III

Information Inequality (Cramér-Rao bound)

1. $\underline{X} \sim p(\underline{x};\theta)$, $\theta \in \Theta \subset \mathbb{R}$.

Define $\underbrace{u(\underline{X};\theta)}_{\text{score fn.}} \equiv \frac{\partial \log p(\underline{X};\theta)}{\partial \theta}$

Can prove: $E_\theta u(\underline{X};\theta) = 0$

Estimating equation $u(\theta;\underline{x}) = 0$.

Define Fisher Information: $I_{\underline{X}}(\theta) \equiv Var_\theta\, u(\underline{X};\theta)$.

Note: 1° (Information inequality) $g(\theta) \equiv E_\theta T(\underline{X})$, then

$$|g'(\theta)| = \left|E_\theta\{T(\underline{X})\, u(\underline{X};\theta)\}\right| \le \sqrt{Var_\theta T(\underline{X})\ I_{\underline{X}}(\theta)}$$

If $T(\underline{X})$ unbiased for θ, $Var_\theta T(\underline{X}) \ge \frac{1}{I_{\underline{X}}(\theta)}$

Note: 1° A UMVU estimator does not have to achieve the info. bound.

$\therefore$ $Var_\theta T(\underline{X}) \ge \frac{(g'(\theta))^2}{I_{\underline{X}}(\theta)}$, equality hold iff $T(\underline{X})$ and $u(\underline{X};\theta)$ are linearly dependent.

(additivity) 2° $X_1, \cdots, X_n$ independent $p_i(x;\theta)$, then $I_{\underline{X}}(\theta) = \sum_{i=1}^n I_{X_i}(\theta)$

--- iid ---, then $I_{\underline{X}}(\theta) = n I_{X_i}(\theta)$

3° $I_{\underline{X}}(\theta) = -E_\theta \frac{\partial^2 \log p(\underline{X};\theta)}{\partial \theta^2}$

$\because$ $E_\theta \frac{\partial^2 \log p}{\partial\theta^2} = E_\theta \frac{p''p-(p')^2}{p^2} = E_\theta \frac{p''}{p} - E_\theta u^2$

$= -I_{\underline{X}}(\theta) + E_\theta \frac{p''}{p}$

while $E_\theta \frac{p''}{p} = \int p''\, d\underline{x} = \frac{\partial^2}{\partial\theta^2}\int p\, d\underline{x} = 0$.

$\therefore$ $-E_\theta \frac{\partial^2 \log p}{\partial \theta^2} = I_{\underline{X}}(\theta)$ □.

Eg: 1° $X \sim N(\mu, \sigma^2)$,

(information method) $\begin{cases} u(X;\mu) = \frac{1}{\sigma^2}(X-\mu) \\ I_X(\mu) = \frac{1}{\sigma^2} \end{cases}$ $\begin{cases} u(\underline{X};\mu) = \frac{n}{\sigma^2}(\bar{X}-\mu) \\ I_{\underline{X}}(\mu) = \frac{n}{\sigma^2} \end{cases}$

$\overline{X}$ is an unbiased estimator for μ.

$Var_\mu \overline{X} = \frac{\sigma^2}{n}$

Let $T(\underline{X})$ be any unbiased estimator for μ, information inequality gives

$$Var_\mu T(\underline{X}) \geq \frac{1}{I_X(\mu)} = \frac{\sigma^2}{n} = Var_\mu \overline{X}$$

$\therefore \overline{X}$ is UMVU for μ.

2° Same as 1°,

$\begin{cases} T(\underline{X}) \text{ is unbiased for } \mu \\ T(\underline{X}) = \overline{X}, \text{ and } u(\underline{X};\mu) = \frac{n}{\sigma^2}(\overline{X}-\mu) \text{ is linearly dependent,} \end{cases}$

$\therefore \overline{X}$ is UMVU for μ.

3° Same as 1°

$$\begin{cases} u(X;\sigma^2) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2 \\ I_X(\sigma^2) = \frac{1}{2\sigma^4} \quad \Rightarrow I_{\underline{X}}(\sigma^2) = \frac{n}{2\sigma^4} \end{cases}$$

$\hat{\sigma}^2$ is an unbiased estimator for σ^2. $\left(\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\right)$

$$Var_{\sigma^2}\hat{\sigma}^2 = \frac{1}{n^2}\sum_{i=1}^n Var_{\sigma^2}(X_i-\mu)^2 = \frac{\sigma^4}{n^2}\sum_{i=1}^n Var\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{2\sigma^4}{n}$$

$\therefore \hat{\sigma}^2$ achieves the C-R bound

$\therefore \hat{\sigma}^2$ is UMVU for σ^2.

4° Same as 1°.

$\begin{cases} T(\underline{X}) = \hat{\sigma}^2 \text{ is unbiased for } \sigma^2 \\ T(\underline{X}) = \hat{\sigma}^2 \text{ and } u(\underline{X};\sigma^2) = -\frac{n}{2\sigma^4} + \frac{n}{2\sigma^4}\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\right] \text{ are} \end{cases}$

linearly dependent,

$\therefore \hat{\sigma}^2$ is UMVU for σ^2.

Eg. : 5° (Location-scale family)

$$Y \sim \frac{1}{\sigma} p\left(\frac{y-\mu}{\sigma}\right)$$

$$u(Y;\mu) = \frac{\frac{\partial}{\partial\mu} p\left(\frac{y-\mu}{\sigma}\right)}{p\left(\frac{y-\mu}{\sigma}\right)}$$

$$I_Y(\mu) = \frac{1}{\sigma^2} \int_{-\infty}^{+\infty} \frac{p'(\xi)^2}{p(\xi)} d\xi$$

For $N(\mu, \sigma^2)$

$$I_Y(\mu) = \frac{1}{\sigma^2} \int_{-\infty}^{+\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{1}{\sigma^2}$$

For Cauchy(μ, σ^2)

$$I_Y(\mu) = \frac{1}{\sigma^2} \frac{4}{\pi} \int_{-\infty}^{+\infty} \frac{y^2}{(1+y^2)^3} dy = \frac{1}{2\sigma^2}$$

6° $X \sim U(0,\theta)$, shown that $T(\underline{X}) = \frac{n+1}{n} X_{(n)}$ UMVU for θ.

$\mathrm{Var}_\theta T(\underline{X}) = \frac{\theta^2}{n(n+2)} \sim O(\frac{1}{n^2})$, goes down faster than $\frac{1}{nI_X(\theta)}$.

(proof: $\mathrm{Var}_\theta T(\underline{X}) = \left(\frac{n+1}{n}\right)^2 \left[E_\theta X_{(n)}^2 - (E_\theta X_{(n)})^2 \right]$

$f_{X_{(n)}}(x) = \frac{n x^{n-1}}{\theta^n}$,

$E_\theta X_{(n)}^2 = \frac{n}{n+2}\theta^2$, $E_\theta X_{(n)} = \frac{n}{n+1}\theta$)

7° $p(x;\alpha) = \alpha x^{\alpha-1}$ $(x \in [0,1])$

$u(X;\alpha) = \frac{1}{\alpha} + \log X \Rightarrow E\log X = -\frac{1}{\alpha}$

$u(\underline{X};\alpha) = \frac{n}{\alpha} + \sum_{i=1}^n \log X_i \Rightarrow E\left(\frac{1}{n}\sum_{i=1}^n \log X_i\right) = -\frac{1}{\alpha}$

$\because$ $T(\underline{X}) = \frac{1}{n}\sum_{i=1}^n \log X_i$ is unbiased for $-\frac{1}{\alpha}$, and

$T(\underline{X}) = \frac{1}{n}\sum_{i=1}^n \log X_i$ and $u(\underline{X};\alpha)$ are linearly dependent.

$\therefore$ $\frac{1}{n}\sum_{i=1}^n \log X_i$ is UMVU for $-\frac{1}{\alpha}$.

2、 *Multi-dimensional Information Inequality:*

$$\underline{X} \sim p(\underline{x};\underline{\theta}), \quad \underline{\theta} \in \Theta \subset \mathbb{R}^d$$

Define $\underline{U}(\underline{X};\underline{\theta}) \equiv \nabla_\theta \log p(\underline{X};\underline{\theta})$, $\boxed{\text{can prove } E_\theta U(\underline{X};\underline{\theta}) = \underline{0}}$

Fisher information $I_X(\underline{\theta}) = E_\theta\left[U(\underline{X};\underline{\theta})\, U^T(\underline{X};\underline{\theta})\right]$

$$\boxed{I_X(\underline{\theta}) = -E_\theta\left[\nabla_\theta U(\underline{X};\underline{\theta})\right]}$$

Estimator $T(\underline{X}) \in \mathbb{R}^r$,

$$g(\underline{\theta}) = E_\theta T(\underline{X})$$

Since
$$\nabla_\theta g(\underline{\theta}) = \int \nabla_\theta p(\underline{x};\underline{\theta})\, T(\underline{x})\, d\underline{x}$$
$$= \int \left(\nabla_\theta \log p(\underline{x};\underline{\theta})\right) T(\underline{x})\, p(\underline{x};\underline{\theta})\, d\underline{x}$$
$$= E_\theta\left[U(\underline{X};\underline{\theta})\, T(\underline{X})\right]$$
$$= Cov\left[U(\underline{X};\underline{\theta}), T(\underline{X})\right]$$

and
$$Var_\theta\left[T(\underline{X}) - (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} U(\underline{X};\underline{\theta})\right] \geq 0$$
$$= Var_\theta T(\underline{X}) - E_\theta\left[T_\theta(\underline{X})\, U^T(\underline{X};\underline{\theta})\right] I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$
$$- (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} E_\theta\left[U(\underline{X};\underline{\theta})\, T_\theta(\underline{X})\right]$$
$$+ (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} E_\theta\left[U(\underline{X};\underline{\theta})\, U^T(\underline{X};\underline{\theta})\right] I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$
$$= Var_\theta T(\underline{X}) - (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$
$$- (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$
$$+ (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} I_X(\underline{\theta}) I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$
$$= Var_\theta T(\underline{X}) - (\nabla_\theta g(\underline{\theta}))^T I_X(\underline{\theta})^{-1} \nabla_\theta g(\underline{\theta})$$

$$\therefore \quad Var_\theta T(\underline{X}) \geq (\nabla_\theta g(\underline{\theta}))^T I_X^{-1}(\underline{\theta})\, \nabla_\theta g(\underline{\theta}).$$

This is the multi-dimensional information inequality.

E.g.: 1° For a 2-D case.

$$\underline{\theta} = (\theta_1, \theta_2), \quad \mathbb{E}T(\underline{X}) = \theta_1.$$

If θ_2 known, $\operatorname{Var} T(\underline{X}) \geq \frac{1}{I_{11}}$

If θ_2 unkown, $\operatorname{Var} T(\underline{X}) \geq \frac{1}{I_{11}^*} = \frac{I_{22}}{I_{11}I_{22} - I_{12}^2}$

with effective information $I_{11}^* = \frac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11}(1-\rho^2)$.

2° For Gaussian distribution,

S^2 is UMVU for σ^2, while

$$\operatorname{Var} S^2 = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n},$$

the information bound is not achieved.

$$I_X(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

3° For Gamma distribution,

$$p(X, \alpha, \beta) = \frac{X^{\alpha-1} e^{-\frac{X}{\beta}}}{\Gamma(\alpha)\, \beta^{\alpha}} \qquad (X > 0)$$

$$\therefore \log p(X, \alpha, \beta) = (\alpha - 1)\log X - \frac{X}{\beta} - \log \Gamma(\alpha) - \alpha \log \beta$$

$$\nabla_{\underline{\theta}} \log p(X, \alpha, \beta) = \left(\log X - \psi(\alpha) - \log \beta, \ \frac{X}{\beta^2} - \frac{\alpha}{\beta}\right)$$

And
$$\begin{cases} \frac{\partial^2}{\partial \alpha^2} \log p(X, \alpha, \beta) = -\varphi(\alpha) \\ \frac{\partial^2}{\partial \beta^2} \log p(X, \alpha, \beta) = -\frac{2X}{\beta^3} + \frac{\alpha}{\beta^2}, \quad -\mathbb{E}\left[-\frac{2X}{\beta^3} + \frac{\alpha}{\beta^2}\right] = \frac{\alpha}{\beta^2} \\ \frac{\partial^2}{\partial \alpha \partial \beta} \log p(X, \alpha, \beta) = \frac{-1}{\beta} \end{cases}$$

$$\therefore I_X(\alpha, \beta) = \begin{pmatrix} \varphi(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix},$$

where
bi-Gamma fn $\psi(\alpha) \equiv \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$
tri-Gamma fn $\varphi(\alpha) \equiv \psi'(\alpha)$.

$\therefore$ If α known, $T(\underline{X})$ estimates β, then $Var_\beta T(\underline{X}) \geq \frac{\beta^2}{n\alpha}$

$\overline{X}$ is linearly dependent on $U_\beta \equiv \frac{\partial}{\partial\beta} \log p(\underline{X}, \alpha, \beta)$

$\because$ $I^{-1} = \frac{\beta^2}{\alpha\psi(\alpha)-1} \begin{bmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \psi(\alpha) \end{bmatrix}$

Can see $\frac{\alpha^2}{n\alpha} < \frac{\beta^2\psi(\alpha)}{\alpha\psi(\alpha)-1}$