MGE-201: Funciones de inferencia

Felipe Osorio

felipe.osorio@uv.cl

Configuración:

Vamos a suponer que el vector aleatorio $\boldsymbol{Y}$ puede ser escrito como

$$\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \dots, \boldsymbol{Y}_n^\top)^\top,$$

donde los $\boldsymbol{Y}_i$'s son independientes. Asumiremos también que $\boldsymbol{Y}$ sigue un modelo estadístico

$$\mathcal{P} = \{\mathsf{P}_\beta : \boldsymbol{\beta} \in \mathcal{B}\}, \qquad \mathcal{B} \in \mathbb{R}^k,$$

y $\mathcal{Y}$ es el espacio muestral.

Definición 1:

Una función $\boldsymbol{g}_n : \mathcal{B} \times \mathcal{Y} \to \mathbb{R}^k$ tal que $\boldsymbol{g}_n(\boldsymbol{\beta}; \cdot)$ es medible para todo $\boldsymbol{\beta} \in \mathcal{B}$ se dice una función de estimación.

Observación:

Para una función de inferencia $\boldsymbol{g}_n$ y una muestra $\boldsymbol{Y} \in \mathcal{Y}$ dadas, es posible obtener un estimador $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{Y})$ como solución de la ecuación

$$\boldsymbol{g}_n(\boldsymbol{\beta}; \boldsymbol{Y}) = \boldsymbol{0}.$$

Ejemplo (Regresión lineal):

Sea $Y_1, \dots, Y_n$ variables independientes tal que $\mathsf{E}(Y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, para $i = 1, \dots, n$, donde $\boldsymbol{x}_i$ es vector de covariables y $\boldsymbol{\beta}$ denota coeficientes de regresión. La ecuación de estimación asociada a OLS es:

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) = \boldsymbol{0}.$$

Note que

$$\mathsf{E}(\boldsymbol{g}_n(\boldsymbol{\beta})) = \mathbf{0}, \tag{1}$$

de ahí que (1) corresponde a una condición de momentos.

Resolviendo $\boldsymbol{g}_n(\boldsymbol{\beta}) = \mathbf{0}$ con relación a $\boldsymbol{\beta}$, obtenemos las ecuaciones

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i Y_i,$$

lo que lleva al estimador OLS,

$$\widehat{\boldsymbol{\beta}}_n = \Big(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i Y_i = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Es decir, OLS es un estimador de momentos.

Ejemplo (Regresión no-lineal):

Considere $Y_1, \dots, Y_n$ independientes con $\mathsf{E}(Y_i) = f(\boldsymbol{x}_i; \boldsymbol{\beta})$ donde $f(\cdot; \boldsymbol{\beta})$ es una función suave de $\boldsymbol{\beta}$. La función de estimación es dada por

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{\partial f(\boldsymbol{x}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)(Y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta})).$$

La función $\boldsymbol{g}_n(\boldsymbol{\beta})$ es fácil de estudiar debido a que es lineal en los residuos

$$r_i(\boldsymbol{\beta}) = Y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta}), \qquad i = 1, \dots, n.$$

Observación:

Considere $Y_1, \dots, Y_n$ variables aleatorias independientes con log-verosimilitud conjunta

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log f(y_i; \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta}).$$

La condición de primer orden adopta la forma:

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\theta=\widehat{\theta}_{\mathsf{ML}}} = \mathbf{0},$$

permite notar que los estimadores ML tienen una interpretación de momentos.[1]

[1] En efecto, $\mathsf{E}(\boldsymbol{U}_n(\boldsymbol{\theta})) = \mathbf{0}$.

Definición 2 (Vector de residuos):

Un vector aleatorio m-dimensional $\boldsymbol{r}(\boldsymbol{\beta};\boldsymbol{Y})$ es llamado un vector de residuos si:

(i) $\boldsymbol{r}$ es insesgado, es decir,

$$\mathsf{E}_{\beta}\{\boldsymbol{r}(\boldsymbol{\beta};\boldsymbol{Y})\} = \boldsymbol{0}.$$

(ii) $\boldsymbol{r}$ tiene matriz de variabilidad

$$\boldsymbol{V}_r(\boldsymbol{\beta}) = \mathsf{Cov}(\boldsymbol{r}(\boldsymbol{\beta};\boldsymbol{Y}))$$

Observación:

- Un vector $\boldsymbol{r}(\boldsymbol{\beta})$ se dice no-singular si $\boldsymbol{V}_r(\boldsymbol{\beta})$ es definida positiva para todo $\boldsymbol{\beta}$.
- Note el énfasis que se hace sobre supuestos de momentos.

Un vector $\boldsymbol{r}(\boldsymbol{\beta})$ es llamado suave si $\boldsymbol{r}(\cdot;\boldsymbol{Y})$ es diferenciable para casi todo $\boldsymbol{Y}$, y la matriz de sensibilidad $m \times k$

$$\boldsymbol{S}_r(\boldsymbol{\beta}) = \mathsf{E}_\beta\left\{\frac{\partial \boldsymbol{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\},$$

existe para todo $\boldsymbol{\beta}$.

Definición 3 (Información de Godambe):

Para un vector residual no-singular y suave, $\boldsymbol{r}(\boldsymbol{\beta})$ se define la matriz de información de Godambe[2] como:

$$\boldsymbol{G}_r(\boldsymbol{\beta}) = \boldsymbol{S}_r^\top(\boldsymbol{\beta})\boldsymbol{V}_r^{-1}(\boldsymbol{\beta})\boldsymbol{S}_r(\boldsymbol{\beta}).$$

[2] Note la forma de sandwich

Sea $\boldsymbol{A}(\boldsymbol{\beta})$ matriz $p \times m$ no estocástica y considere

$$\boldsymbol{q}(\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{r}(\boldsymbol{\beta}).$$

De ahí que

$$\boldsymbol{V}_q(\boldsymbol{\beta}) = \mathsf{Cov}(\boldsymbol{q}(\boldsymbol{\beta})) = \boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{V}_r(\boldsymbol{\beta})\boldsymbol{A}^\top(\boldsymbol{\beta}).$$

Resultado 1:

Para $\boldsymbol{A}(\boldsymbol{\beta})$ matriz $p \times m$ no estocástica y $\boldsymbol{r}(\boldsymbol{\beta})$ regular. Entonces

$$\boldsymbol{S}_q(\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{S}_r(\boldsymbol{\beta}).$$

Demostración:

En efecto,

$$\begin{aligned}
\boldsymbol{S}_q(\boldsymbol{\beta}) &= \mathsf{E}\{\nabla_\beta \boldsymbol{q}(\boldsymbol{\beta})\} = \mathsf{E}\{[\nabla_\beta \boldsymbol{A}(\boldsymbol{\beta})]\boldsymbol{r}(\boldsymbol{\beta}) + \boldsymbol{A}(\boldsymbol{\beta})\nabla_\beta \boldsymbol{r}(\boldsymbol{\beta})\} \\
&= [\nabla_\beta \boldsymbol{A}(\boldsymbol{\beta})]\,\mathsf{E}\{\boldsymbol{r}(\boldsymbol{\beta})\} + \boldsymbol{A}(\boldsymbol{\beta})\,\mathsf{E}\{\nabla_\beta \boldsymbol{r}(\boldsymbol{\beta})\} \\
&= \boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{S}_r(\boldsymbol{\beta})
\end{aligned}$$

Considere $\boldsymbol{r}(\boldsymbol{\beta})$ y $\boldsymbol{s}(\boldsymbol{\beta})$ vectores residuales m-dimensionales. Entonces,

$$\boldsymbol{S}_{r+s}(\boldsymbol{\beta}) = \boldsymbol{S}_r(\boldsymbol{\beta}) + \boldsymbol{S}_s(\boldsymbol{\beta}).$$

La linealidad de $\boldsymbol{S}(\boldsymbol{\beta})$ es evidentemente, pues

$$\nabla_\beta[\boldsymbol{r}(\boldsymbol{\beta}) + \boldsymbol{s}(\boldsymbol{\beta})] = \nabla_\beta \boldsymbol{r}(\boldsymbol{\beta}) + \nabla_\beta \boldsymbol{s}(\boldsymbol{\beta}).$$

Además, si $\boldsymbol{r}(\boldsymbol{\beta})$ y $\boldsymbol{s}(\boldsymbol{\beta})$ son independientes. Entonces,

$$\boldsymbol{V}_{r+s}(\boldsymbol{\beta}) = \boldsymbol{V}_r(\boldsymbol{\beta}) + \boldsymbol{V}_s(\boldsymbol{\beta}).$$

Considere una re-parametrización $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\gamma})$, donde $\boldsymbol{\gamma}$ es un vector l-dimensional $(l \leq k)$, y sea

$$\boldsymbol{q}(\boldsymbol{\gamma}) = \boldsymbol{r}(\boldsymbol{\beta}(\boldsymbol{\gamma}))$$

Entonces,

$$\boldsymbol{V}_q(\boldsymbol{\gamma}) = \boldsymbol{V}_r(\boldsymbol{\beta}(\boldsymbol{\gamma})), \qquad \boldsymbol{S}_q(\boldsymbol{\gamma}) = \boldsymbol{S}_r(\boldsymbol{\beta}(\boldsymbol{\gamma}))\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma}).$$

En efecto

$$\begin{aligned}\boldsymbol{S}_q(\boldsymbol{\gamma}) &= \mathsf{E}_\gamma\{\nabla_\gamma \boldsymbol{q}(\boldsymbol{\gamma})\} = \mathsf{E}_\gamma\{\nabla_\gamma \boldsymbol{r}(\boldsymbol{\beta}(\boldsymbol{\gamma}))\} = \mathsf{E}_\gamma\{\nabla_\beta \boldsymbol{r}(\boldsymbol{\beta}(\boldsymbol{\gamma}))\,\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma})\}\\ &= \mathsf{E}_\gamma\{\nabla_\beta \boldsymbol{r}(\boldsymbol{\beta})\}\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma}) = \boldsymbol{S}_r(\boldsymbol{\beta}(\boldsymbol{\gamma}))\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma})\end{aligned}$$

Adicionalmente, si $\boldsymbol{r}$ es no-singular, entonces

$$\begin{aligned}\boldsymbol{G}_q(\boldsymbol{\gamma}) &= \boldsymbol{S}_q^\top(\boldsymbol{\gamma})\boldsymbol{V}_q^{-1}(\boldsymbol{\gamma})\boldsymbol{S}_q(\boldsymbol{\gamma})\\ &= \nabla_\gamma^\top \boldsymbol{\beta}(\boldsymbol{\gamma})\boldsymbol{S}_r^\top(\boldsymbol{\gamma})\boldsymbol{V}_r^{-1}(\boldsymbol{\gamma})\boldsymbol{S}_r(\boldsymbol{\beta}(\boldsymbol{\gamma}))\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma})\\ &= \nabla_\gamma^\top \boldsymbol{\beta}(\boldsymbol{\gamma})\boldsymbol{G}_r(\boldsymbol{\beta}(\boldsymbol{\gamma}))\nabla_\gamma \boldsymbol{\beta}(\boldsymbol{\gamma})\end{aligned}$$

Definición 4:

Un vector residual suave m-dimensional $\boldsymbol{r}(\boldsymbol{\beta})$ es dicho regular, si $k \leq m$ y la matriz de sensibilidad $\boldsymbol{S}_r(\boldsymbol{\beta})$ tiene rango k para todo $\boldsymbol{\beta}$.

Resultado 2:

Un vector residual regular $\boldsymbol{r}(\boldsymbol{\beta})$ es no-singular.

Demostración:

(Por contradicción) Suponga que $\boldsymbol{r}(\boldsymbol{\beta})$ es singular, en cuyo caso $\boldsymbol{V}_r(\boldsymbol{\beta}) = \mathrm{Cov}(\boldsymbol{r}(\boldsymbol{\beta}))$ no es definida positiva. Entonces, existe un vector no nulo $\boldsymbol{c}$ tal que la combinación $\boldsymbol{c}^\top \boldsymbol{r}(\boldsymbol{\beta})$ es cero casi seguramente. De ahí que

$$\boldsymbol{0} = \boldsymbol{S}_{c^\top r}(\boldsymbol{\beta}) = \boldsymbol{c}^\top \boldsymbol{S}_r(\boldsymbol{\beta}),$$

y por tanto, $\boldsymbol{S}_r(\boldsymbol{\beta})$ es singular, lo que contradice que $\boldsymbol{r}(\boldsymbol{\beta})$ sea regular. De ahí que $\boldsymbol{V}_r(\boldsymbol{\beta})$ debe ser definida positiva.

Para un vector residual regular $\boldsymbol{r}(\boldsymbol{\beta})$, la matriz de información de Godambe

$$\boldsymbol{G}_r(\boldsymbol{\beta}) = \boldsymbol{S}_r^\top(\boldsymbol{\beta})\boldsymbol{V}_r^{-1}(\boldsymbol{\beta})\boldsymbol{S}_r(\boldsymbol{\beta}),$$

es definida positiva.

Para función de estimación regular $\boldsymbol{g}_n : \mathbb{R}^k \to \mathbb{R}^k$, la matriz de sensibilidad será cuadrada aunque no necesariamente simétrica.

Suponga $\boldsymbol{q}_n(\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta})$ con $\boldsymbol{A}(\boldsymbol{\beta})$ matriz no singular $k \times k$. Entonces

$$\begin{aligned}\boldsymbol{G}_q(\boldsymbol{\beta}) &= \boldsymbol{S}_g^\top(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})\{\boldsymbol{A}^{-\top}(\boldsymbol{\beta})\boldsymbol{V}_g^{-1}(\boldsymbol{\beta})\boldsymbol{A}^{-1}(\boldsymbol{\beta})\}\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{S}_g(\boldsymbol{\beta}) \\ &\boldsymbol{S}_g^\top(\boldsymbol{\beta})\boldsymbol{V}_g^{-1}(\boldsymbol{\beta})\boldsymbol{S}_g(\boldsymbol{\beta}) = \boldsymbol{G}_g(\boldsymbol{\beta}),\end{aligned}$$

de ahí que $\boldsymbol{g}$ y $\boldsymbol{q}$ son equivalentes.

Definición 5 (Algoritmo Newton-scoring):

Considere la expansión en series de Taylor de $\boldsymbol{g}(\boldsymbol{\beta})$ en torno de $\boldsymbol{\beta}_0$, tenemos

$$\boldsymbol{g}(\boldsymbol{\beta}) \approx \boldsymbol{g}(\boldsymbol{\beta}_0) + \frac{\partial \boldsymbol{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\Big|_{\beta=\beta_0}(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

como $\boldsymbol{g}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$ y substituyendo $\dot{\boldsymbol{g}}(\boldsymbol{\beta}_0)$ por $\boldsymbol{S}_g(\boldsymbol{\beta}_0)$, sigue que

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 - \boldsymbol{S}_g^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}(\boldsymbol{\beta}_0), \qquad (2)$$

esto sugiere considerar:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{S}_g^{-1}(\boldsymbol{\beta}^{(t)})\boldsymbol{g}(\boldsymbol{\beta}^{(t)}),$$

para llevar a cabo la estimación de parámetros.

Observación:

Este procedimiento fue propuesto por Jørgensen y Knudsen (2004)[3] quienes lo denominaron algoritmo Newton-scoring.

[3] Scandinavian Journal of Statistics **31**, 93-114.

Ejemplo:

Considere la función de estimación de minimos cuadrados no lineales

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{\beta})(Y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta})),$$

con $\boldsymbol{f}_i(\boldsymbol{\beta}) = \partial f(\boldsymbol{x}_i; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. En efecto, $\mathsf{E}\{\boldsymbol{g}_n(\boldsymbol{\beta})\} = \mathbf{0}$, y

$$\begin{aligned} \boldsymbol{V}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{\beta}) \operatorname{var}(Y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta})) \boldsymbol{f}_i^\top(\boldsymbol{\beta}) = \sigma^2 \sum_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{\beta}) \boldsymbol{f}_i^\top(\boldsymbol{\beta}) \\ &= \sigma^2 \boldsymbol{F}^\top \boldsymbol{F}, \end{aligned}$$

donde $\boldsymbol{F} = (\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)^\top$. Mientras que

$$\boldsymbol{S}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{\beta}) \boldsymbol{f}_i^\top(\boldsymbol{\beta}).$$

De este modo,

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{F}^\top \boldsymbol{F}.$$

Definición 6:

Sea $\boldsymbol{r}(\boldsymbol{\beta})$ vector de residuos no singular. Defina la función quasi-score como:

$$\boldsymbol{q}(\boldsymbol{\beta}) = -\boldsymbol{S}_r^\top(\boldsymbol{\beta})\boldsymbol{V}_r^{-1}(\boldsymbol{\beta})\boldsymbol{r}(\boldsymbol{\beta}),$$

donde $\boldsymbol{S}_r$ y $\boldsymbol{V}_r$ son las matrices de sensibilidad y variabilidad de $\boldsymbol{r}(\boldsymbol{\beta})$.

La matriz $k \times m$,

$$\boldsymbol{A}(\boldsymbol{\beta}) = -\boldsymbol{S}_r^\top(\boldsymbol{\beta})\boldsymbol{V}_r^{-1}(\boldsymbol{\beta}),$$

es llamada matriz de pesos de Crowder (1987).[4] Además, las matrices de sensibilidad y variabilidad de $\boldsymbol{q}(\boldsymbol{\beta})$ son

$$\boldsymbol{S}_q = -\boldsymbol{S}_r^\top\boldsymbol{V}_r^{-1}\boldsymbol{S}_r = -\boldsymbol{G}_r$$
$$\boldsymbol{V}_q = \boldsymbol{S}_r^\top\boldsymbol{V}_r^{-1}\boldsymbol{V}_r\boldsymbol{V}_r^{-1}\boldsymbol{S}_r = \boldsymbol{G}_r.$$

Es decir, $\boldsymbol{V}_q = -\boldsymbol{S}_q$, y de ahí que

$$\boldsymbol{G}_q = \boldsymbol{G}_r.$$

[4] Biometrika **74**, 591-597.

Funciones de inferencia

Observación:

Cualquier función de estimación regular $\boldsymbol{g}(\boldsymbol{\beta})$ satisfaciendo

$$-\boldsymbol{S}_g(\boldsymbol{\beta}) = \boldsymbol{V}_g(\boldsymbol{\beta}),$$

es una función quasi-score.

En particular, la función score

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log p(\boldsymbol{\beta}),$$

es una función quasi-score.

Suponga $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ vectores aleatorios independientes y sea

$$\boldsymbol{r}_i(\boldsymbol{\mu}_i), \quad \boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\beta}),$$

vectores m_i-dimensionales. Considere la siguiente definición.

Definición 7:

Una función de estimación lineal es definida como:

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{W}_i(\boldsymbol{\beta})\boldsymbol{r}_i(\boldsymbol{\mu}_i),$$

donde $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\beta})$ es una matriz no aleatoria $k \times m_i$.

Ejemplo (mínimos cuadrados):

En este caso tenemos

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Ejemplo (M-estimación):

Considere Y_i con distribución simétrica, la función de estimación M es

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \psi(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

donde ψ es una función impar.

Definición 8:

Basado en la Ecuación (2) podemos definir una función de estimación normalizada como:

$$\overline{\boldsymbol{g}}(\boldsymbol{\beta}) = -\boldsymbol{S}_g^{-1}(\boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{\beta}).$$

Previo (Ordenamiento de Löwner):

Si $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices de covarianza de la misma dimensión. Entonces $\boldsymbol{A} \geq \boldsymbol{B}$ si y sólo si $\boldsymbol{A} - \boldsymbol{B}$ es semidefinida positiva.[5]

[5] Esto corresponde a un ordenamiento parcial en el espacio de matrices semidefinidas positivas.

Resultado 3:

Sea $\overline{\boldsymbol{g}}_1$ y $\overline{\boldsymbol{g}}_2$ dos funciones de estimación normalizadas con matrices de información de Godambe $\boldsymbol{K}_1$ y $\boldsymbol{K}_2$, respectivamente. La condición

$$\mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_2) = \boldsymbol{K}_1, \tag{3}$$

implica las siguientes tres condiciones:

(a) $\overline{\boldsymbol{g}}_1$ y $\overline{\boldsymbol{g}}_2 - \overline{\boldsymbol{g}}_1$ son no correlacionadas.

(b) $\boldsymbol{K}_2 \geq \boldsymbol{K}_1$.

(c) $\mathsf{var}(\boldsymbol{c}^\top \overline{\boldsymbol{g}}_2) \geq \mathsf{var}(\boldsymbol{c}^\top \overline{\boldsymbol{g}}_1)$ para cualquier vector $\boldsymbol{c} \in \mathbb{R}^k$.

Demostración:

Note que podemos escribir

$$\overline{\boldsymbol{g}}_2 = \overline{\boldsymbol{g}}_1 + (\overline{\boldsymbol{g}}_2 - \overline{\boldsymbol{g}}_1). \tag{4}$$

Si $\mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_2) = \boldsymbol{K}_1$ entonces

$$\begin{aligned}\mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_2 - \overline{\boldsymbol{g}}_1) &= \mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_2) - \mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_1) \\ &= \mathsf{Cov}(\overline{\boldsymbol{g}}_1, \overline{\boldsymbol{g}}_2) - \boldsymbol{K}_1 = \boldsymbol{0},\end{aligned}$$

lo que prueba (a). Tomando covarianzas en ambos lados de (4) lleva a

$$\boldsymbol{K}_2 = \boldsymbol{K}_1 + \boldsymbol{A}, \qquad \boldsymbol{A} = \mathsf{Cov}(\overline{\boldsymbol{g}}_2 - \overline{\boldsymbol{g}}_1),$$

como $\boldsymbol{A}$ es semidefinida positiva, sigue (b). Además, tenemos que para cualquier $\boldsymbol{c}$,

$$\begin{aligned}0 \leq \boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{c} &= \boldsymbol{c}^\top (\boldsymbol{K}_2 - \boldsymbol{K}_1)\boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{K}_2 \boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{K}_1 \boldsymbol{c} \\ &= \mathsf{var}(\boldsymbol{c}^\top \overline{\boldsymbol{g}}_2) - \mathsf{var}(\boldsymbol{c}^\top \overline{\boldsymbol{g}}_1),\end{aligned}$$

lo que implica (c).

Funciones de inferencia

Suponga que tenemos una función de estimación lineal

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{W}_i(\boldsymbol{\beta})\boldsymbol{r}_i(\boldsymbol{\mu}_i).$$

De este modo,

$$\boldsymbol{S}_g = \sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{S}_{r_i} = \sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{S}_i \boldsymbol{D}_i,$$

mientras que

$$\boldsymbol{V}_g = \sum_{i=1}^{n} \operatorname{Cov}\big(\boldsymbol{W}_i(\boldsymbol{\beta})\boldsymbol{r}_i(\boldsymbol{\mu}_i)\big) = \sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{V}_i \boldsymbol{W}_i^{\top}.$$

Es decir,

$$\boldsymbol{G}_g(\boldsymbol{\beta}) = \boldsymbol{S}_g^{\top}(\boldsymbol{\beta})\boldsymbol{V}_q^{-1}(\boldsymbol{\beta})\boldsymbol{S}_g(\boldsymbol{\beta}).$$

Resultado 4 (Optimalidad de Crowder):

Suponga $\boldsymbol{q}(\boldsymbol{\beta})$ función quasi-score

$$\boldsymbol{q}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i(\boldsymbol{\mu}_i), \qquad \boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^\top.$$

Entonces $\boldsymbol{q}(\boldsymbol{\beta})$ es óptima para todas las funciones de estimación lineales en el sentido que

$$\boldsymbol{G}_g^{-1}(\boldsymbol{\beta}) \geq \boldsymbol{G}_q^{-1}(\boldsymbol{\beta}),$$

para cualquier función de estimación lineal $\boldsymbol{g}(\boldsymbol{\beta})$, con la igualdad si y sólo si

$$\overline{\boldsymbol{g}}(\boldsymbol{\beta}) = \overline{\boldsymbol{q}}(\boldsymbol{\beta}),$$

esto es, si $\boldsymbol{g}(\boldsymbol{\beta})$ y $\boldsymbol{q}(\boldsymbol{\beta})$ son equivalentes.

Demostración:

Deseamos mostrar la Ecuación (3) del Resultado 3. En efecto

$$
\begin{aligned}
\operatorname{Cov}(\boldsymbol{q},\boldsymbol{g}) = \mathsf{E}(\boldsymbol{q}\boldsymbol{g}^\top) &= -\,\mathsf{E}\Big[\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i\Big(\sum_{j=1}^n \boldsymbol{W}_j\boldsymbol{r}_j\Big)\Big]\\
&= -\,\mathsf{E}\Big[\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i \sum_{j=1}^n \boldsymbol{r}_j^\top \boldsymbol{W}_j^\top\Big]\\
&= -\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1} \sum_{j=1}^n \mathsf{E}(\boldsymbol{r}_i\boldsymbol{r}_j^\top)\boldsymbol{W}_j^\top,
\end{aligned}
$$

como $\boldsymbol{r}_i$ y $\boldsymbol{r}_j$ son no correlacionados para $i \neq j$, sigue que

$$
\begin{aligned}
\mathsf{E}(\boldsymbol{q}\boldsymbol{g}^\top) &= -\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1}\,\mathsf{E}(\boldsymbol{r}_i\boldsymbol{r}_i^\top)\boldsymbol{W}_j^\top = -\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{V}_i\boldsymbol{W}_j^\top\\
&= -\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{S}_i^\top \boldsymbol{W}_i^\top = -\boldsymbol{S}_g^\top
\end{aligned}
$$

Esto permite calcular

$$\begin{aligned}\operatorname{Cov}(\overline{\boldsymbol{q}}, \overline{\boldsymbol{g}}) &= \mathsf{E}\{\boldsymbol{S}_q^{-1}\boldsymbol{q}(\boldsymbol{S}_g^{-1}\boldsymbol{g})^\top\} = \boldsymbol{S}_q^{-1}\,\mathsf{E}(\boldsymbol{q}\boldsymbol{g}^\top)\boldsymbol{S}_g^{-\top} \\ &= -\boldsymbol{S}_q^{-1}\boldsymbol{S}_g^\top\boldsymbol{S}_g^{-\top} = -\boldsymbol{S}_q^{-1} = \boldsymbol{G}_q^{-1},\end{aligned}$$

y esto implica la optimalidad de $\boldsymbol{q}$, lo que concluye la prueba.

Corolario (Desigualdad de Godambe):

Asuma que $\boldsymbol{g}$ es función de inferencia regular. Entonces

$$\boldsymbol{G}_g^{-1}(\boldsymbol{\theta}) - \boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0},$$

para todo $\boldsymbol{\theta} \in \Theta$, donde la igualdad se cumple si y sólo si $\boldsymbol{g}$ y $\boldsymbol{U}$ son funciones de estimación equivalentes, con $\boldsymbol{U}$ la función score.

Asuma una función de estimación lineal

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{r}_i(\boldsymbol{\mu}_i),$$

donde la notación enfatiza la dependencia de n. Asimismo $\boldsymbol{S}_n(\boldsymbol{\beta})$ y $\boldsymbol{V}_n(\boldsymbol{\beta})$ denotan las matrices de sensibilidad y variabilidad de $\boldsymbol{g}_n$.

Además, suponga que

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{S}_n(\boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{\beta}), \qquad \lim_{n\to\infty} \frac{1}{n}\boldsymbol{V}_n(\boldsymbol{\beta}) = \boldsymbol{V}(\boldsymbol{\beta}).$$

Resultado 5:

Si $\boldsymbol{S}(\boldsymbol{\beta})$ es no singular, entonces $\boldsymbol{V}(\boldsymbol{\beta})$ es no singular.

Demostración:

La prueba se lleva a cabo por contradicción. Suponga que $\boldsymbol{V}(\boldsymbol{\beta})$ es singular. Entonces existe un vector $\boldsymbol{c}$ tal que

$$\boldsymbol{c}^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{c} = 0,$$

y de ahí que

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{c}^\top \boldsymbol{V}_n(\boldsymbol{\beta})\boldsymbol{c} = 0,$$

lo cual implica que

$$\frac{1}{n}\boldsymbol{c}^\top \boldsymbol{g}_n(\boldsymbol{\beta}) \xrightarrow{\mathsf{p}} 0.$$

Asumiendo la continuidad del operador de sensibilidad $\boldsymbol{S}(\cdot)$, tenemos

$$\begin{aligned}\boldsymbol{0} &= \boldsymbol{S}\Big[\lim_{n\to\infty} \boldsymbol{c}^\top \boldsymbol{g}_n(\boldsymbol{\beta})\Big] = \lim_{n\to\infty} \boldsymbol{S}\Big[\frac{1}{n}\boldsymbol{c}^\top \boldsymbol{g}_n(\boldsymbol{\beta})\Big] \\ &= \boldsymbol{c}^\top \lim_{n\to\infty} \frac{1}{n}\boldsymbol{S}_n(\boldsymbol{\beta}) = \boldsymbol{c}^\top \boldsymbol{S}(\boldsymbol{\beta})\end{aligned}$$

lo que es una contradicción pues $\boldsymbol{S}(\boldsymbol{\beta})$ ha sido asumida no singular.

Observación:

Note que

$$\begin{aligned}\frac{1}{n}\boldsymbol{G}_n(\boldsymbol{\beta}) &= \frac{1}{n}\boldsymbol{S}_n^\top(\boldsymbol{\beta})\Big[\frac{1}{n}\boldsymbol{V}_n(\boldsymbol{\beta})\Big]^{-1}\frac{1}{n}\boldsymbol{S}_n(\boldsymbol{\beta}) \\ &\to \boldsymbol{S}^\top(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{G}(\boldsymbol{\beta}),\end{aligned}$$

es decir, la información de Godambe promedio converge a un límite finito.

Resultado 6:

Sea $\{\widehat{\boldsymbol{\theta}}_n\}_{n\geq 1}$ una secuencia de raíces de las ecuaciones de estimación

$$\boldsymbol{g}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\theta};\boldsymbol{Y}_i) = \boldsymbol{0},$$

desde la condición de insesgamiento se tiene la consistencia $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{p}} \boldsymbol{\theta}_0$.

Supuestos: (Yuan y Jennrich, 1998)[6]

C1: $\boldsymbol{g}_n(\boldsymbol{\theta}) \to \mathbf{0}$ con probabilidad 1.

C2: Existe una vecindad de $\boldsymbol{\theta}_0$ tal que, con prob. 1, $\boldsymbol{g}_n(\boldsymbol{\theta})$ es diferenciable y $\dot{\boldsymbol{g}}_n(\boldsymbol{\theta})$ converge a un límite no estocástico que es no singular en $\boldsymbol{\theta}_0$.

C3: $\dfrac{1}{\sqrt{n}}\boldsymbol{g}_n(\boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}_0))$.

Resultado 7 (Normalidad asintótica):

Si $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$ y $\boldsymbol{g}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$ con probabilidad 1. Entonces, bajo los supuestos C1 a C3, tenemos que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_k\big(\mathbf{0}, \boldsymbol{G}_g^{-1}(\boldsymbol{\theta}_0)\big).$$

[6] Journal of Multivariate Analysis 65, 245-260.

Esbozo de la demostración:

Tenemos,

$$\frac{1}{\sqrt{n}}\boldsymbol{g}_n(\boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta})).$$

Usando que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{\mathsf{a}}{=} \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}) = -\Big[\frac{1}{n}\boldsymbol{S}_n(\boldsymbol{\theta})\Big]^{-1}\frac{1}{\sqrt{n}}\boldsymbol{g}_n(\boldsymbol{\theta})$$

$$\xrightarrow{\mathsf{D}} \mathsf{N}_k\big(\mathbf{0}, \boldsymbol{S}^{-1}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{S}^{-\top}(\boldsymbol{\theta})\big),$$

es decir

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}_k\big(\mathbf{0}, \boldsymbol{G}^{-1}(\boldsymbol{\theta})\big).$$

Ejemplo (Modelo mal especificado):

Asuma que disponemos de $X_1, \dots, X_n$ variables IID generadas desde $g(x; \boldsymbol{\theta})$. Aunque los datos son modelados considerando

$$\mathcal{P}_f = \{f(x; \boldsymbol{\theta})^{\otimes n} : \boldsymbol{\theta} \in \Theta\},$$

con funciones de log-verosimilitud y score asociadas

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i; \boldsymbol{\theta}), \qquad \boldsymbol{U}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x_i; \boldsymbol{\theta}),$$

Siguiendo a Kent (1982),[7] tenemos $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$, y

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \boldsymbol{S}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{V}(\boldsymbol{\theta}_0)\boldsymbol{S}^{-1}(\boldsymbol{\theta}_0)),$$

con

$$\boldsymbol{V}(\boldsymbol{\theta}) = \mathsf{E}_g\{\boldsymbol{U}_n(\boldsymbol{\theta})\boldsymbol{U}_n^\top(\boldsymbol{\theta})\}, \qquad \boldsymbol{S}(\boldsymbol{\theta}) = \mathsf{E}_g\Big\{-\frac{\partial \boldsymbol{U}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big\}.$$

[7] Biometrika **69**, 19-27.