Máxima verosimilitud

Gabriel V. Montes-Rojas

Máxima verosimilitud: Introducción

- Máxima verosimilitud es un caso particular de un estimador M.
- En este caso se asume que **se conoce** la función de densidad condicional de las observaciones, y a partir de ello se construye un modelo paramétrico.
- Se maximiza la verosimilitud (likelihood) de una muestra $\{(y_i, \mathbf{x}_i) : i = 1, 2, ..., N\}$ para un modelo correctamente especificado de la densidad condicional $f(y|\mathbf{x}; \boldsymbol{\theta})$:

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \prod_{i=1}^{N} f(y_i|\mathbf{x}_i; \boldsymbol{\theta}).$$

- Si tomamos logaritmos, maximizar la verosimilitud es lo mismo que maximizar el logaritmo de la función de verosimilitud (log-likelihood). Sin embargo, simplifica mucho el problema dado que la suma es más simple que la multiplicación. Definamos $\ell(y_i|\mathbf{x}_i; \boldsymbol{\theta}) \equiv \ell_i(\boldsymbol{\theta}) = -q((y_i, \mathbf{x}_i), \boldsymbol{\theta})$ y $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$. Entonces,

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} log\ f(y_i|\mathbf{X}; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \ell(y_i|\mathbf{x}_i; \boldsymbol{\theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}).$$

Identificación

La identificación requiere que $\theta_0 \in \Theta$ *sea la única solución a*

$$\max_{\theta \in \Theta} E[\ell_i(\theta)],$$

o

$$E[\ell_i(\theta_0)|\mathbf{x}_i] \geq E[\ell_i(\theta)|\mathbf{x}_i], \theta \in \Theta.$$

Entonces,

$$\hat{\theta} = arg \max_{\theta \in \Theta} E[\ell_i(\theta)],$$

es el **estimador de máxima verosimilitud condicional** *(CMLE, conditional maximum likelihood estimator) de* θ*, donde la esperanza se toma con respecto a la distribución conjunta de* $(y_i, \mathbf{x}_i)$*.*

Consistencia

Debido a que máxima verosimilitud es un caso especial de estimadores M, la consistencia requiere los mismos elementos.

Consistencia de CMLE: *Sea* $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, 2, ...\}$ *una muestra aleatoria con* $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^K$, $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^G$. *Sea* $\boldsymbol{\Theta} \subset \mathbb{R}^P$ *el espacio de los parámetros y denotemos el modelo paramétrico de la densidad condicional como* $\{f(\cdot|\mathbf{x}, \boldsymbol{\theta}) : \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.
Asumamos que
(a) $\{f(\cdot|\mathbf{x}, \boldsymbol{\theta})$ *es la densidad verdadera con respecto a la medida* $v(d\mathbf{y})$ *para todo* $\mathbf{x}$ *y* $\boldsymbol{\theta}$;
(b) $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ *es laúnica solución de* $\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} E[\ell_i(\boldsymbol{\theta})|\mathbf{x}_i]$;
(c) $\boldsymbol{\Theta}$ *es compacto;*
(d) para cada $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\ell(., \boldsymbol{\theta})$ *is Borel measurable on* $\mathcal{Y} \times \mathcal{X}$;
(e) para cada $(\mathbf{x}, \mathbf{y}) \in \mathcal{Y} \times \mathcal{X}$, $\ell(\mathbf{x}, \mathbf{y}, .)$ *es continuo en* $\boldsymbol{\Theta}$; *y*
(d) $|\ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})| < b(\mathbf{x}, \mathbf{y})$ *para todo* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *donde b es una función no negativa en* $\mathcal{W}$ *tal que* $E[b(\mathbf{x}, \mathbf{y})] < \infty$.
Entonces existe una solución al problema de máxima verosimilitud y plim $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

Función score

- Para la normalidad asintótica se requiere ciertos supuestos sobre $\ell_i(\boldsymbol{\theta})$, en particular, se nececita $\boldsymbol{\theta}_0 \in int(\boldsymbol{\Theta})$ y que sea dos veces continuamente diferenciable.
- Entonces definamos el score para la observación i como el vector de derivadas parciales

$$\mathbf{s}_i(\boldsymbol{\theta}) \equiv \nabla_\theta \ell_i(\boldsymbol{\theta})' = \left(\frac{\partial \ell_i}{\theta_1}(\boldsymbol{\theta}), \frac{\partial \ell_i}{\theta_2}(\boldsymbol{\theta}), ..., \frac{\partial \ell_i}{\theta_P}(\boldsymbol{\theta}) \right)'$$

Igualdad de la matriz de información

La igualdad de la matriz de información (Information matrix equality) es

$$-E[\mathbf{H}_i(\theta_0)] = E[\mathbf{s}_i(\theta_0)\mathbf{s}_i(\theta_0)'].$$

Prueba: Para $\theta_0 \in int(\Theta)$, θ_0 satisface

$$\begin{aligned} E[\mathbf{s}_i(\theta_0)|\mathbf{x}_i] &= \mathbf{0} = \int_{\mathcal{Y}} \mathbf{s}_i(\theta_0) f(y|\mathbf{x_i};\theta_0) v(dy) \\ &= \int_{\mathcal{Y}} \frac{\partial \ell_i(\theta)}{\partial \theta}|_{\theta=\theta_0} f(y|\mathbf{x_i};\theta_0) v(dy) \\ &= \int_{\mathcal{Y}} \frac{\partial log\ f(y|\mathbf{x_i};\theta)}{\partial \theta}|_{\theta=\theta_0} f(y|\mathbf{x_i};\theta_0) v(dy) \\ &= \int_{\mathcal{Y}} \frac{\partial f(y|\mathbf{x_i};\theta)}{\partial \theta}|_{\theta=\theta_0} v(dy) \end{aligned}$$

Entonces usando $\nabla_\theta E[\mathbf{s}_i(\theta_0)|\mathbf{x}_i] = \mathbf{0}$ y asumiendo que la derivada y la integral (esperanza) se pueden intercambiar, tenemos

$$\int_{\mathcal{Y}} \frac{\partial^2 f(y|\mathbf{x_i};\theta)}{\partial \theta \partial \theta'}|_{\theta=\theta_0} v(dy) = \mathbf{0}.$$

Igualdad de la matriz de información

Prueba: (cont.) Entonces,

$$\begin{aligned}
\frac{\partial^2 log\ f(y|\mathbf{x_i};\theta)}{\partial\theta\partial\theta'} &= \frac{\partial \frac{\partial log\ f(y|\mathbf{x_i};\theta)}{\partial\theta}}{\partial\theta'} \\
&= \frac{\partial\left(\frac{\partial f(y|\mathbf{x_i};\theta)}{\partial\theta} f^{-1}(y|\mathbf{x_i};\theta)\right)}{\partial\theta'} \\
&= \frac{\partial^2 f(y|\mathbf{x_i};\theta)}{\partial\theta\partial\theta'} f^{-1}(y|\mathbf{x}_i;\theta) - \frac{\partial f(y|\mathbf{x_i};\theta)}{\partial\theta} f^{-2}(y|\mathbf{x}_i;\theta)\frac{\partial f(y|\mathbf{x_i};\theta)}{\partial\theta'} \\
&= \frac{\partial^2 f(y|\mathbf{x_i};\theta)}{\partial\theta\partial\theta'} f^{-1}(y|\mathbf{x}_i;\theta) - \frac{\partial log\ f(y|\mathbf{x_i};\theta)}{\partial\theta}\frac{\partial log\ f(y|\mathbf{x_i};\theta)}{\partial\theta'} \\
&= \frac{\partial^2 f(y|\mathbf{x_i};\theta)}{\partial\theta\partial\theta'} f^{-1}(y|\mathbf{x}_i;\theta) - \mathbf{s}_i(\theta)\mathbf{s}_i(\theta)'
\end{aligned}$$

Notemos que $\mathbf{H}_i(\theta) \equiv \nabla_\theta \mathbf{s}_i(\theta) = \nabla^2_\theta \ell_i(\theta)$. Entonces aplicando esperanzas a la última igualdad usando $\theta = \theta_0$

$$-E\left[\frac{\partial^2 log\ f(y|\mathbf{x_i};\theta)}{\partial\theta\partial\theta'}\right]_{\theta=\theta_0} = -E[\mathbf{H}_i(\theta_0)] = E[\mathbf{s}_i(\theta_0)\mathbf{s}_i(\theta_0)']$$

Igualdad de la matriz de información generalizada

Sea $\mathbf{g}(\mathbf{w}, \boldsymbol{\theta})$ *un vector* $Q \times 1$ *de mapeo* $\mathcal{W} \times \boldsymbol{\Theta} \mapsto \mathbf{g}(\mathbf{w}, \boldsymbol{\theta})$ *y asumimos que* $\mathbf{g}(\mathbf{w}, \boldsymbol{\theta}_0)$ *es diferenciable para* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. *Sea* $\mathbf{g}_i(\boldsymbol{\theta}) \equiv \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta})$. *Finalmente, asumamos que* $E_\theta[\nabla_\theta \mathbf{g}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$. *Igualdad de la matriz de información generalizada es*

$$-E[\nabla_\theta \mathbf{g}_i(\boldsymbol{\theta}_0)] = E[\mathbf{g}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'].$$

Prueba: Consideremos

$$E_\theta[\mathbf{g}_i(\boldsymbol{\theta})] = \int_{\mathcal{W}} \mathbf{g}(\mathbf{w}, \boldsymbol{\theta}) f(\mathbf{w}) v(d\mathbf{w}) = \mathbf{0}$$

para $\boldsymbol{\theta}$. Tomando derivadas con rspecto a $\boldsymbol{\theta}$ y asumiendo que derivada y esperanza se pueden intercambiar,

$$\int_{\mathcal{W}} \nabla_\theta \mathbf{g}(\mathbf{w}, \boldsymbol{\theta}) f(\mathbf{w}) v(d\mathbf{w}) + \int_{\mathcal{W}} \mathbf{g}(\mathbf{w}, \boldsymbol{\theta}) \nabla_\theta f(\mathbf{w}) v(d\mathbf{w}) = \mathbf{0},$$

entonces

$$E_\theta[\nabla_\theta \mathbf{g}_i(\boldsymbol{\theta})] + E_\theta[\mathbf{g}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})'] = \mathbf{0}.$$

porque $\nabla_\theta f(\mathbf{w}) = \mathbf{s}(\mathbf{w}, \boldsymbol{\theta})' f(\mathbf{w})$. Reemplazando $\boldsymbol{\theta}_0$ obtenemos el resultado.

Normalidad asintótica

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}_0^{-1}),$$

donde $\mathbf{A}_0 = \mathbf{B}_0$ *y* $\mathbf{A}_0 = -E[\mathbf{H}_i(\theta_0)]$.

Eficiencia

- Para estimadores asintóticamente normales, el "mejor" estimador es aquel consistente y con menor varianza. La eficiencia de un estimador se refiere a la varianza.
- Para cualquier estimador M

$$Avar(\hat{\theta}) = \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}.$$

- Sin embargo, sólo para MLE $Avar(\hat{\theta}) = \mathbf{A}_0^{-1}$. Entonces MLE es eficiente.
- Cràmer-Rao Lower Bound: Sea $\hat{\theta}$ un estimador *insesgado* para θ_0. Entonces,

$$Var(\hat{\theta}) - (\mathcal{I}_i(\theta_0))^{-1}$$

es definida semipositiva (para el caso unidimensional $Var(\hat{\theta}) \geq (\mathcal{I}_i(\theta_0))^{-1}$) donde $\mathcal{I}_i(\theta_0) \equiv E[\mathbf{s}_i(\theta_0)\mathbf{s}_i(\theta_0)']$ es la matriz de información de Fisher. Ésta matriz es una forma de medir la cantidad de observación de una variable aleatoria **w** contiene sobre el parámetro θ_0 sobre el cual la probbilidad de **w** depende.
El requerimiento de insesgadez implica que se pueden obtener mejores estimadores, pero no insesgados.

Ejemplo MCO

- Supongamos que $f(y_i|\mathbf{x_i};\boldsymbol{\beta},\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{2\sigma^2}\right)$. Sea $\boldsymbol{\theta} = (\boldsymbol{\beta},\sigma)$.
- Entonces (usamos C para una constante que no tiene parámetros)

$$\ell_i(\boldsymbol{\theta}) = C - \ln(\sigma) - \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{2\sigma^2}, \mathbf{s}_i(\boldsymbol{\theta}) = \begin{pmatrix} \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})}{\sigma^2}\mathbf{x_i'} \\ -\sigma^{-1} + \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{\sigma^3} \end{pmatrix},$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{1}{\sigma^2}\mathbf{x_i'x_i} & -2\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})}{\sigma^3}\mathbf{x_i'} \\ . & \sigma^{-2} - 3\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{\sigma^4} \end{pmatrix},$$

$$\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})' = \begin{pmatrix} \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{\sigma^4}\mathbf{x_i'x_i} & -\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})}{\sigma^3}\mathbf{x_i'} + \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^3}{\sigma^5}\mathbf{x_i'} \\ . & \sigma^{-2} + \frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^4}{\sigma^6} - 2\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{\sigma^4} \end{pmatrix},$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\ell_i(\boldsymbol{\theta}) = NC - N\ln(\sigma) - \sum_{i=1}^{N}\frac{(y_i-\mathbf{x_i}\boldsymbol{\beta})^2}{2\sigma^2},$$

Ejemplo MCO

- Probar que

$$E[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'] = -E[\mathbf{H}_i(\theta_0)]$$

  Ayuda: Para $u \sim N(0, \sigma^2)$, $E(u^2) = \sigma^2$, $E(u^3) = 0$ y $E(u^4) = 3\sigma^4$.
- Probar que $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{MLE}$.
  Ayuda: $\hat{\sigma}^2_{MLE} = N^{-1}\sum_{i=1}^{N}(y_i - \mathbf{x_i}\hat{\boldsymbol{\beta}})^2$.

Referencias

Estas notas están basadas en

- *Capítulos 12 y 13 de Wooldridge.*
- *Newey, W.K., y McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," en Handbook of Econometrics, Volumen 4, ed. R.F. Engle y D. McFadden. Amsterdam: North Holland, 2111–2245.*
- *Van der Vaart, A.W. (1998), Asymptotic Statistics. Cambridge University Press.*