Diseño SIC

Daniel Czarnievicz

2017

Estrategia de selección

Se toma una muestra bajo diseño SI de tamaño fijo n_I de los N_I clusters en U_I. Luego todos los elementos en los clusters seleccionados son relevados.

El estimador $\hat{t}_\pi$

$$\star \; \hat{t}_\pi = N_I \, \bar{t}_{s_I} = N_I \sum\nolimits_{s_I} \frac{t_{y_i}}{n_I}$$

$$\star \; V_{SIC}(\hat{t}_\pi) = \frac{N_I^2}{n_I}(1 - f_I) S^2_{t_{U_I}}$$

$$\text{donde } S^2_{t_{U_I}} = \frac{1}{N_I - 1} \sum\nolimits_{U_I} \left(t_{y_i} - \bar{t}_{U_I} \right)^2 \text{ y } \bar{t}_{U_I} = \sum\nolimits_{U_I} \frac{t_{y_i}}{N_I}$$

$$\star \; \hat{V}_{SIC}(\hat{t}_\pi) = \frac{N_I^2}{n_I}(1 - f_I) S^2_{t_{s_I}}$$

$$\text{donde } S^2_{t_{s_I}} = \frac{1}{n_I - 1} \sum\nolimits_{s_I} \left(t_{y_i} - \bar{t}_{s_I} \right)^2 \text{ y } \bar{t}_{s_I} = \sum\nolimits_{s_I} \frac{t_{y_i}}{n_I}$$

Efecto diseño

Sea $\delta = 1 - \frac{S^2_{y_W}}{S^2_{y_U}}$ el coeficiente de homogeneidad donde:

- $S^2_{y_W} = \frac{1}{N - N_I} \sum\nolimits_{U_I} \sum\nolimits_{U_i} \left(y_k - \bar{y}_{U_i} \right)^2$ es la pooled-variance interna de los clusters
- $\bar{y}_{U_i} = \frac{1}{N_I} \sum\nolimits_{U_i} y_k$ es la media del i-ésimo cluster

Si $S^2_{y_{U_i}} = \frac{1}{N_i - 1} \sum\nolimits_{U_i} \left(y_k - \bar{y}_{U_i} \right)^2$ es la varianza de y en el cluster U_i, entonces:

$$S^2_{y_W} = \frac{\sum_{U_I} (N_I - 1) S^2_{y_{U_i}}}{\sum_{U_I} (N_i - 1)}$$

Por lo tanto, $S^2_{y_W}$ es el promedio ponderado de las varianzas $S^2_{y_{U_i}}$ en los N_I clusters, siendo $N_i - 1$ los respectivos pesos.

$-\frac{N_I - 1}{N - N_I} \leq \delta \leq 1$

- $\delta > 0 \Leftrightarrow S^2_{y_W} < S^2_{y_U}$
- $\delta = 0 \Leftrightarrow S^2_{y_W} = S^2_{y_U}$
- $\delta < 0 \Leftrightarrow S^2_{y_W} > S^2_{y_U}$

Un δ pequeño implica que los elementos en el mismo cluster son disimiles en y_k. Un δ grande implica que los elementos en el mismo cluster tienen valores parecidos de y_k. Si $\delta = 1$, entonces la variación interna de todos los clusters es 0. Si $\delta = -\dfrac{N_I - 1}{N - N_I}$ entonces $\bar{y}_{U_I}$ es igual para todos los clusters.

Sea $\bar{N}$ el promedio de elementos por cluster: $\bar{N} = \dfrac{N}{N_I}$. Sea $K_I = \dfrac{N_I^2}{n_I}(1 - f_I)$. Ses COV la covarianza entre N_i y $N_i\,\bar{y}_{U_i}^2$ tal que: $COV = \dfrac{1}{N_I - 1}\sum_{U_I}(N_i - \bar{N})N_i\,\bar{y}_{U_i}^2$. Por lo tanto: $S_{t_{U_I}}^2 = \bar{N}\,S_{y_U}^2\left(1 + \dfrac{N - N_I}{N_I - 1}\,\delta\right) + COV$. Con esto podemos entonces expresar la varianza como:

$$\star\ V_{SIC}(\hat{t}_\pi) = \left(1 + \frac{N - N_I}{N_I - 1}\,\delta\right)\bar{N}\,K_I S_{y_U}^2 + K_I\,COV$$

Luego, dado que $E(n_S) = n_I\,\bar{N} = n_I\,\dfrac{N}{N_I} = n$ podemos comprar el diseño SIC con el diseño SI de tamaño $n = n_I\,\bar{N}$, reescribiendo la varianza del SIC como:

$$\star\ V_{SIC}(\hat{t}_\pi) = \left(1 + \frac{N - N_I}{N_I - 1}\,\delta\right)V_{SI}(\hat{t}_\pi) + K_I\,COV$$

$$\star\ deff(SIC, \hat{t}_\pi) = \frac{V_{SIC}(\hat{t}_\pi)}{V_{SI}(\hat{t}_\pi)} = 1 + \frac{N - N_I}{N_I - 1}\,\delta + \frac{COV}{\bar{N}\,S_{y_U}^2}$$

Si todos los clusters son de igual tamaño, entonces $COV = 0$, por lo que $V_{SIC}(\hat{t}_\pi) < V_{SI}(\hat{t}_\pi) \Leftrightarrow \delta < 0$, lo que requiere una variación intra-clusters lo suficientemente grande. Si el tamaño de los clusters no es fijo y la correlación entre N_i y $N_i\,\bar{y}_{U_i}^2$ es positiva, el incremento de varianza debido a la selección por clusters puede empeorar significativamente dado que $K_I\,COV$ puede ser grande.