Estudio del comportamiento de estadísticas para datos binarios correlacionados en muestras pequeñas*

Hachuel, Leticia[1]; Boggio, Gabriela[1]; Wojdyla, Daniel; Cuesta, Cristina; Servy, Elsa

Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

[1]*Consejo de Investigaciones de la Universidad Nacional de Rosario*

Objetivo

Comparación de algoritmos de generación de datos binarios correlacionados y evaluación del comportamiento de estadísticas que consideran la falta de independencia de las observaciones mediante estudios de Montecarlo.

Algoritmos de generación de datos

Algoritmo de Park, Park y Shin

Para generar un vector aleatorio de k variables binarias se considera un conjunto de k variables aleatorias Poisson Z_1, ..., Z_k que son, a la vez, sumas parciales de variables Poisson independientes, $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$.

Los valores esperados y la estructura de las correlaciones de las variables binarias Y_v v=1,...,k definidas por $Y_{vi}=I_{\{0\}}(Z_{vi})$ se obtienen controlando el esquema de aparición simultánea de $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$ y las magnitudes de β_1, ..., β_τ. El algoritmo describe cómo determinar τ (número de variables X's), β_1, ..., β_τ y el esquema de sumas parciales a fin de generar un vector aleatorio binario k-dimensional $(Y_1, ..., Y_k)'$ con un vector de medias pre-especificado $P=(\pi_1, \pi_2,... \pi_k)'$ y matriz de correlación también pre-especificada R.

Algoritmo de Servy, Hachuel y Wojdyla

Cada vector k-dimensional de variables binarias se genera por los k primeros pasos de una cadena de Markov. La primera respuesta se elige de acuerdo a una distribución de probabilidades $\{a_{u(1)}, u(1) \varepsilon R; R=\{0,1\}\}$. Las respuestas siguientes se simulan según una cadena de Markov, cuyas probabilidades iniciales son $\{a_{u(1)}, u(1) \varepsilon R\}$ y matriz de transición $M=((p_{\alpha\beta}))$ $\alpha,\beta \varepsilon R$ donde $p_{\alpha\beta}$ designa a la probabilidad condicional de que $u(v) = \beta$ dado que $u(v-1) = \alpha$ para v = 2,...,k.

Para formar otro conglomerado, es decir otro vector de respuestas binarias correlacionadas, se inicia otra cadena similar independiente de la anterior.

Compatibilización de los parámetros de ambos algoritmos

i) Aplicar el algoritmo Servy et al.

ii) Elegir aquellos casos que originan que los valores de $R=(\rho_{12}, \rho_{13}, \rho_{23})$ sean positivos. Calcular $F=(P_{000}, P_{001}, P_{010}, P_{100}, P_{011}, P_{101}, P_{110}, P_{111})$ y $P=(\pi_1, \pi_2, \pi_3)$.

iii) Considerar estos valores de P y R como datos iniciales para el algoritmo Park et al.

Evaluación de los modelos

Una vez elegidos los casos compatibles en términos de los parámetros, se compara la consistencia de ambos algoritmos generando muestras y calculando las estimaciones de las probabilidades marginales y de las correlaciones entre pares de posiciones.

Se eligieron tres escenarios paramétricos para la evaluación de los modelos con baja, media y alta correlación. Los resultados fueros satisfactorios en cuanto a exactitud y precisión de las estimaciones. Se presentan los resultados para uno de los escenarios .

Promedios y desvíos est. de las estimaciones de las componentes de P y R

n	Park et al.			Servy et al.		
	π_1=0.25	π_2=0.30	π_3=0.34	π_1=0.25	π_2=0.30	π_3=0.34
15	0.25 (0.11)	0.31 (0.12)	0.34 (0.12)	0.25 (0.12)	0.31 (0.12)	0.35 (0.12)
30	0.25 (0.08)	0.30 (0.09)	0.34 (0.09)	0.25 (0.08)	0.31 (0.08)	0.35 (0.09)
50	0.26 (0.06)	0.30 (0.07)	0.34 (0.07)	0.25 (0.06)	0.30 (0.07)	0.34 (0.07)
70	0.25 (0.05)	0.30 (0.06)	0.34 (0.06)	0.25 (0.05)	0.30 (0.06)	0.34 (0.06)
100	0.25 (0.04)	0.30 (0.05)	0.34 (0.05)	0.25 (0.04)	0.30 (0.05)	0.34 (0.05)
n	ρ_{12}=0.76	ρ_{13}=0.59	ρ_{23}=0.77	ρ_{12} =0.76	ρ_{13}=0.59	ρ_{23}=0.77
15	0.76 (0.19)	0.60 (0.23)	0.78 (0.18)	0.75 (0.20)	0.57 (0.23)	0.77 (0.18)
30	0.76 (0.13)	0.58 (0.16)	0.77 (0.13)	0.75 (0.14)	0.58 (0.16)	0.78 (0.12)
50	0.76 (0.10)	0.59 (0.12)	0.77 (0.09)	0.76 (0.11)	0.59 (0.12)	0.77 (0.09)
70	0.76 (0.09)	0.59 (0.10)	0.77 (0.08)	0.76 (0.09)	0.57 (0.10)	0.77 (0.08)
100	0.76 (0.07)	0.59 (0.08)	0.78 (0.07)	0.76 (0.07)	0.57 (0.09)	0.77 (0.07)

Comparación de estadísticas

Se estudia el control del error tipo I de las siguientes estadísticas:

X^2_{SG} : Score generalizada

X^2_{WG} : Wald generalizada

X^2_{WGc}: Wald generalizada corregida

Se calculan bajo tres estructuras de correlación: Independencia, AR(1) y fija.

Diseño del estudio de Montecarlo

El comportamiento de las estadísticas se estudia para el caso particular de comprobar la hipótesis Ho: $\beta_2 = 0$ cuando se ajusta el siguiente modelo:

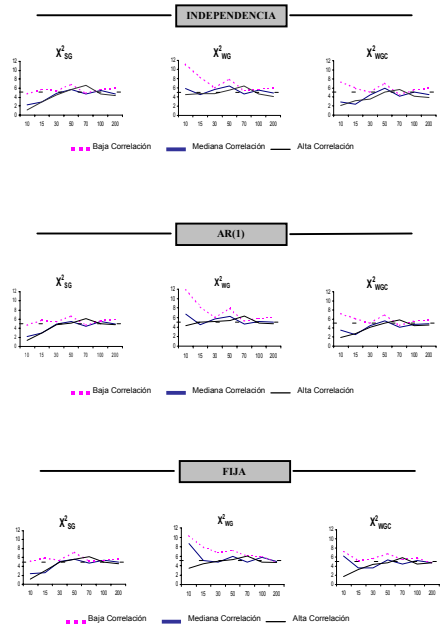
$ln\{\pi /(1-\pi)\} = \beta_1 + \beta_2 X$ con X=0,1

a datos generados por el algoritmo de Servy et al. Para ello:

- Se simulan muestras simples aleatorias de n= 10, 15, 30, 50, 100 y 200 conglomerados de tamaño fijo k=3.
- En cada muestra se calculan las estadísticas y se evalúa el rechazo o no de la hipótesis a un nivel del 5%.
- Se repite el procedimiento 1000 veces y se calcula el porcentaje real de rechazo.

Resultados

Comportamiento de las estadísticas según tamaño de la muestra, de acuerdo a la intensidad de la correlación intragrupo y a las diferentes especificaciones de la matriz de correlación de trabajo.



Discusión

En la evaluación de cuán estrechamente las muestras generadas por los modelos son capaces de estimar los parámetros especificados, los resultados obtenidos han mostrado una notable concordancia en la estimación de las correlaciones intra-conglomerados y en los valores de las probabilidades marginales.

Con respecto al comportamiento de las estadísticas, las características más salientes son la liberalidad de las estadísticas tipo Wald y el efecto conservador que producen altas correlaciones intragrupo tanto en las estadísticas tipo Wald como Score.

"Estudio del comportamiento de estadísticas para medidas repetidas en muestras pequeñas bajo escenarios múltiples" Proyecto de Investigación de la Secretaría de Ciencia y Tecnología de la UNR – 2001/2002.