MAT-466: Funciones de inferencia

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

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

Observación:

Para una función de inferencia $\boldsymbol{\Psi}_n$ 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

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

Ejemplo (Regresión lineal):

Sea $Y_1, \dots, Y_n$ variables independientes tal que $Y_i \sim \mathsf{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ donde $\boldsymbol{x}_i$ es vector de covariables asociada al i-ésimo individuo y $\boldsymbol{\beta}$ coeficientes de regresión. El estimador de mínimos cuadrados es

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

Funciones de inferencia

Ejemplo (Regresión no-lineal):

Considere $Y_1, \dots, Y_n$ independientes con $Y_i \sim \mathsf{N}(f(\boldsymbol{x}_i; \boldsymbol{\beta}), \sigma^2)$ donde $f(\cdot; \boldsymbol{\beta})$ es una función suave de $\boldsymbol{\beta}$. La función de estimación es dada por

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

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

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

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,

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

(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[1] como:

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

[1] 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,

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

Funciones de inferencia

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

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

Si además $\boldsymbol{r}(\boldsymbol{\beta})$ 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 $\boldsymbol{V}_r(\boldsymbol{\beta}) = \operatorname{Cov}(\boldsymbol{r}(\boldsymbol{\beta}))$ 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 de ahí que $\boldsymbol{S}_r(\boldsymbol{\beta})$ 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 $\boldsymbol{\Psi}_n : \mathbb{R}^k \to \mathbb{R}^k$, la matriz de sensibilidad será cuadrada aunque no necesariamente simétrica.

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

$$\begin{aligned}\boldsymbol{G}_\Phi(\boldsymbol{\beta}) = \boldsymbol{S}_\Psi^\top(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})\{\boldsymbol{A}^{-\top}(\boldsymbol{\beta})\boldsymbol{V}_\Psi^{-1}(\boldsymbol{\beta})\boldsymbol{A}^{-1}(\boldsymbol{\beta})\}\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{S}_\Psi(\boldsymbol{\beta})\\ \boldsymbol{S}_\Psi^\top(\boldsymbol{\beta})\boldsymbol{V}_\Psi^{-1}(\boldsymbol{\beta})\boldsymbol{S}_\Psi(\boldsymbol{\beta}) = \boldsymbol{G}_\Psi(\boldsymbol{\beta}),\end{aligned}$$

de ahí que $\boldsymbol{\Phi}$ y $\boldsymbol{\Psi}$ son equivalentes.

Definición 5 (Algoritmo Newton-scoring):

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

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

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 - \boldsymbol{S}_\Psi^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Psi}(\boldsymbol{\beta}_0), \tag{1}$$

esto sugiere considerar:

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

para llevar a cabo la estimación de parámetros.

Observación:

Este procedimiento fue propuesto por Jørgensen y Knudsen (2004)[2] quienes lo denominaron algoritmo Newton-scoring.

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

Ejemplo:

Considere la función de estimación de minimos cuadrados no lineales

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

con $\boldsymbol{f}_i(\boldsymbol{\beta}) = \partial f(\boldsymbol{x}_i;\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. En efecto, $\mathsf{E}\{\boldsymbol{\Psi}_n(\boldsymbol{\beta})\} = \mathbf{0}$, 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).[3] 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

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

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

Observación:

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

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

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.

Ejemplo (GLM):

Suponga que

$$\boldsymbol{g}(\boldsymbol{\beta}) = \phi \sum_{i=1}^{n} \Big(\frac{\omega_i}{V_i}\Big)^{1/2} \boldsymbol{x}_i (Y_i - \mu_i),$$

donde $\mathsf{E}(Y_i) = \mu_i$, con $h(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Ejemplo (GEE):

Considere

$$\boldsymbol{g}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{C}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

donde $\boldsymbol{Y}_i$ es un vector aleatorio con media $\boldsymbol{\mu}_i$, $\boldsymbol{D}_i = \nabla_\beta \boldsymbol{\mu}_i$ es la matriz de modelo local y $\boldsymbol{C}_i$ es la matriz de covarianza de trabajo.

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.[4]

Definición 8:

Basado en la Ecuación (1) podemos definir una función de estimación normalizada como:

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

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