Leticia Hachuel
Gabriela Boggio
Virginia Borra
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

EVALUACIÓN DE LOS SUPUESTOS SOBRE LOS EFECTOS ALEATORIOS DEL MODELO LOGÍSTICO NORMAL: SU APLICACIÓN EN EL ESTUDIO DEL LUPUS SISTÉMICO ERITEMATOSO

1. INTRODUCCIÓN

La estimación en los modelos lineales generalizados mixtos para medidas repetidas no normales se basa en la teoría de máxima verosimilitud, la cual supone que el modelo de probabilidad está correctamente especificado.

Se sabe que los resultados obtenidos mediante el ajuste de dichos modelos no son siempre robustos bajo errores en la especificación de la estructura de los efectos aleatorios. Por lo tanto el uso de una herramienta de diagnóstico para la detección de esta falta de especificación adecuada resulta de importancia (Alonso et al., 2008; Agresti et al., 2004; Letiere et al., 2000).

En este trabajo se presenta la aplicación de un conjunto de tests sugeridos por Alonso et al. (2008) a un modelo lineal generalizado mixto particular, el modelo logístico normal en el estudio relacionado con la demora en el diagnóstico del lupus sistémico eritematoso (LES). La demora en el diagnóstico de esta enfermedad requiere una especial atención ya que la dilación en el diagnóstico de esta enfermedad puede provocar severos daños y mayor mortalidad a largo plazo.

En la sección siguiente se describe brevemente el modelo estadístico utilizado y las pruebas estadísticas acerca de la distribución de los efectos aleatorios sugeridas.

2. METODOLOGÍA

2.1. El modelo GLMM

En los últimos años ha habido un creciente uso de modelos con efectos aleatorios entre los cuales se encuentran los modelos lineales generalizados mixtos para medidas repetidas (GLMM).

La premisa básica de los modelos lineales generalizados mixtos 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, pudiendo ser continua, binaria o de conteo. Asociado con cada Y_{ij} hay un vector (fila) X_{ij} de covariables de dimen-

sión 1xp, las cuales 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 correspondiente al j-ésimo individuo del i-ésimo grupo, un modelo logístico para Y_{ij} con interceptos aleatorios 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:

$$\ln\left\{\frac{Pr(Y_{ij}=1/b_i)}{1-Pr(Y_{ij}=1/b_i)}\right\}=\eta_{ij}=\mathbf{X}_{ij}\boldsymbol{\beta}+b_i \qquad (2)$$

Es decir la media condicional de Y_{ij} se relaciona con el predictor lineal a través del enlace logit.

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

En general, para los efectos aleatorios se elige la distribución normal por conveniencia en relación con aspectos computacionales y en concordancia con el modelo lineal.

2.2. Los tests

Los tests presentados por Alonso et al. (2008) están basados en la propiedad mostrada por White (1982), según la cual si el modelo está especificado correctamente, se verifica:

$B(\xi_0)+A(\xi_0)=0$, siendo:

$$A(\xi)=E\left\{\frac{\partial^2 \log f(y_i,\xi)}{\partial\xi_k\partial\xi_l}\right\}\ \text{y}\ B(\xi)=E\left\{\frac{\partial \log f(y_i,\xi)}{\partial\xi_k}.\frac{\partial \log f(y_i,\xi)}{\partial\xi_l}\right\}.$$

Por lo tanto, cualquier desviación de los supuestos del modelo se espera distorsionen esta igualdad.

En base a esta relación los autores demostraron que bajo la hipóteis nula de que el modelo está correctamente especificado, se verifica que:

1. $\frac{n}{2p}\left[\delta_{d1}(n)\right]^2 \sim \chi_1^2$,

2. $\frac{n}{2p}\left[\delta_{d2}(n)-1\right]^2 \sim \chi_1^2$,

3. $\frac{\left[n\delta_{dt}(n)\right]^2}{2\sigma_{\delta_n}} \sim \chi_1^2$.

Los elementos de estos tests se definen de la siguiente manera:

$$\delta_{d1}(n) = \log\left|B_n(\xi_0)\left[-A_n^{-1}(\xi_0)\right]\right|,$$

$$\delta_{d2}(n) = \left|B_n(\xi_0)\right|.\left|-A_n^{-1}(\xi_0)\right|,$$

$$\delta_{dt}(n) = \frac{tr\left[B_n(\xi_0)\right]}{tr\left[-A_n(\xi_0)\right]} - \frac{\left|B_n(\xi_0)\right|}{\left|-A_n(\xi_0)\right|},$$

$$A_n(\xi) = \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i,\xi)}{\partial\xi_k \partial\xi_l}\right\},$$

$$B_n(\xi) = \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i,\xi)}{\partial\xi_k}.\frac{\partial \log f(y_i,\xi)}{\partial\xi_l}\right\},$$

k, l =1,..., p con p igual al número de parámetros del modelo y

ξ_0 es el vector de px1 parámetros del modelo cuando no existen errores en la especificación del modelo,

$$\sigma_{\delta_n} = \sum_{k=1}^{p}\left(\frac{\gamma_k}{\sum_l \gamma_l} - 1\right)^2$$

γ_k son los autovalores de la matriz $-A(\xi_0)$.

Los dos primeros tests miden el alejamiento de la igualdad entre $B(\xi_0) = -A(\xi_0)$ usando el determinante de la matriz $B_n(\xi_0)\left[-A_n^{-1}(\xi_0)\right]$, es decir, se utiliza el determinante como una forma de medir la distancia entre dichas matrices. El tercer test combina el determinante con la traza en la misma estadística para cuantificar dicha distancia.

En las aplicaciones prácticas ξ_0 se reemplaza por un estimador consistente bajo la hipótesis nula $\hat{\xi}_n$.

Una de las ventajas de la herramienta propuesta es su relativamente fácil implementación usando los procedimientos NLMIXED e IML de SAS como sugieren los autores en su publicación del año 2008.

3. **EL PROBLEMA**

El lupus sistémico eritematoso (LES) es una enfermedad autoinmune resultante de la conjunción de factores genéticos, hormonales y medioambientales. Varios estudios mostraron una asociación importante entre el desarrollo de LES y características sociodemográficas tales como edad, sexo y nivel de educación. Además se ha demostrado que la actividad de la enfermedad, infecciones, tratamientos y daños orgánicos influyen en el pronóstico de la enfermedad.

Un aspecto que resulta de interés es la demora en el diagnóstico de la enfermedad ya que algunas manifestaciones clínicas tales como las cardiológicas y cutáneas pueden dilatar el diagnóstico provocando daños más severos y mayor mortalidad a largo plazo. Es por ello que resulta importante identificar aquellas manifestaciones clínicas previas al diagnóstico

que influyen en la demora en el diagnóstico de la enfermedad teniendo en cuenta además características sociodemográficas.

El Grupo Latinoamericano de Estudios de Lupus (GLADEL) lleva adelante un estudio de carácter multicéntrico que incluye 34 centros distribuidos en 9 países latinoamericanos, los cuales registran la información sobre características de los pacientes, manifestaciones clínicas presentes durante la enfermedad y tratamientos recibidos, en una base de datos estandarizada.

En este trabajo se ponen a prueba los supuestos acerca de los efectos aleatorios en un modelo logístico normal ajustado a datos proporcionados por GLADEL para identificar posibles factores que influyen en el momento que se realiza el diagnóstico de LES.

4. **APLICACIÓN**

La demora en el diagnóstico se define como el tiempo transcurrido entre el primer síntoma que puede ser asociado con la enfermedad y la fecha del diagnóstico definitivo de la enfermedad. Para un primer estudio los especialistas sugirieron considerar como variable respuesta dicha demora de acuerdo a la siguiente clasificación dicotómica: menor o igual que 6 meses, mayor que 6 meses.

Se consideraron las siguientes variables explicativas: género (masculino, femenino); grupo étnico (blanco, afrolatinoamericano -ALA-, mestizo); edad del paciente al momento del diagnóstico; y la presencia o ausencia de las siguientes manifestaciones clínicas previas al diagnóstico de LES: fiebre, fenómeno de Raynaud, síndrome Sicca y trombosis vascular.

Con el objeto de identificar factores que influyan en la probabilidad de que la demora en el diagnóstico sea mayor a 6 meses, se ajusta un modelo logístico normal, es decir un modelo logit para la respuesta binaria *demora en el diagnóstico de LES* con intercepto aleatorio y las variables explicativas recién definidas. El carácter multicéntrico de la información puede conducir a que la variabilidad de la demora sea mayor entre pacientes de diferentes centros que entre los que asisten a un mismo centro. Esta heterogeneidad se tiene en cuenta a través del modelo específico con efectos aleatorios presentado (Agresti, 2002; Agresti y Hartzel, 2000).

Todas las variables explicativas del modelo son significativas pero no alcanzan a describir la completa variabilidad de la respuesta ya que el efecto aleatorio tiene variancia también significativa (tabla 1).

En relación a los coeficientes del modelo asociados a las manifestaciones clínicas, se observa que la presencia de fiebre favorece el diagnóstico de LES, mientras que el hecho contrario ocurre para las otras características clínicas consideradas.

Tabla 1: Estimaciones de los parámetros del modelo

	Categoría	Estimación	Error estándar	Probabilidad asociada
Intercepto		-1,8323	0,3930	<0,0001
Género(a)	Femenino	0,4140	0,1904	0,0362

Grupo étnico(b)	Mestiza	0,1165	0,1661	0,4874
	ALA	-0,4070	0,2041	0,0536
Edad al diagnóstico	Comp. Lineal	0,0789	0,0192	0,0002
	Comp. Cuadrática	-0,0008	0,0003	0,0041
Fiebre(c)	Si	-0,4327	0,1163	0,0007
Fenómeno de Raynaud(c)	Si	0,5026	0,1419	0,0011
Síndrome de Sicca(c)	Si	0,5991	0,2868	0,0437
Trombosis vascular(c)	Si	0,8098	0,3587	0,0299
Variancia del efecto aleatorio		0,2354	0,0940	

Sin embargo, este trabajo tiene como precupación central la evaluación del enfoque metodológico utilizado, por lo que interesa verificar si la distribución supuesta para los efectos aleatorios no afecta las propiedades de las estimaciones de los parámetros de los efectos fijos.

Se ponen a prueba los tres tests propuestos por Alonso et al. (2008) presentados en la sección 2.2 con los siguientes resultados:

$\delta_{d1}(n) = 3{,}43$, con una probabilidad asociada igual a 0,063,

$\delta_{d2}(n) = 0{,}99$, con una probabilidad asociada igual a 0,321 y

$\delta_{dt}(n) = 0{,}43$, con una probabilidad asociada igual a 0,510.

Salvo el primer test que presenta una probabilidad asociada bastante cercana al nivel del 5%, los otros dos proporcionan evidencia suficiente como para asegurar la distribución normal de los efectos aleatorios.

5. **CONSIDERACIONES FINALES**

En este trabajo se pusieron a prueba tres tests propuestos recientemente en la literatura para detectar la falta de especificación en la estructura de los efectos aleatorios en un modelo lineal generalizado mixto. A pesar de que los resultados hallados no proveen evidencia de una mala especficación, es necesaria cierta cautela al momento de interpretarla. Si bien el tamaño de la muestra no parece conflictivo, estas herramientas de diagnóstico pueden fallar en la detección de algun tipo particular de error en la especificación. Los autores recomiendan por lo tanto completar el estudio considerando distribuciones alternativas para los efectos aleatorios; si se obtiene bajo todas ellas resultados similares es factible incrementar la confianza en ellos mientras que la disparidad provocaría un aumento de la precaución a la hora de realizar conclusiones. Esta tarea está siendo abordada al momento de esta presentación conjuntamente con estudios de simulación ante distintos escenarios a fin de detetectar aquellos más conflictivos.

Paralelamente, los especialistas en LES han sugerido estudiar dicha demora de acuerdo a una clasificación del tiempo en 4 categorías de orden creciente, de manera de abarcar situaciones progresivas de demora desde diagnósticos tempranos hasta muy tardíos. La característica de la variable respuesta orientará el análisis al ajuste de un modelo para respuesta ordinal que también suponga la inclusión de interceptos aleatorios. Ello ha motivado

que se comience a trabajar en la posibilidad de adaptar los tests presentados en esta oportunidad par el caso de modelos ordinales.

4. REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2002). Categorical Data Analysis, 2nd ed. John Wiley & Sons.

Agresti, A.; Caffo, B.; Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduce efficiency and possible remedies. Computational Statistics & Data Analysis, 47: 639-653.

Agresti, A.; Hartzel, J. (2000). Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine*, 19: 1115-1139.

Alonso, A.; Letiere, S.; Molenberghs, G. (2008). A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. Computational Statistics and Data Analysis 52: 4474-4486.

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

Letiere, S.; Alonso, A. A.; Molenberghs, G. (2000). The impact of a misspecified random-effects distribution on the estimation and the performance of the inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27(16): 3125-3144.