Leticia Hachuel
Gabriela Boggio
Guillermina Harvey
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

ESTUDIO COMPARATIVO DE MÉTODOS DE ESTIMACIÓN EN UN MODELO LOGIT MIXTO[1]

1. INTRODUCCIÓN

En muchas aplicaciones los individuos bajo estudio presentan algún tipo de agrupamiento que provoca que las observaciones provenientes de sujetos de un mismo grupo tiendan a estar correlacionadas. En el análisis estadístico de este tipo de datos frecuentemente se busca modelar respuestas como función de covariables, ajustando los resultados por la correlación potencial de las respuestas de individuos dentro de un mismo grupo. Un enfoque posible para ello es el que incluye en el modelo efectos no observados que varían aleatoriamente y reciben el nombre de efectos aleatorios (Goldstein, 2003; Fitzmaurice et al., 2004).

Estos modelos constituyen la clase de los denominados modelos lineales generalizados mixtos (MLGM), los cuales admiten variables respuestas no normales y permiten modelar una función de la media a través de efectos fijos, asociados a variables medidas tanto a nivel individual como grupal, y de efectos aleatorios en el predictor lineal, los que generalmente se suponen distribuidos normalmente.

El enfoque clásico de estimación de estos modelos es el de máximo-verosimilitud ya que es posible especificar la función de probabilidad paramétrica completa. Otra alternativa de estimación proviene del paradigma bayesiano, el cual requiere la especificación de dos partes: un modelo estadístico que describe la distribución de los datos dadas cantidades desconocidas y una distribución a priori que describe creencias acerca de esas cantidades desconocidas, parámetros, independiente de los datos.

Si bien el supuesto de distribución normal de los efectos aleatorios tiene ciertas ventajas, persiste la incertidumbre acerca de cómo la falta de una correcta especificación de la distribución de probabilidad de estos efectos afecta la estimación de los parámetros del modelo.

En este trabajo se realiza un estudio por simulación para evaluar la influencia de una mala especificación de la distribución de probabilidad de los efectos aleatorios sobre la estimación de los parámetros de un MLGM comparando ambos enfoques de estimación.

En la sección siguiente se describe brevemente el modelo estadístico utilizado, seguidamente se presenta el estudio de simulación elegido y los resultados hallados.

[1] En este trabajo participan en calidad de auxiliares de investigación los alumnos de la Licenciatura en Estadística Joaquín Trombetti y Diego Marfetán.

2. METODOLOGÍA

2.1. Modelo lineal generalizado mixto

La premisa básica de los MLGM es que la correlación entre las unidades de un mismo grupo puede pensarse que surge por el hecho de compartir un conjunto de efectos aleatorios. Condicional sobre los efectos aleatorios, las observaciones de diferentes grupos se suponen independientes y con una distribución de probabilidad perteneciente a la familia exponencial.

Sea Y_{ij} la respuesta para el j-ésimo individuo del i-ésimo grupo, la cual puede ser continua, binaria o de conteo. Asociado con cada Y_{ij} hay un vector (fila) X_{ij} de dimensión 1x (p+1), que contiene el valor 1 asociado al intercepto y el valor de p covariables que pueden variar de grupo a grupo o bien de individuo a individuo dentro de cada grupo.

Para el caso particular en que Y_{ij} es una respuesta binaria, un posible MLGM simple es un modelo logístico con interceptos aleatorios que se especifica de la siguiente forma (Fitzmaurice et al., 2004):

1.- Condicional sobre un único efecto aleatorio, b_i, las Y_{ij} son independientes y tienen una distribución de probabilidad Bernoulli, con

$$Var(Y_{ij}/b_i)= E(Y_{ij}/b_i)\ \{1- E(Y_{ij}/b_i)\}\ \ (\phi=1). \qquad (1)$$

2.- La media condicional de Y_{ij} depende de efectos fijos y aleatorios a través de la siguiente expresión:

$$\log\left\{\frac{Pr(Y_{ij}=1/b_i)}{1-Pr(Y_{ij}=1/b_i)}\right\}=X_{ij}\beta+b_i\,. \qquad (2)$$

Es decir la media condicional de Y_{ij} se relaciona con el predictor lineal a través del enlace logit, siendo β el vector de (p+1) parámetros fijos.

3.- El único efecto aleatorio b_i tiene una distribución de probabilidad univariada con media cero y variancia σ_b^2 que usualmente se supone Normal.

2.2. Métodos de estimación

Una vez postulado un MLGM, el método de estimación máximo-verosímil requiere especificar la función de verosimilitud que se considera función de los parámetros dadas las observaciones, esto es:

$$L(\beta,\sigma/y)=\int f(y/b;\beta)f(b;\sigma)db\,. \qquad (3)$$

Esta expresión generalmente es difícil de resolver y se complica aún más cuando el número de efectos aleatorios aumenta.

Se pueden utilizar métodos de integración numérica para aproximar la verosimilitud. El error inducido por reemplazar la integral por una suma finita, como lo hacen los métodos de cuadratura de Gauss-Hermite, se hace cada vez más difícil de controlar a medida que la dimensión de la integral aumenta. Las aproximaciones convergen a las estimaciones máximo-verosímiles (MV) a medida que el número de puntos de cuadratura se incrementa de una manera apropiada para la integración numérica (Song, 2007).

Por otro lado, el enfoque bayesiano trata a los parámetros como variables aleatorias con una determinada distribución de probabilidad y la inferencia estadística descansa en la distribución a posteriori de los parámetros, dados los datos. Para obtener estas distri-

buciones a posteriori se deben elegir previamente las distribuciones a priori de los parámetros del modelo, las que se combinan luego con la función de verosimilitud de la siguiente manera, según el teorema de Bayes (Dobson, 2008; Bernardo & Smith, 1994):

$$h(\beta,\sigma / y) = \frac{f(y / \beta,\sigma) f(\beta,\sigma)}{f(y)} \qquad (4)$$

El denominador de esta expresión depende sólo de los datos por lo que resulta irrelevante para la inferencia sobre los parámetros; de ahí que la distribución a priori multiplicada por la función de verosimilitud determina finalmente la distribución a posteriori.

Contrariamente al enfoque frecuentista, el bayesiano no diferencia análisis para muestras grandes o pequeñas, ya que la inferencia descansa en la distribución a posteriori sin tener en cuenta el tamaño de la muestra.

Excepto en casos particulares, no hay una expresión cerrada para la distribución a posteriori, por lo que se usan métodos de simulación para aproximarla. El método más conocido es el MCMC, por el cual se diseña un proceso estocástico con forma de Cadena de Markov de manera tal que la distribución estacionaria sea la distribución a posteriori del parámetro de interés.

El proceso comienza con la selección de estimaciones iniciales y se desprecia un primer tramo de la cadena hasta que la distribución sea cercana a la estacionaria. Los restantes valores simulados se supone que proveen información acerca de la distribución a posteriori. Como las observaciones sucesivas de la cadena de Markov están correlacionadas, se poda el proceso para obtener observaciones levemente correlacionadas. Finalmente cuando se tienen las suficientes observaciones después del descarte de forma de aproximarse lo mejor posible a la distribución a posteriori, se obtienen medidas resumen de interés como la media y desviación estándar. Alcanzado este punto, los métodos bayesianos de inferencia son paralelos a aquéllos de la inferencia frecuentista.

La metodología bayesiana necesita especificar distribuciones a priori para todos los parámetros del modelo, lo cual puede llegar a ser una complicación en el caso particular abordado ya que, además de especificar una distribución a priori para los parámetros fijos del modelo, es necesario especificar la distribución a priori del hiperparámetro relacionado con la variancia del efecto aleatorio.

3. ESTUDIO DE SIMULACIÓN

Se diseña un estudio por simulación a fin de comparar el comportamiento de los estimadores de los parámetros de un MLGM particular: un modelo logístico con intercepto aleatorio y una única variable explicativa, el cual se formaliza de la siguiente manera:

$$\log\left\{\frac{Pr(Y_{ij}=1/b_i)}{1-Pr(Y_{ij}=1/b_i)}\right\} = \eta_{ij} = \beta_0 + \beta_1 X_{ij} + b_i . \qquad (5)$$

Los estimadores a evaluar son los obtenidos por el método de máxima verosimilitud (MV) implementado en el procedimiento GLIMMIX de SAS y el método bayesiano proporcionado por el procedimiento MCMC del mismo programa computacional.

Para llevar adelante el estudio por simulación se generan datos a partir del algoritmo empleado por Evans y Hosmer (2004), que considera el modelo (5). En él, X_{ij} es el valor asumido por una variable tipo Bernoulli con probabilidad igual a 0.5 y b_i es el valor asumido por una variable aleatoria Normal(0,1) en un caso -caso a)- o Exponencial(1) en otro caso -caso b)-.

Para valores predeterminados de β_0 y β_1 y valores aleatorios para b_i y X_{ij}, se determinan $\pi_{ij} = P(Y_{ij} = 1/ b_i)$, a partir de las cuales se obtienen valores binarios 0 ó 1 comparando cada probabilidad con un valor elegido al azar de una distribución Uniforme definida en el intervalo [0;1].

A los valores binarios generados por el algoritmo recién descrito se les ajusta el modelo logístico mixto (5), con un coeficiente aleatorio que siempre se supone distribuido Normal, ya sea que los datos generados provengan del caso a) o b). Se obtienen las estimaciones de los coeficientes del modelo y de la variancia de los efectos aleatorios.

Se repite este proceso de generación de datos y ajuste de modelos 1000 veces. Se calcula el promedio de las estimaciones de los parámetros del modelo a través de las 1000 muestras de acuerdo a los dos métodos de estimación presentados y para las dos especificaciones de la distribución de los efectos aleatorios. Los escenarios de análisis se definieron teniendo en cuenta diferentes valores y combinaciones de:

n: número de grupos en la muestra.

k: número de individuos dentro de cada grupo.

En este trabajo se presentan los resultados considerando en el modelo β_0=1 y β_1=0.8 y combinaciones entre valores de k = 5, 10, 30 y valores de n = 10, 30. Los escenarios definidos por estos valores de k y n se eligieron en función de la falta de consenso acerca del número mínimo de grupos necesarios para garantizar los resultados del ajuste de modelos de regresión jerárquicos (Austin, 2010).

La elección de este modelo simple con una única covariable permite evaluar la influencia de la distribución de los efectos aleatorios sobre la estimación de los parámetros del modelo con una demanda computacional aceptable teniendo en cuenta sobretodo los requerimientos computacionales propios de los métodos bayesianos.

Para la estimación de este modelo mediante el procedimiento MCMC se eligieron distribuciones a priori no informativas para los parámetros. Para los efectos fijos se supusieron distribuciones normales con media 0 y variancia igual a 1000. Los interceptos aleatorios siempre se supusieron con distribución Normal y la distribución a priori de su desvío estándar se supuso Uniforme en el intervalo (0, 100). Para cada parámetro se calculó la media de la distribución a posteriori respectiva. Se eligió la estadística de Geweke para comprobar la convergencia de las cadenas para los parámetros de interés y se calculó la variancia de los efectos aleatorios elevando al cuadrado la media de la distribución a posteriori del desvío estándar.

4. RESULTADOS

El parámetro del modelo de mayor interés es el asociado a la covariable incluida en el modelo. Asimismo se presentan los resultados hallados para la estimación de la componente de variancia, esto es la variancia de la distribución de los efectos aleatorios.

Caso a): Correcta especificación de la distribución de los efectos aleatorios

En relación al coeficiente de regresión, las estimaciones por ambos métodos presentan los mayores sesgos cuando se consideran 5 sujetos por grupo en una muestra de 10 grupos. Sin embargo, las estimaciones máximo-verosímiles son mucho menos sesgadas y este sesgo se vuelve despreciable sobre todo cuando aumenta el número de grupos y de individuos por grupo.

En cuanto a la estimación MV de la componente de variancia, es en general buena con sesgos levemente positivos para escenarios con escaso número de individuos cualquiera sea el número de grupos. Por el contrario, la estimación bayesiana presenta grandes sesgos positivos aún con tamaños de muestra moderados y el sesgo continúa siendo grande aún con k igual a 30 y n igual a 10 (Tabla 1).

Tabla 1: Estimaciones del coeficiente de regresión y componente de variancia en caso a)

		β_1		σ_b^2	
		Estimación MV	**Estimación Bayesiana**	**Estimación MV**	**Estimación Bayesiana**
n	k	**Promedio**	**Promedio**	**Promedio**	**Promedio**
10	5	0,856	1,497	1,166	11,082
	10	0,837	0,929	1,040	3,291
	30	0,814	0,820	0,910	1,798
30	5	0,834	0,888	1,080	1,787
	10	0,802	0,826	1,024	1,292
	30	0,805	0,812	0,995	1,157

Caso b): Incorrecta especificación de la distribución de los efectos aleatorios

Los resultados hallados cuando se generan datos con una distribución exponencial para los efectos aleatorios pero el ajuste supone distribución normal, no se diferencian mayormente de los encontrados para el caso a) en lo que respecta a la estimación del coeficiente de regresión. En cambio hay notorios sesgos negativos en la estimación MV de la componente de variancia. El método bayesiano proporciona una muy mala estimación para k igual a 5 y n igual a 10 y en el resto de los escenarios, sesgos importantes en ambos sentidos (Tabla 2).

Tabla 2: Estimaciones del coeficiente de regresión y componente de variancia en caso b)

		β_1		σ_b^2	
		Estimación MV	**Estimación Bayesiana**	**Estimación MV**	**Estimación Bayesiana**
n	k	**Promedio**	**Promedio**	**Promedio**	**Promedio**
10	5	0,840	1,410	0,639	6,075
	10	0,830	0,913	0,512	1,995
	30	0,806	0,839	0,565	1,285
30	5	0,834	0,887	0,540	0,887
	10	0,809	0,851	0,492	0,661
	30	0,804	0,800	0,560	0,681

Resumiendo, a partir de los escenarios considerados resultan más favorables las estimaciones proporcionadas por el método máximo-verosímil. No se observa influencia significativa de la incorrecta especificación de la distribución de los efectos aleatorios sobre la estimación del coeficiente de regresión; sí en cambio, la influencia es fuerte en la estimación de la componente de variancia.

5. CONSIDERACIONES FINALES

En este trabajo se estudia el comportamiento de los estimadores obtenidos por los métodos de estimación de máxima-verosimilitud y Bayes de los parámetros de un modelo logístico simple con intercepto aleatorio.

El estudio por simulación pone en evidencia que el método máximo verosímil funciona correctamente aún para muestras chicas y el uso del procedimiento MCMC tiende a dar resultados que tienen mayor sesgo comparado con el máximo-verosímil tanto en relación al coeficiente de regresión como a la variancia de los efectos aleatorios, sobre todo cuando el número de observaciones por grupo es chico.

Esta etapa del estudio presenta limitaciones que se deben puntualizar. En primer lugar, debido al intenso trabajo computacional requerido por las simulaciones cuando se utilizan métodos bayesianos resulta difícil examinar modelos de regresión más complejos con mayor número de covariables o componentes de variancia. Otra limitación, asociada al mismo método, concierne al uso de sólo distribuciones a priori no informativas, lo cual puede influenciar sobre las distribuciones a posteriori cuando el tamaño del conjunto de datos es chico o moderado (Austin, 2010). Por otro lado, el modelo puesto a prueba en esta presentación no presenta un alto grado de correlación entre las observaciones de un mismo grupo. Se está trabajando en complejizar el modelo de generación de datos multiplicando el coeficiente aleatorio por una constante cuya magnitud presenta una relación directa con el grado de correlación entre las respuestas de los individuos de un mismo grupo. De esta manera se podrá evaluar si los resultados varían según el grado de asociación entre las respuestas de unidades de un mismo grupo.

6. REFERENCIAS BIBLIOGRÁFICAS

Austin, P. C. (2010). "Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures". *The International Journal of Biostatistics*, 6(1): 1-18.

Bernardo, J. M.; Smith, A.F.M. (1994). *Bayesian Theory*. John Wiley & sons.

Dobson, A.; Barnett, A. (2008). "An Introduction to Generalized Linear Models", 3rd ed. Chapman & Hall/CRC.

Evans, S. R.; Hosmer, D. W. (2004). "Goodness of Fit Tests for Logistic GEE Models: Simulation Results". *Communications in Statistics: Simulation and Computation*, 33(1): 247-258.

Fitzmaurice, G.; Laird, N.; Ware, J. (2004). *Applied longitudinal analysis*. John Wiley & Sons.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. Kendall's Library of Statistics. London.

SAS Institute, Inc. (2011). SAS/STAT User's guide, version 9.3. Cary, NC, USA.

Song, P. X. K. (2007). *Correlated data analysis: modeling, analytics and applications*. Springer, New York.