X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

MODELOS LINEALES GENERALIZADOS MIXTOS ALGUNOS CASOS PRÁCTICOS

Llorenç Badiella[1]

[1]Servei d'Estadística Aplicada, Univ. Autònoma de Barcelona

RESUMEN

Los modelos mixtos generalizados son una propuesta de análisis estadístico que proporciona un entorno óptimo para responder a les cuestiones de un estudio con diseño experimental más o menos complejo: permite analizar diferentes tipos de variable respuesta (recuentos, proporciones, variables de escala, etc.) modelizando simultáneamente el valor esperado del fenómeno estudiado y su variabilidad. En particular, permite modelizar experimentos en presencia de observaciones correlacionadas. En la actualidad se ha producido un auge en el uso de estas técnicas motivado por las crecientes necesidades de los investigadores de responder a preguntas sofisticadas, además de la creciente eficiencia del software estadístico. En el presente trabajo se repasan diferentes propuestas de modelización, su metodología y algunos detalles técnicos mediante la exploración de diversos ejemplos pertenecientes a los ámbitos de la ecología, medicina y deporte. Los ejemplos tratados son ejemplos reales surgidos de diferentes estudios en los que ha colaborado el Servicio de Estadística Aplicada de la UAB.

Palabras e frases crave: GLMM, factores aleatorios, estructura de correlaciones, aplicaciones.

1. INTRODUCCIÓN

En general, los modelos estadísticos paramétricos se basan en el análisis de relaciones lineales entre cierta variable de interés (la variable respuesta) y ciertas variables de control que contribuyen a explicar su comportamiento (variables explicativas).

Estas características suelen medirse en sujetos, individuos o en general unidades experimentales que son independientes entre sí. Estas relaciones además de ser lineales son evaluadas en términos aditivos. A menudo se asume, de forma muy razonable, que el error no explicado tiene la propiedad de seguir una distribución Normal homogénea.

Podría suceder que la variable objetivo sólo tomara valores en un intervalo o bien que no sea continua, o ni siquiera cuantitativa. De esta manera los errores no

pueden ser Normales. Esta generalización da lugar a los modelos lineales generalizados (GLM). Otra posible generalización consiste en reducir restricciones sobre los errores manteniendo sin embargo la propiedad de Normalidad, es decir, contemplando errores no independientes o heterogéneos. Esta generalización, que permite dotar de estructura a la variabilidad de los errores del modelo, da lugar a los modelos mixtos (MIXED). Los modelos lineales generalizados y los modelos mixtos pueden ser fusionados dando lugar a los modelos lineales generalizados mixtos (GLMM) adaptando las propiedades de ambas propuestas de modelización. Ahora, este enfoque será aplicable en multitud de diseños y ámbitos de conocimiento distintos.

2. MODELOS LINEALES MIXTOS

A continuación se presentan las propuestas de modelización más comúnmente empleadas para analizar una variable objetivo con distribución Normal y observaciones no independientes:

2.1 Modelo para un sólo factor aleatorio

El modelo lineal con un sólo factor aleatorio asumiría que en primer lugar se han seleccionado una serie de niveles a_i de una determinada población heterogénea con cierta variabilidad entre ellas, denotada por σ_A^2. En cada uno de estos niveles preseleccionados a_i se han realizado una serie de mediciones, más homogéneas entre sí. La variabilidad entre estas sub-réplicas (condicionando a las variables explicativas) es σ^2.

La notación estándar empleada en este modelo es la siguiente:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + ... + \beta_p X_{pij} + a_i + \epsilon_{ij} \tag{1}$$

donde se asume que $\epsilon_{ij} \sim N(0, \sigma^2)$ y $a_i \sim N(0, \sigma_A^2)$.

2.2 Modelo de *intercepts* y *slopes* aleatorios

Este modelo es adecuado para analizar estudios con datos longitudinales. Cada nivel del factor aleatorio (habitualmente individuos) son analizados en determinadas ocasiones a lo largo del tiempo, aunque a menudo, de forma desestructurada. Se asume que la variable objetivo puede describirse mediante un patrón de evolución (una recta) poblacional (efectos fijos) y un patrón individualizado (efectos aleatorios). Su formulación es la siguiente:

$$Y_{ij} = \beta_0 + a_i + (\beta_1 + b_i) X_{1ij} + \epsilon_{ij} \tag{2}$$

donde se asume que $\epsilon_j \sim N(0, \sigma^2)$, $a_i \sim N(0, \sigma_A^2)$ y $b_i \sim N(0, \sigma_B^2)$.

2.3 Modelos de estructuras de correlaciones

Los modelos mixtos planteados de forma marginal y empleando notación matricial corresponden a:

$$Y \sim N(X\beta, R) \tag{3}$$

donde R es una matriz normalmente definida por bloques que pueden tener estructuras muy peculiares: sin estructura, con simetría compuesta, AR(1), espacial, etc. e incluso podrían contemplar variabilidad heterogénea. Cada bloque corresponde a un individuo o nivel del factor aleatorio.

En este sentido, existen dos versiones diferentes del mismo planteamiento: el modelo mixto de efectos aleatorios (modelo tipo G) y el modelo mixto de estructura de correlaciones (modelo tipo R).

3. MODELOS LINEALES MIXTOS GENERALIZADOS

El modelo lineal generalizado permite considerar distribuciones para la variable respuesta entre aquéllas que pertenecen a la familia de distribuciones exponencial de un parámetro.

Con el objetivo de especificar adecuadamente el modelo surgen de nuevo los dos enfoques distintos presentados anteriormente: Modelo marginal (tipo G) o Modelo condicional (tipo G). Ahora, los métodos de ajuste e interpretación van a diferir substancialmente, dando lugar a planteamientos parecidos pero no equivalentes.

El modelo marginal (R) se concreta del modo siguiente:

$$\begin{aligned} g(E[Y]) &= X\beta \\ \text{Var}[Y] &= V(\mu)\phi \\ \text{Corr}[Y_{ij}, Y_{ij'}] &= \alpha_{jj'} \end{aligned} \tag{4}$$

donde g es cierta función de enlace que permite trabajar de forma lineal el valor esperado, $V(\mu)$ corresponde a la función varianza asociada a la propia distribución, ϕ un parámetro de escala complementario y finalmente, la estructura de correlaciones se define a partir de los parámetros $\alpha_{jj'}$.

El modelo condicional (G) es el siguiente:

$$\begin{aligned} \text{E}[Y|u] &= g^{-1}(X\beta + Zu) \\ u &\sim N(0, G) \end{aligned} \tag{5}$$

donde se asume que condicional a los efectos aleatorios se trata de un modelo lineal generalizado simple y que la distribución de los efectos aleatorios u es Normal, suposición normalmente muy razonable.

4. APLICACIONES

4.1 Modelo Mixto: Ansiolítico

Un investigador ha seleccionado 10 individuos, que han sido sometidos a distintas pruebas para medir su ansiedad. En cuatro periodos diferentes se evalúan 4 dosis de un mismo tratamiento ansiolítico. En cada periodo, el efecto del tratamiento es medido en dos condiciones diferentes: una situación control y una situación de ansiedad provocada. La variable respuesta consiste en una medida eléctrica del sobresalto

producido por ciertos estímulos. El investigador desea evaluar el efecto de la dosis en ambas condiciones experimentales (ver figura 1). Será necesario considerar una determinada estructura de correlaciones puesto que existen medidas repetidas.

4.2 Datos longitudinales: Pino Rojo

Se pretende comparar el crecimiento del pino rojo según diferentes períodos. 35 parcelas fueron seleccionadas, y en cada parcela se escogieron un número de árboles cuyas fechas de nacimiento pertenecieran a los intervalos de tiempo 1900-1925, 1926-1950, 1951-1975, 1976-1990. El crecimiento fue evaluado examinando la amplitud de los diferentes anillos año a año (ver figura 2). Se desea comparar el patrón de crecimiento según el periodo de nacimiento.

4.3 Recuentos longitudinales: Brotes clínicos

Se desea analizar el patrón de evolución de una patología auto-inmune. La información de los pacientes es recogida semestralmente. La variable objetivo consiste en el número de brotes clínicos experimentados desde la última visita realizada. Se selecciona únicamente a aquellos pacientes con al menos 3 visitas realizadas, en total se incluyen 277 sujetos (ver figura 3). Finalmente, se desea analizar la influencia de la edad en el momento de diagnóstico o del sexo en la severidad de la enfermedad.

4.4 Modelo GLMM Recuentos: Combates Taekwondo

En el presente estudio, perteneciente al ámbito del Taekwondo, se desea medir y analizar los factores que influyen en la cantidad de movimientos ofensivos realizados por los contrincantes así como su éxito. Se dispone de datos correspondientes a 32 combates diferentes pertenecientes a distintas categorías de peso. En total, se dispone de 2246 movimientos ofensivos. Los datos han sido agrupados según intervalos de tiempo de 1 minuto de duración obteniendo dos variables objetivo: el recuento de movimientos ofensivos realizados por minuto y el recuento de movimientos con éxito por minuto (ver figura 4).

5. REFERENCIAS

Lee. Y., Nelder, J.A. (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society Series B*, **58**, 619-656.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*, Chapman & Hall/CRC Boca Raton.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Second Ed. Chapman and Hall/CRC, London.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*, SAS Institute Inc..

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Zuur, A.F. et al. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer.

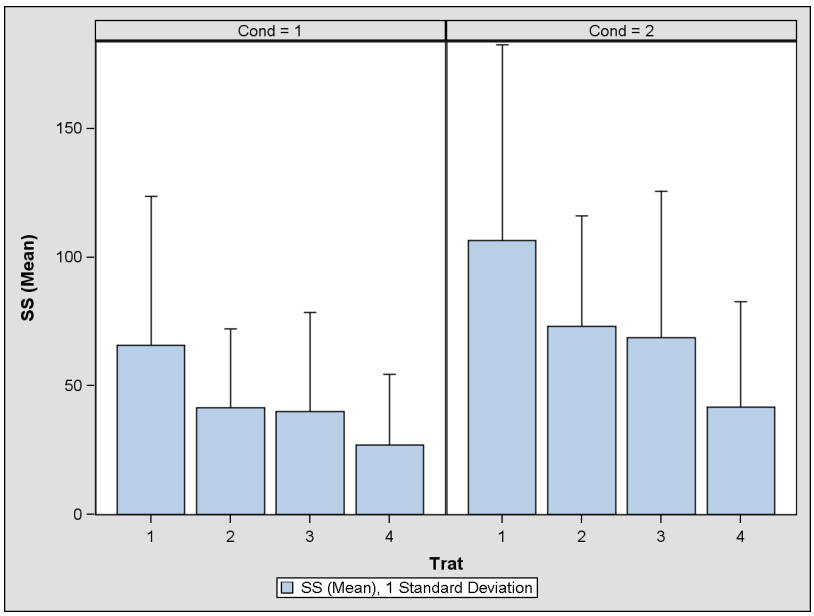

Figure 1: Estudio Ansiolíticos

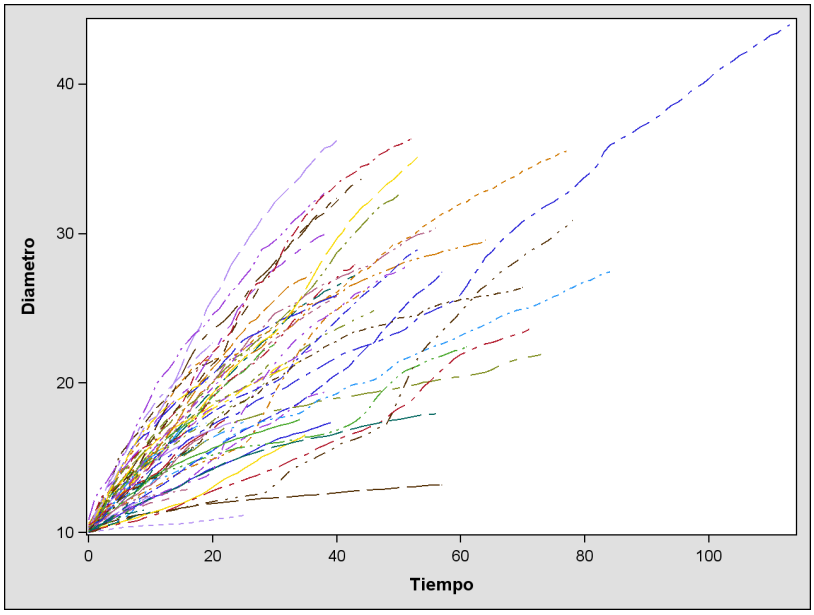

Figure 2: Estudio Pino Rojo

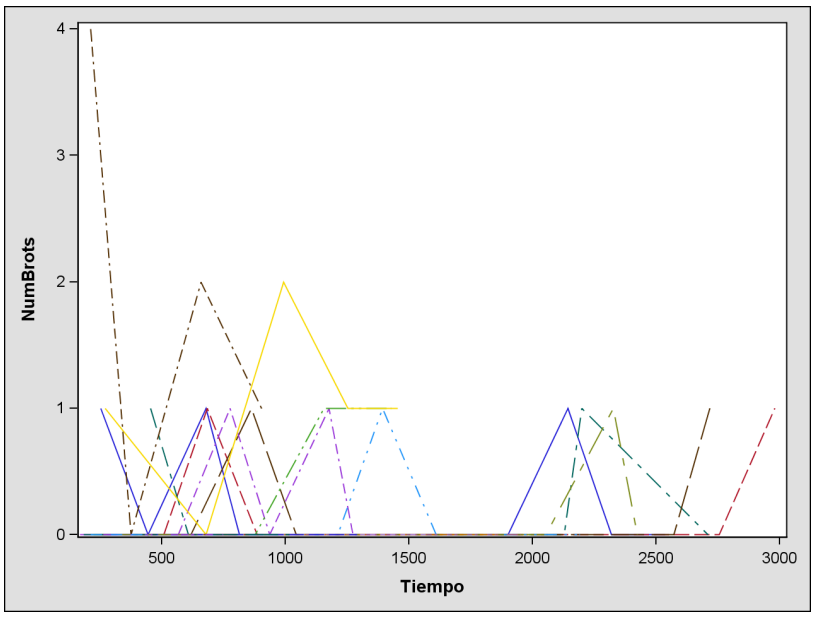

Figure 3: Estudio Brotes Clínicos

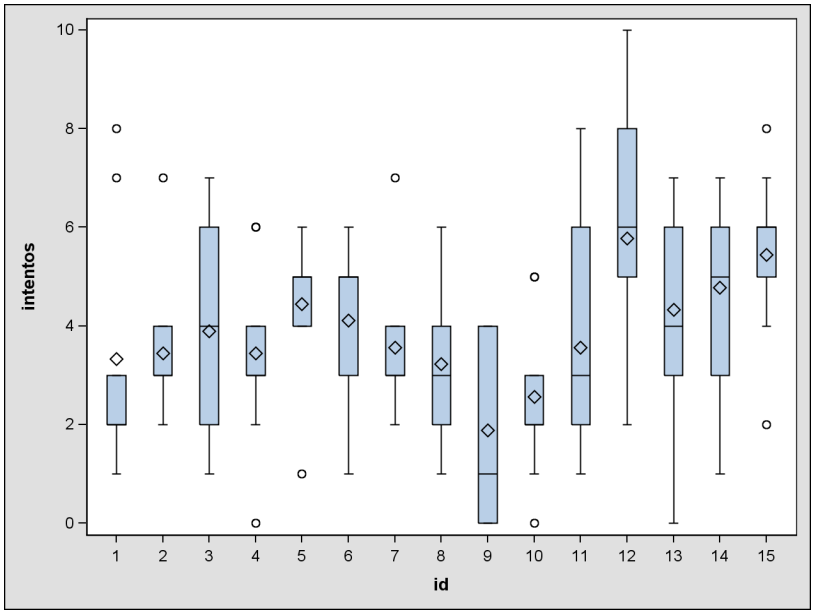

Figure 4: Combates Taekwondo