MGE-201: Propiedades de la verosimilitud

Felipe Osorio

felipe.osorio@uv.cl

Definición 1 (Función de verosimilitud):

Para una observación $\boldsymbol{x}$ fijada de un vector aleatorio $\boldsymbol{X}$ con densidad $f(\cdot;\boldsymbol{\theta})$. La función de verosimilitud

$$L(\cdot;\boldsymbol{x}) : \Theta \to \mathbb{R}_+,$$

es definida como

$$L(\boldsymbol{\theta};\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

Observación:

La verosimilitud corresponde a la densidad conjunta de los datos que se desea analizar.

Ejemplo:

Sea $X_1, \dots, X_n$ variables aleatorias IID con distribución $\mathsf{N}(\theta, 1)$. Entonces,

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^n (2\pi)^{-1/2} \exp\big\{ - \tfrac{1}{2}(x_i - \theta)^2 \big\} = (2\pi)^{-n/2} \exp\Big\{ - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \Big\}$$
$$\propto \exp\Big\{ - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \Big\}.$$

Ahora,

$$\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n (x_i - \overline{x} + \overline{x} - \theta)^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - \theta)^2,$$

pues $\sum_{i=1}^n (x_i - \overline{x})(\overline{x} - \theta) = 0$. De este modo,

$$L(\theta; \boldsymbol{x}) \propto \exp\Big\{ - \frac{1}{2} \Big[\sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - \theta)^2 \Big] \Big\}$$
$$\propto \exp\Big\{ - \frac{n}{2} (\overline{x} - \theta)^2 \Big\}.$$

Es decir, $L(\theta; \boldsymbol{x})$ es proporcional a una densidad $\mathsf{N}_1(\overline{x}, 1/n)$.

Función de verosimilitud

Observación:

Es conveniente usar la función de log-verosimilitud dada por

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log L(\boldsymbol{\theta}; \boldsymbol{x}) = \log f(\boldsymbol{x}; \boldsymbol{\theta}).$$

Ejemplo:

Sea $X \sim \mathsf{N}(\theta, 1)$. Entonces,

$$\ell(\theta; X) = -\frac{1}{2}\log 2\pi - \frac{1}{2}(X - \theta)^2,$$

de este modo, la variable aleatoria

$$-2\ell(\theta; X) - \log 2\pi = (X - \theta)^2 \sim \chi^2(1).$$

Función de verosimilitud

Considere $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, vectores aleatorios IID. Entonces la verosimilitud de los datos combinados es:

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i; \boldsymbol{\theta}), \qquad \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}).$$

Ejemplo:

Sea $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ vectores aleatorios independientes $\mathsf{N}_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i(\boldsymbol{\phi}))$ con $\boldsymbol{X}_i$ matriz $m_i \times p$. Así,

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\boldsymbol{y}_i; \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^n \log|2\pi\boldsymbol{\Sigma}_i(\boldsymbol{\phi})| - \frac{1}{2}\sum_{i=1}^n D_i^2(\boldsymbol{\theta}),$$

con

$$D_i^2(\boldsymbol{\theta}) = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\phi})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

para $i = 1, \dots, n$, con $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top)^\top$.

Lema 1 (Principio de verosimilitud):

Sea $\boldsymbol{\theta}_0$ el parámetro subyacente (o verdadero). Entonces,

$$\mathsf{E}_{\theta_0}\{\ell(\boldsymbol{\theta};\boldsymbol{X})\} \leq \mathsf{E}_{\theta_0}\{\ell(\boldsymbol{\theta}_0;\boldsymbol{X})\}, \qquad \forall\, \boldsymbol{\theta}_0, \boldsymbol{\theta} \in \Theta.$$

Demostración:

En efecto,[1]

$$\begin{aligned}
\mathsf{E}_{\theta_0}\{\log f(\boldsymbol{x};\boldsymbol{\theta}_0)\} - \mathsf{E}_{\theta_0}\{\log f(\boldsymbol{x};\boldsymbol{\theta})\} &= -\,\mathsf{E}_{\theta_0}\left\{\log\frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{x};\boldsymbol{\theta}_0)}\right\} \\
&\geq -\log \mathsf{E}_{\theta_0}\left\{\frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{x};\boldsymbol{\theta}_0)}\right\} = \log\int \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{x};\boldsymbol{\theta}_0)}\, f(\boldsymbol{x};\boldsymbol{\theta}_0)\,\mathrm{d}\boldsymbol{x} \\
&= \log\int f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = \log 1 = 0.
\end{aligned}$$

[1] Recuerde que, para h convexa $\mathsf{E}(h(Y)) \geq h(\mathsf{E}(Y))$.

Supuesto A1:

Las distribuciones $\{\mathsf{P}_\theta : \boldsymbol{\theta} \in \Theta\}$ tienen soporte común, de modo que el conjunto

$$A = \{\boldsymbol{x} : f(\boldsymbol{x}; \boldsymbol{\theta}) \geq 0\},$$

no depende de $\boldsymbol{\theta}$.

Contraejemplos:

Considere $X \sim \mathsf{U}(0, \theta)$, con $\theta \in (0, \infty)$ cuya densidad es

$$f(x; \theta) = \frac{1}{\theta} I_{[0,\theta]}(x).$$

También la familia de distribuciones exponencial con dos parámetros $Y \sim \mathsf{Exp}(a, b)$,

$$f(y; a, b) = \frac{1}{b} \exp\Big(-\frac{(x-a)}{b}\Big) I_{[a,\infty)}(y), \qquad a, b > 0.$$

Supuesto A2:

El espacio paramétrico $\Theta \subset \mathbb{R}^p$ es un conjunto abierto.

Supuesto A3:

Para todo $\boldsymbol{x} \in A$, la función de log-verosimilitud es 3-veces continuamente diferenciable con respecto a $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$.

Definición 2 (Función score):

Suponga las condiciones A1 a A3 para todo $\boldsymbol{x} \in A$, se define la función score como el vector de derivadas parciales de la log-verosimilitud

$$\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}) = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta}} = \Big(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_p}\Big)^\top.$$

Supuesto A4:

Suponga que existen funciones integrables $F_1(x)$, $F_2(x)$ y $H(x)$ tal que

$$\int H(\boldsymbol{x}) f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} < M,$$

para $M \in \mathbb{R}$ un valor apropiado y que se satisface

$$\left|\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i}\right| < F_1(\boldsymbol{x}), \qquad \left|\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right| < F_2(\boldsymbol{x}),$$

$$\left|\frac{\partial^3 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < H(\boldsymbol{x}), \qquad i, j, k = 1, \dots, p.$$

Esta condición implica que podemos intercambiar las operaciones de integración y diferenciación. Por ejemplo,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_A f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = \int_A \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x}.$$

Resultado 1:

Bajo las condiciones A1 a A4, tenemos

$$\mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} = \mathbf{0}, \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

Demostración:

Considere,

$$\begin{aligned}
\mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} &= \int_A \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \int_A \frac{1}{f(\boldsymbol{x};\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\, f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \int_A \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} = \frac{\partial}{\partial \boldsymbol{\theta}} \int_A f(\boldsymbol{x};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \\
&= \mathbf{0}
\end{aligned}$$

Ejemplo:

Considere $X \sim \mathsf{N}(\theta, 1)$. Entonces,

$$U(\theta; X) = \frac{\partial}{\partial\theta}\Big[-\frac{1}{2}\log 2\pi - \frac{1}{2}(X-\theta)^2\Big] = X - \theta.$$

De este modo, $U(\theta; X) \sim \mathsf{N}(0, 1)$. Luego, es directo que

$$\mathsf{E}\{U(\theta; X)\} = \mathsf{E}(X - \theta) = 0.$$

Definición 3 (Matriz de información de Fisher):

Suponga las condiciones A1 a A3. Entonces la matriz de información de Fisher se define como:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{Cov}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\} = \mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{X})\boldsymbol{U}^\top(\boldsymbol{\theta};\boldsymbol{X})\}.$$

Es decir, $\boldsymbol{\mathcal{F}}(\boldsymbol{\theta})$ tiene elementos

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = \mathsf{E}_\theta\left\{\frac{\partial}{\partial\theta_i}\ell(\boldsymbol{\theta};\boldsymbol{X})\frac{\partial}{\partial\theta_j}\ell(\boldsymbol{\theta};\boldsymbol{X})\right\}.$$

Observación:

En ocasiones escribimos $\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta})$ pero la información no es aleatoria.

Ejemplo:

Sean $X_1, \dots, X_n$ variables aleatorias $\mathsf{N}(\mu, \sigma^2)$ con σ^2 conocido. Entonces,

$$L(\mu) \propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\Big\},$$

así

$$\ell(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + c,$$

con c una contante. Además,

$$U(\mu; \boldsymbol{X}) = \dot{\ell}(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu).$$

De este modo,[2]

$$\mathcal{F}_n(\mu) = \mathsf{var}\{U(\mu; \boldsymbol{X})\} = \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathsf{var}(X_i - \mu) = \frac{n}{\sigma^2}.$$

[2] Hay más información, conforme $\sigma^2 \to 0$ o $n \to \infty$.

Resultado 2:

Suponga las condiciones A1 a A4, Entonces

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{-\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{X})\} = \mathsf{E}_\theta\left\{ -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right\}.$$

Demostración:

Revisar página 32 de las notas de clase.[3]

Observación:

Este resultado permite obtener la matriz de información de Fisher de dos maneras equivalentes en modelos regulares (esto es, bajo los Supuestos A1 a A4). Es decir,

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{X})\boldsymbol{U}^\top(\boldsymbol{\theta}; \boldsymbol{X})\} = \mathsf{E}_\theta\left\{ -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right\}.$$

[3] Disponible en: `http://fosorios.mat.utfsm.cl/files/lectures/Inferencia.pdf`

Información observada

Definición 4 (Información observada):

La matriz

$$\boldsymbol{J}(\boldsymbol{\theta};\boldsymbol{X}) = -\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top},$$

se denomina información observada.

Observación:

En efecto,

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \mathsf{E}_\theta\{\boldsymbol{J}(\boldsymbol{\theta};\boldsymbol{X})\},$$

que también es llamada matriz de información esperada.

Ejemplo:

Considere $X \sim \mathsf{Bin}(n, \theta)$ con $\theta \in (0, 1)$. De este modo,

$$\ell(\theta; x) = \log \binom{n}{x} + (n - x) \log(1 - \theta) + x \log \theta,$$

así,

$$U(\theta; x) = -\frac{n - x}{1 - \theta} + \frac{x}{\theta},$$

obteniendo la derivada de $U(\theta; x)$,

$$U'(\theta; x) = -\frac{n - x}{(1 - \theta)^2} - \frac{x}{\theta^2}.$$

Además, como $\mathsf{E}(X) = n\theta$, sigue que

$$\begin{aligned} \mathcal{F}(\theta) &= \mathsf{E}_\theta\{-U'(\theta; X)\} = \frac{n - \mathsf{E}(X)}{(1 - \theta)^2} + \frac{\mathsf{E}(X)}{\theta^2} = \frac{n - n\theta}{(1 - \theta)^2} + \frac{n\theta}{\theta} \\ &= n\Big(\frac{1}{1 - \theta} + \frac{1}{\theta}\Big) = \frac{n}{\theta(1 - \theta)}. \end{aligned}$$

Ejemplo:

Sea $X \sim$ FE 1-paramétrica. En este caso, tenemos

$$\ell(\eta; x) = \log f(x; \eta) = \eta T(x) - b(\eta) + \log h(x),$$

De ahí que,

$$U(\eta; x) = T(x) - b'(\eta),$$

Usando el Resultado 2 visto en la Sesión 4,[4]

$$\mathsf{E}\{U(\eta; X)\} = \mathsf{E}\{T(X)\} - b'(\eta) = 0.$$

Además,

$$U'(\eta; x) = -\frac{\mathrm{d}^2\, b(\eta)}{\mathrm{d}\eta^2} = -b''(\eta),$$

obtenemos la información de Fisher:

$$\mathcal{F}(\eta) = \mathsf{E}\{-U'(\eta)\} = b''(\eta) = \mathsf{var}(T).$$

[4] URL: https://github.com/faosorios/Curso-Inferencia/blob/main/diapositivas/MAT206_slides-04.pdf

Resultado 3:

Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ vectores aleatorios independientes con informaciones de Fisher $\boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta})$ y $\boldsymbol{\mathcal{F}}_Y(\boldsymbol{\theta})$, respectivamente. La información de Fisher contenida en $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ es dada por

$$\boldsymbol{\mathcal{F}}_Z(\boldsymbol{\theta}) = \boldsymbol{\mathcal{F}}_X(\boldsymbol{\theta}) + \boldsymbol{\mathcal{F}}_Y(\boldsymbol{\theta}).$$

Corolario:

Sea $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ copias IID de una variable aleatoria X con información de Fisher (común) $\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta})$. Entonces, la información contenida en la muestra es:

$$\boldsymbol{\mathcal{F}}_n(\boldsymbol{\theta}) = n\boldsymbol{\mathcal{F}}_1(\boldsymbol{\theta}).$$

Observación:

Suponga $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ vectores aleatorios independientes[5]. Entonces,

$$\ell_n(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}; \boldsymbol{X}_i), \quad \boldsymbol{U}_n(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^n \boldsymbol{U}_i(\boldsymbol{\theta}; \boldsymbol{X}_i), \quad \boldsymbol{\mathcal{F}}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{\mathcal{F}}_i(\boldsymbol{\theta}),$$

con $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_n^\top)^\top$.

[5] no necesariamente copias

Resultado 4 (Dependencia de la parametrización):

Sea X una variable aleatoria con distribución P_θ. Suponga otra parametrización dada por $\theta = h(\phi)$ con h diferenciable. Entonces, la información de Fisher con relación a ϕ es dada por

$$\mathcal{F}_X^*(\phi) = \mathcal{F}_X(h(\phi))\{h'(\phi)\}^2,$$

donde $\mathcal{F}_X(\theta)$ es la información que $X \sim \mathsf{P}_\theta$ contiene respecto de θ y $\mathcal{F}_X^*(\phi)$ es la información que $X \sim \mathsf{P}_{h(\phi)}$ contiene sobre ϕ.

Demostración:

La función de log-verosimilitud con relación a $\mathsf{P}_{h(\phi)}$ es:

$$\ell^*(\phi; x) = \log f(x; h(\phi)).$$

De este modo, la función score asume la forma:

$$U^*(\phi; X) = \frac{\partial}{\partial\phi} \log f(x; h(\phi)) = \frac{\partial}{\partial\theta} \log f(x; \theta)\Big|_{\theta=h(\phi)} h'(\phi),$$

luego,

$$\begin{aligned}\mathcal{F}^*_X(\phi) &= \mathsf{var}_\phi\{U^*(\phi; X)\} = \mathsf{var}_\phi\{U(h(\phi); X)\, h'(\phi)\} \\ &= \{h'(\phi)\}^2\, \mathsf{var}_\phi\{U(h(\phi); X)\} = \{h'(\phi)\}^2 \mathcal{F}_X(h(\phi)).\end{aligned}$$

Ejemplo:

Considere la distribución Poi(θ). Así,

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

Note que el parámetro natural es $\eta = \log\theta$. Luego,

$$f(x;\eta) = \exp(x\eta - e^{\eta})/x!$$

En la parametrización original, tenemos que

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

De este modo,

$$\mathcal{F}_X(\theta) = \mathsf{var}_\theta\{U(\theta;X)\} = \mathsf{var}_\theta\left(\frac{X}{\theta} - 1\right) = \frac{1}{\theta^2}\,\mathsf{var}_\theta(X) = \frac{1}{\theta}.$$

Por otro lado, $\theta = e^{\eta} = h(\eta)$. Así,

$$\mathcal{F}^*_X(\eta) = \mathcal{F}_X(h(\eta))\{h'(\eta)\}^2 = \frac{1}{e^{\eta}}(e^{\eta})^2 = e^{\eta}.$$

Observación:

Considere $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\phi})$ con $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^p$ $(p \leq k)$ y sea $\boldsymbol{H}(\boldsymbol{\phi}) = \partial \boldsymbol{h}(\boldsymbol{\phi}) / \partial \boldsymbol{\phi}^\top$ matriz de rango completo. Entonces,

$$\boldsymbol{\mathcal{F}}_*(\boldsymbol{\phi}) = \boldsymbol{H}^\top(\boldsymbol{\phi}) \boldsymbol{\mathcal{F}}(\boldsymbol{h}(\boldsymbol{\phi})) \boldsymbol{H}(\boldsymbol{\phi}).$$