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} = p^S (1-p)^{n-S},$$

where $S = \sum_{i=1}^{n} X_i$.

Log-likelihood:

$$\ell_n(p) = S \log p + (n - S) \log(1-p).$$

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. Find MLE of the parameter p.

Probability mass function: $f(x;p) = p^x(1-p)^{1-x}$ where $x = 0, 1$.

Likelihood function:

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i;p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S},$$

where $S = \sum_{i=1}^{n} X_i$.

Log-likelihood:

$$\ell_n(p) = S\log p + (n-S)\log(1-p).$$

From this we get that MLE estimate is $\widehat{p}_n = S/n$.

One More Example

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
$\theta = (\mu, \sigma)$ is our parameter of interest.

One More Example

Example

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
$\theta = (\mu, \sigma)$ is our parameter of interest.
Likelihood function has the following form (up to some constants):

$$\begin{aligned}\mathcal{L}_n(\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\right\} = \\ &= \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right\} = \\ &= \frac{1}{\sigma^n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right\},\end{aligned}$$

where $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i, \quad S^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
(continued on the next slide)

One More Example

Example (continued)

The last statement follows from:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2,$$

which can be seen from:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - \mu)^2.$$

Log-likelihood becomes:

$$\ell_n(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}.$$

Let $\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = 0$ and $\frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = 0$, then MLE gives $\widehat{\mu}_n = \overline{X}$ and $\widehat{\sigma}_n = S$.

Properties of MLEs

- consistent, that is $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_*$, where θ_* is true value of the parameter θ;
- *equivariant*: if $\widehat{\theta}_n$ is MLE for θ then $g(\widehat{\theta}_n)$ is MLE for $g(\theta)$;
- asymptotically normal: $(\widehat{\theta}_n - \theta_*)/\widehat{se} \rightsquigarrow \mathcal{N}(0,1)$;
- asymptotically optimal or efficient (for a sufficient sample size it has lower variance).

Remark: the properties of MLEs stated above hold when $f(x;\theta)$ is sufficiently regular. In "difficult" cases MLEs "lose" these properties.

Consistency of MLE

- Maximizing $\ell_n(\theta)$ is equivalent to maximizing $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$,
- because $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_*))$ and $\ell_n(\theta_*)$ is a constant.
- Then

$$\mathbb{E}_{\theta_*}\left(\log \frac{f(x;\theta)}{f(x;\theta_*)}\right) = \int \log\left(\frac{f(x;\theta)}{f(x;\theta_*)}\right) f(x;\theta_*)dx =$$
$$= -\int \log\left(\frac{f(x;\theta_*)}{f(x;\theta)}\right) f(x;\theta_*)dx = -K(\theta_*,\theta).$$

- Hence, $M_n(\theta) \approx -K(\theta_*,\theta)$ attains its maximum at θ_* since $-K(\theta_*,\theta_*) = 0$ and $-K(\theta_*,\theta) < 0$ for $\theta \neq \theta_*$.
- We need to show that MLE estimate converges *in probability* to the true value of the parameter.

Consistency of MLE

Theorem

- *Let θ_* denote the true value of the parameter θ. Define*

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

and $M(\theta) = -K(\theta_*, \theta)$.

- *Suppose that $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$*
- *and for every $\epsilon > 0$ $\sup_{\theta:\,|\theta-\theta_*|\geq\epsilon} M(\theta) < M(\theta_*)$.*
- *Let $\widehat{\theta}_n$ denote the MLE, then*

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_*.$$

Skoltech
Skolkovo Institute of Science and Technology

Consistency of MLE

Proof: since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n(\widehat{\theta}_n) \geq M_n(\theta_*)$. It follows that,

$$
\begin{aligned}
M(\theta_*) - M(\widehat{\theta}_n) &= M_n(\theta_*) - M(\widehat{\theta}_n) + M(\theta_*) - M_n(\theta_*) \\
&\leq M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_*) - M_n(\theta_*) \\
&\leq \sup_{\theta} |M_n(\theta) - M(\theta)| + M(\theta_*) - M_n(\theta_*) \xrightarrow{\mathbb{P}} 0.
\end{aligned}
$$

Then for any $\delta > 0$:

$$
\mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_*) - \delta\right) \to 0.
$$

Skoltech

Skolkovo Institute of Science and Technology

Consistency of MLE

- Take arbitrary $\epsilon > 0$.
- By assumption there exists $\delta > 0$, for which the inequality $|\theta - \theta_*| \geq \epsilon$ implies that $M(\theta) < M(\theta_*) - \delta$.
- Hence,

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta_*| > \epsilon\right) \leq \mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_*) - \delta\right) \to 0.$$

Equivariance

Theorem

Let $\tau = g(\theta)$ be a function of the parameter θ and let $\widehat{\theta}_n$ be its MLE estimate.

Then $\widehat{\tau}_n = g(\widehat{\theta}_n)$ is MLE for $\tau = g(\theta)$.

Proof:

- Assume that g is one to one mapping. i.e. an inverse function $h = g^{-1}$ exists.
- For any τ and $\theta = h(\tau)$:

$$\mathcal{L}_n(\theta) = \mathcal{L}_n\big(h(\tau)\big).$$

- Due to the mapping being one to one the maximizers coincide:

$$\hat{\theta}_n = h(\hat{\tau}_n)$$

and consequently

$$\hat{\tau}_n = g(\hat{\theta}_n).$$

Equivariance

Example

- Let $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$.
- MLE of θ equals $\widehat{\theta}_n = \overline{X}_n$.
- Let $\tau = e^{\theta}$.
- Then MLE for τ equals $\widehat{\tau}_n = e^{\widehat{\theta}_n} = e^{\overline{X}}$.

Asymptotic Normality

- Let $s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$ be a *score function*.
- Then Fisher information is defined as:

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right) = \sum_{i=1}^{n} \mathbb{V}_\theta\big(s(X_i;\theta)\big).$$

Lemma

Let $f(x;\theta)$ be continuously differentiable in θ. Then it holds

$$\mathbb{E}_\theta\big(s(X;\theta)\big) = 0 \ \ \textit{and} \ \ \mathbb{V}_\theta\big(s(X;\theta)\big) = \mathbb{E}_\theta\big(s^2(X;\theta)\big).$$

Proof: We know that $1 = \int f(x;\theta)dx$.

Then

$$0 = \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \int \frac{\partial}{\partial\theta} f(x;\theta)dx = \int \frac{\partial f(x;\theta)}{\partial\theta}\cdot\frac{1}{f(x;\theta)} f(x;\theta)dx$$

$$= \int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx = \int s(x;\theta) f(x;\theta)dx = \mathbb{E}_\theta\left[s(X;\theta)\right].$$

Asymptotic Normality

Theorem

The following equality holds: $I_n(\theta) = nI(\theta)$.

Also,

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right) = -\int\left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right) f(x;\theta)dx.$$

Asymptotic Normality

Theorem

Let $se = \sqrt{\mathbb{V}(\widehat{\theta}_n)}$. *Under appropriate regularity conditions, the following hold:*

1. $se \approx \sqrt{1/I_n(\theta_*)}$ *and* $\frac{\widehat{\theta}_n - \theta_*}{se} \rightsquigarrow \mathcal{N}(0,1)$,
2. *Let* $\widehat{se} = \sqrt{1/I_n(\widehat{\theta}_n)}$. *Then* $\frac{\widehat{\theta}_n - \theta_*}{\widehat{se}} \rightsquigarrow \mathcal{N}(0,1)$.

Asymptotic Normality

Proof:
Let $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

Then $0 = \ell_n'(\widehat{\theta}_n) \approx \ell_n'(\theta_*) + (\widehat{\theta}_n - \theta_*)\ell_n''(\theta_*)$.

We obtain

$$\widehat{\theta}_n - \theta_* \approx -\ell_n'(\theta_*)/\ell_n''(\theta_*),$$
$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \approx \frac{\frac{1}{\sqrt{n}}\ell_n'(\theta_*)}{-\frac{1}{n}\ell_n''(\theta_*)}. \tag{1}$$

Asymptotic Normality

Let $Y_i = \frac{\partial \log f(X_i;\theta)}{\partial \theta}$.

From the Lemma on the Slide 32 it follows that $\mathbb{E}(Y_i) = 0$ and $\mathbb{V}(Y_i) = I(\theta_*)$.

Then, according to CLT, for the numerator in (1) it holds that

$$n^{-1/2} \sum_{i=1}^{n} Y_i = \sqrt{n}\overline{Y} = \sqrt{n}(\overline{Y} - 0) \rightsquigarrow W \sim \mathcal{N}(0, I(\theta_*)).$$

Asymptotic Normality

Define $A_i = -\partial^2 \log f(X_i;\theta)/\partial\theta^2$. Then $\mathbb{E}(A_i) = I(\theta_*)$ and for the denominator in (1) it holds that $\overline{A} \xrightarrow{\mathbb{P}} I(\theta_*)$.
Hence,

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \frac{W}{I(\theta_*)} \stackrel{d}{=} \mathcal{N}\left(0, \frac{1}{I(\theta_*)}\right).$$

Assume that $I(\theta)$ is a continuous function of its argument, then $I(\widehat{\theta}_n) \xrightarrow{\mathbb{P}} I(\theta_*)$ and

$$\frac{\widehat{\theta}_n - \theta_*}{\widehat{se}} = \sqrt{n} I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_*) = \left\{\sqrt{n} I^{1/2}(\theta_*)(\widehat{\theta}_n - \theta_*)\right\} \sqrt{\frac{I(\widehat{\theta}_n)}{I(\theta_*)}}.$$

The first factor converges in distribution to $\mathcal{N}(0,1)$, the second – to 1. $\square$

Cramér–Rao bound

Theorem

Let $\hat{\theta}_n = g(X_1, \dots, X_n)$ be an unbiased estimate of θ_*. Then

$$\mathbb{V}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta_*)}.$$

Proof: Since estimate $\hat{\theta}_n$ is unbiased for $\theta = \theta_*$, we have:

$$1 = \int g(x_1, \dots, x_n) \frac{\partial f(x_1, \dots, x_n; \theta)}{\partial \theta} dx_1 \dots dx_n$$

$$= \int g(x_1, \dots, x_n) \frac{\partial \log f(x_1, \dots, x_n; \theta)}{\partial \theta} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n$$

$$= \mathbb{E}_\theta \, \hat{\theta}_n \, s(X_1, \dots, X_n; \theta) = \mathbb{E}_\theta \, \hat{\theta}_n \left[\sum_{i=1}^n s(X_i; \theta) \right].$$

Cramér–Rao bound

Proof (continued): Hence we arrive at:

$$\mathbb{E}_\theta \, \hat{\theta}_n \left[\sum_{i=1}^n s(X_i;\theta) \right] = 1.$$

But, since $\mathbb{E}_\theta\big(s(X;\theta)\big) = 0$:

$$\mathbb{E}_\theta \, \hat{\theta}_n \left[\sum_{i=1}^n s(X_i;\theta) \right] = \mathbb{E}_\theta \, (\hat{\theta}_n - \theta) \left[\sum_{i=1}^n s(X_i;\theta) \right] = Cov_\theta\big(\hat{\theta}_n, \sum_{i=1}^n s(X_i;\theta)\big).$$

Using $I_n(\theta) = \mathbb{V}_\theta\big(\sum_{i=1}^n s(X_i;\theta)\big)$ and $Cov_\theta(X,Y) \leq \sqrt{\mathbb{V}_\theta(X)\mathbb{V}_\theta(Y)}$ we get

$$1 \leq \mathbb{V}_\theta(\hat{\theta}_n) I_n(\theta). \; \square$$

Optimality

- Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$.
- MLE for θ is $\widehat{\theta}_n = \overline{X}_n$.
- Denote the *sample median* by $\widetilde{\theta}_n$, it can also be used to estimate θ.

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \sigma^2);$$
$$\sqrt{n}(\widetilde{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}\left(0, \sigma^2 \frac{\pi}{2}\right).$$

- More generally, consider two estimators T_n and U_n and suppose that

$$\sqrt{n}(T_n - \theta_*) \rightsquigarrow \mathcal{N}(0, t^2);$$
$$\sqrt{n}(U_n - \theta_*) \rightsquigarrow \mathcal{N}(0, u^2).$$

Optimality

We define the *asymptotic relative efficiency* of U_n to T_n by $ARE(U,T) = t^2/u^2$.

In the Normal example above, $ARE(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = 0.63$.

Asymptotic relative efficiency can be interpreted as a fraction of the data that is "effectively" used for the estimation.

Theorem

Let $\widehat{\theta}_n$ be the MLE and $\widetilde{\theta}_n$ is any other estimator. Then, assuming appropriate regularity conditions (e.g. that the assumed parametric model is true), this

$$ARE(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1.$$

Thus, MLE has the smallest (asymptotic) variance and we say that the MLE is efficient or asymptotically optimal.

Delta-method

- Let $\tau = g(\theta)$ where g is a smooth function.
- MLE for τ is $\widehat{\tau} = g(\widehat{\theta})$.
- What is the distribution[1] of $\widehat{\tau}$?

Theorem

If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$, then

$$\frac{\widehat{\tau}_n - \tau_*}{\widehat{se}(\widehat{\tau}_n)} \rightsquigarrow \mathcal{N}(0, 1),$$

where $\widehat{\tau}_n = g(\widehat{\theta}_n)$ and $\widehat{se}(\widehat{\tau}_n) = |g'(\widehat{\theta}_n)| \, \widehat{se}(\widehat{\theta}_n)$.

[1] Distribution of an estimator is also called its sampling distribution

Theorem (continued)

Proof:

$$\begin{aligned}\widehat{\tau}_n = g(\widehat{\theta}_n) &\approx g(\theta_*) + (\widehat{\theta}_n - \theta_*)g'(\theta_*) = \tau_* + (\widehat{\theta}_n - \theta_*)g'(\theta_*),\\ \sqrt{n}(\widehat{\tau}_n - \tau_*) &\approx \sqrt{n}(\widehat{\theta}_n - \theta_*)g'(\theta_*),\\ \frac{\sqrt{nI(\theta_*)}(\widehat{\tau}_n - \tau_*)}{g'(\theta_*)} &\approx \sqrt{nI(\theta_*)}(\widehat{\theta}_n - \theta_*).\end{aligned}$$

(finished on the next slide)

Theorem (continued)

We know that $\sqrt{nI(\theta_*)}(\widehat{\theta}_n - \theta_*)$ *converges in distribution to* $\mathcal{N}(0,1)$. *Then*

$$\frac{\sqrt{nI(\theta_*)}(\widehat{\tau}_n - \tau_*)}{g'(\theta_*)} \rightsquigarrow \mathcal{N}(0,1).$$

From this we can conclude that

$$\widehat{\tau}_n \approx \mathcal{N}\left(\tau_*, se^2(\widehat{\tau}_n)\right),\ se^2(\widehat{\tau}_n) = \frac{\left(g'(\theta_*)\right)^2}{nI(\theta_*)}.$$

If we replace θ_* *with* $\widehat{\theta}_n$ *the result will still hold.*

Example

Example

- Let $X_1, \ldots, X_n \sim Bernoulli(p)$.
- Statistic to estimate:
$$\psi = g(p) = \log \frac{p}{1-p}.$$
- Fisher information is
$$I(p) = \frac{1}{p(1-p)}.$$
- Estimate of the standard error is
$$\widehat{se} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

Example

Example (continued)

MLE of ψ is

$$\widehat{\psi}_n = \log \frac{\widehat{p}_n}{1-\widehat{p}_n}.$$

Since $g'(p) = 1/(p(1-p))$, according to the delta method we obtain

$$\widehat{se}(\widehat{\psi}_n) = |g'(\widehat{p}_n)|\ \widehat{se}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

Example

- Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
- Suppose that μ is known and σ is unknown.
- We want to estimate $\psi = \log \sigma$.
- The log-likelihood is

$$\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

(continued on the next slide)

Example

Example (continued)

Taking the derivative and equating it to zero we get:

$$\widehat{\sigma}_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}}.$$

To get the standard error we need the Fisher information. First,

$$\begin{aligned} \log f(X;\sigma) &= -\log\sigma - \frac{(X-\mu)^2}{2\sigma^2}, \\ \frac{\partial^2 \log f(X;\sigma)}{\partial\sigma^2} &= \frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4}, \\ I(\sigma) &= -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}. \end{aligned}$$

(continued on the next slide)

Example

Example (continued)

- Thus, we obtain
$$\widehat{se} = \frac{\widehat{\sigma}_n}{\sqrt{2n}}.$$
- Let $\psi = g(\sigma) = \log \sigma$, then $\widehat{\psi}_n = \log \widehat{\sigma}_n$.
- Since $g'(\sigma) = 1/\sigma$, we have
$$\widehat{se}(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}_n} \frac{\widehat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}.$$

We can extend the idea of the delta method to the models with several parameters. Let $\theta = (\theta_1, \dots, \theta_k)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)$ be the MLE for θ. Log-likelihood will be

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

Let

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial \theta_j^2}, \;\; H_{jk} = \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_k}.$$

Define the Fisher Information Matrix by

$$I_n(\theta) = -\begin{pmatrix} \mathbb{E}_\theta(H_{11}) & \mathbb{E}_\theta(H_{12}) & \cdots & \mathbb{E}_\theta(H_{1k}) \\ \mathbb{E}_\theta(H_{21}) & \mathbb{E}_\theta(H_{22}) & \cdots & \mathbb{E}_\theta(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}_\theta(H_{k1}) & \mathbb{E}_\theta(H_{k2}) & \cdots & \mathbb{E}_\theta(H_{kk}) \end{pmatrix}.$$

Define the precision matrix:

$$J_n(\theta) = I_n^{-1}(\theta).$$

Theorem

Under appropriate regularity conditions,

$$\widehat{\theta} - \theta_* \rightsquigarrow \mathcal{N}(0, J_n).$$

Also, if $\widehat{\theta}_j$ *is the* j*-th component of vector* $\widehat{\theta}$

$$\frac{\widehat{\theta}_j - \theta_{j,*}}{\widehat{se}_j} \rightsquigarrow \mathcal{N}(0, 1),$$

where $\widehat{se}_j^2 = J_n(j, j)$ *is the* j*-th diagonal element of the matrix* J_n*.*
The approximate covariance is

$$Cov(\widehat{\theta}_j, \widehat{\theta}_k) \approx J_n(j, k).$$

Let $\tau = g(\theta_1, \ldots, \theta_k)$ be a function of the parameters and let its gradient be

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_k} \end{pmatrix}.$$

Theorem

Suppose that $\nabla g(\widehat{\theta}) \neq 0$. *Let* $\widehat{\tau} = g(\widehat{\theta})$. *Then*

$$\frac{\widehat{\tau} - \tau_*}{\widehat{se}(\widehat{\tau})} \rightsquigarrow \mathcal{N}(0, 1),$$

where $\widehat{se}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)}$, $\widehat{J}_n = J_n(\widehat{\theta})$, $\widehat{\nabla} g = \nabla g(\widehat{\theta})$.

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\tau = g(\mu, \sigma) = \sigma/\mu$. The Fisher Information Matrix is

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

$$J_n = I_n^{-1}(\mu, \sigma) = \frac{1}{n} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}, \; \nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix},$$

$$\widehat{se}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\widehat{\mu}^4} + \frac{\widehat{\sigma}^2}{2\widehat{\mu}^2}}.$$

Thank you for your attention!