MAT-466: Ecuaciones de estimación generalizadas

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Ecuaciones de estimación generalizadas

Considere $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_K$ vectores aleatorios independientes, con $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ y suponga

$$f(y; \theta_{ij}, \phi) = \exp[\phi\{y\theta_{ij} - b(\theta_{ij})\} + c(y; \phi)],$$

donde $\mathsf{E}(Y_{ij}) = \mu_{ij} = b'(\theta_{ij})$, $\mathsf{var}(Y_{ij}) = \phi^{-1} V_{ij}$, $V_{ij} = \mathsf{d}\,\mu_{ij}/\,\mathsf{d}\,\theta_{ij}$. Además,

$$g(\mu_{ij}) = \eta_{ij}, \qquad \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}.$$

Bajo el supuesto de independencia, las funciones score son dadas por

$$\boldsymbol{U}_K(\boldsymbol{\beta}) = \phi \sum_{i=1}^K \Big(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\Big)^\top \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{1}$$

estas ecuaciones pueden ser reescritas como:

$$\boldsymbol{U}_K(\boldsymbol{\beta}) = \phi \sum_{i=1}^K \boldsymbol{X}_i^\top \boldsymbol{\Delta}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

donde $\boldsymbol{\Delta}_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i^\top = \operatorname{diag}(\partial \eta_{i1}/\partial \mu_{i1}, \dots, \partial \eta_{in_i}/\partial \mu_{in_i})$.

Sea $\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}$ la solución de la ecuación (1), esto lleva al siguiente resultado.

Resultado 1:

El estimador $\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}$ es consistente y $\sqrt{K}(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}} - \boldsymbol{\beta})$ es asintóticamente normal con media cero y matriz de covarianza

$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}) = \lim_{K\to\infty} K\{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1}\boldsymbol{B}(\boldsymbol{\beta})\{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1},$$

donde

$$\boldsymbol{A}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{X}_i^\top \boldsymbol{\Delta}_i \boldsymbol{V}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i, \qquad \boldsymbol{B}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{X}_i^\top \boldsymbol{\Delta}_i \, \mathsf{Cov}(\boldsymbol{Y}_i) \boldsymbol{\Delta}_i \boldsymbol{X}_i.$$

Ecuaciones de estimación generalizadas

Observación:

Un estimador consistente para $\operatorname{Cov}(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}})$ es obtenido mediante substituir $\operatorname{Cov}(\boldsymbol{Y}_i)$ por $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}))(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}))^\top$. Es decir,

$$\widehat{\operatorname{Cov}}(\widehat{\boldsymbol{\beta}}_{\mathsf{Ind}}) = \{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1}\Big(\sum_{i=1}^{K} \boldsymbol{X}_i^\top \widehat{\boldsymbol{\Delta}}_i \widehat{\boldsymbol{r}}_i \widehat{\boldsymbol{r}}_i^\top \widehat{\boldsymbol{\Delta}}_i \boldsymbol{X}_i\Big)\{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1},$$

donde $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})$.

Sea $\boldsymbol{R}_i(\boldsymbol{\alpha})$ matriz simétrica tal que corresponde a una matriz de correlación y considere $\boldsymbol{\alpha}$ un vector que caracteriza $\boldsymbol{R}_i(\boldsymbol{\alpha})$ que es llamada matriz de correlación de trabajo.

Sea

$$\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \phi \boldsymbol{V}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{V}_i^{1/2}.$$

Esto llevó a Liang y Zeger (1986)[1] a definir:

$$\boldsymbol{\Psi}_K(\boldsymbol{\beta}) = \phi \sum_{i=1}^{K} \Big(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\Big)^\top \{\boldsymbol{V}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{V}_i^{1/2}\}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \qquad (2)$$

Usando que $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^\top$ puede ser escrito como

$$\boldsymbol{D}_i = \boldsymbol{V}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i, \qquad \text{o bien} \qquad \boldsymbol{D}_i = \boldsymbol{W}_i^{1/2} \boldsymbol{V}_i^{1/2} \boldsymbol{X}_i,$$

lleva a diversas formas de definir $\boldsymbol{\Psi}_K(\boldsymbol{\beta})$.

[1] Biometrika **73**, 13-22

El algoritmo de estimación, adopta la forma:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \Big(\sum_{i=1}^{K} \boldsymbol{D}_i^{(r)\top} \boldsymbol{\Sigma}_i^{-(r)} \boldsymbol{D}_i^{(r)}\Big)^{-1} \sum_{i=1}^{K} \boldsymbol{D}_i^{(r)\top} \boldsymbol{\Sigma}_i^{-(r)} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i^{(r)}),$$

que puede ser escrito como un problema de IGLS, como:

$$\boldsymbol{\beta}^{(r+1)} = \Big(\sum_{i=1}^{K} \boldsymbol{D}_i^{(r)\top} \boldsymbol{\Sigma}_i^{-(r)} \boldsymbol{D}_i^{(r)}\Big)^{-1} \sum_{i=1}^{K} \boldsymbol{D}_i^{(r)\top} \boldsymbol{\Sigma}_i^{-(r)} \boldsymbol{Z}_i^{*},$$

con $\boldsymbol{Z}_i^{*} = \boldsymbol{D}_i \boldsymbol{\beta} + \boldsymbol{Y}_i - \boldsymbol{\mu}_i$.

Además, podemos escribir

$$\boldsymbol{\beta}^{(r+1)} = \Big(\sum_{i=1}^{K} \boldsymbol{X}_i^{\top} \boldsymbol{W}_i^{1/2} \boldsymbol{R}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{W}_i^{1/2} \boldsymbol{X}_i\Big)^{-1} \sum_{i=1}^{K} \boldsymbol{X}_i^{\top} \boldsymbol{W}_i^{1/2} \boldsymbol{R}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{W}_i^{1/2} \boldsymbol{Z}_i,$$

con $\boldsymbol{Z}_i = \boldsymbol{\eta}_i + \boldsymbol{W}_i^{-1/2} \boldsymbol{V}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$.

Observación:

- Note que $\widehat{\boldsymbol{\beta}}_{\text{GEE}}$ como solución de $\boldsymbol{\Psi}_K(\boldsymbol{\beta}) = \mathbf{0}$ no depende de ϕ.
- Podemos hacer que $\boldsymbol{\Psi}_K(\boldsymbol{\beta})$ depende solamente de $\boldsymbol{\beta}$ mediante substituir $\boldsymbol{\alpha}$ por un estimador $\sqrt{K}$-consistente, es decir

$$\sqrt{K}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = O_p(\mathbf{1}).$$

De ahí que

$$\boldsymbol{\Psi}_K^*(\boldsymbol{\beta}) = \sum_{i=1}^K \Big(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\Big)^\top \{\boldsymbol{V}_i^{1/2}\boldsymbol{R}_i(\widehat{\boldsymbol{\alpha}})\boldsymbol{V}_i^{1/2}\}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

Recuerde que:

Para $\{X_n\}$, $\{Y_n\}$ secuencias de variables aleatorias entonces $X_n = O_p(Y_n)$ si existe $\epsilon > 0$, $M = M(\epsilon)$ y $n_0 = n_0(\epsilon)$, tal que

$$\mathsf{P}(|X_n| \leq M|Y_n|) \geq 1 - \epsilon, \qquad \text{para todo } n > n_0.$$

Resultado 2:

Suponga que

(i) $\widehat{\boldsymbol{\alpha}}$ es $K^{1/2}$-consistente dado $\boldsymbol{\beta}$ y ϕ.

(ii) $\widehat{\phi}$ es $K^{1/2}$-consistente dado $\boldsymbol{\beta}$.

(iii) $|\partial\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)| \leq K(\boldsymbol{Y}, \boldsymbol{\beta})$ que es $O_p(1)$.

Entonces $\sqrt{K}(\widehat{\boldsymbol{\beta}}_{\text{GEE}} - \boldsymbol{\beta})$ es asintóticamente normal con media cero y matriz de covarianza

$$\operatorname{Cov}(\widehat{\boldsymbol{\beta}}_{\text{GEE}}) = \lim_{K\to\infty} K\{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1}\boldsymbol{B}(\boldsymbol{\beta})\{\boldsymbol{A}(\boldsymbol{\beta})\}^{-1},$$

donde

$$\boldsymbol{A}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i, \qquad \boldsymbol{B}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \operatorname{Cov}(\boldsymbol{Y}_i) \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i.$$

Ecuaciones de estimación generalizadas

Sea

$$r_{ij} = (Y_{ij} - \widehat{\mu}_{ij})/\{V(\widehat{\mu}_{ij})\}^{1/2}, \qquad \widehat{\mu}_{ij} = \mu_{ij}(\widehat{\boldsymbol{\beta}}),$$

para $i = 1, \ldots, K$; $j = 1, \ldots, n_i$.

Entonces, podemos considerar

$$\widehat{\phi}^{-1} = \frac{1}{N-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} r_{ij}^2,$$

con $N = \sum_{i=1}^{K} n_i$. Este estimador es $K^{1/2}$-consistente siempre que los cuartos momentos de Y_{ij} sean finitos.

Suponga, sin pérdida de generalidad, que $n_i = n$. Considere los siguientes ejemplos,

Ejemplo:

Considere $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}$, entonces obtenemos las ecuaciones de estimación de independencia dadas en (1).

Ejemplo:

Sea $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$ donde $\alpha_j = \text{corr}(Y_{ij}, Y_{i,j+1})$, $j = 1, \dots, n-1$. Un estimador natural para α_j es

$$\widehat{\alpha}_j = \frac{\phi}{K - p} \sum_{i=1}^{K} r_{ij} r_{i,j+1}.$$

Ejemplo:

Suponga corr$(Y_{ij}, Y_{ij'}) = \alpha$ para todo $j \neq j'$. Esto corresponde a la estructura de equicorrelación. De este modo,

$$\widehat{\alpha} = \frac{\phi}{\sum_{i=1}^{K} n_i(n_i - 1)/2 - p} \sum_{i=1}^{K} \sum_{j>j'} r_{ij} r_{ij'}$$

Ejemplo:

Sea corr$(Y_{ij}, Y_{ij'}) = \alpha^{|j-j'|}$, que es llamada estructura de correlación AR-1. Como E$(r_{ij} r_{ij'}) \approx \alpha^{|j-j'|}$ entonces α puede ser estimado desde la regresión de $\log(r_{ij} r_{ij'})$ sobre $\log |j - j'|$.

Algunos paquetes en R para estimación de GEE:

gee: Generalized Estimation Equation Solver.
(Carey, V.J., portado a R por Lumley, T. y Ripley, B).

geepack: Solve Generalized Estimating Equations.
McDaniel, L.S., Henderson, N.C., Rathouz, P.J. (2013).
Fast pure R implementation of GEE: application of the Matrix package.
The R Journal 5/1, 181-187.

geeM: Generalized Estimating Equation Package.
Halekoh, U., Højsgaard, S., Yan, J. (2006).
The R package geepack for generalized estimating equations.
Journal of Statistical Software 15/2, 1-11.

Datos de incontinencia urinaria (Preisser y Qaqish, 1999)

- Datos provienen de un estudio para evaluar el impacto de la incontinencia urinaria sobre la vida de pacientes ancianos (GUIDE) mayores de 76 años.

- La respuesta es binaria, indicando si el individuo siente que su rutina diaria se ve afectada por pérdidas accidentales de orina.

- Datos obtenidos para 137 ancianos agrupados en 38 prácticas médicas (cluster), los datos son desbalanceados (de 1 a 8 pacientes por cluster).

- Se dispone de 5 regresores, sexo, edad, accidentes diarios, severidad y número de veces que usa el baño diariamente.

- Se consideró un enlace logístico para el siguiente modelo

$$\operatorname{logit}(\mu_{ij}) = \beta_0 + \beta_1 \, \mathsf{Sexo} + \beta_2 \, \mathsf{Edad} + \beta_3 \, \mathsf{Accidentes} + \beta_4 \, \mathsf{Severo} + \beta_5 \, \mathsf{Baño}.$$

  Además se asumió una estructura de equicorrelación para $\boldsymbol{R}_i(\alpha)$.

Datos de incontinencia urinaria (Preisser y Qaqish, 1999)

Resultados del ajuste mediante GEE para los datos GUIDE.

Variable	Estimación	Err.Est.	Z
Intercepto	-3.054	0.959	-3.185
Sexo	-0.745	0.600	-1.242
Edad	-0.676	0.561	-1.205
Accidentes	0.392	0.093	4.202
Severo	0.812	0.359	2.263
Baño	0.108	0.099	1.090

Se obtuvo además $\widehat{\alpha} = 0.0932$.

EX UMBRA IN SOLEM