Métodos para la evaluación de los modelos lineales y no lineales mixtos para datos longitudinales

María del Carmen Garcia[12], Elsa Servy,[1] Liliana Koegel [1] ,Cecilia Rapelli[1]

(1): Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. (2) Consejo de Investigaciones Universidad Nacional de Rosario.

OBJETIVO

Los estudios longitudinales, en los cuales se efectúan mediciones repetidas a diferentes unidades en el tiempo, juegan un rol importante en investigación. Los modelos lineales y no lineales mixtos resultan adecuados para el análisis de este tipo de datos. La aplicación de los modelos mixtos requiere la identificación de un modelo parco para describir la relación entre las variables. Inicialmente, se especifican los efectos fijos y aleatorios y la estructura de covariancias y los supuestos sobre la estructura básica de los datos. Posteriormente se estiman de la forma más eficiente posible los parámetros y se comprueba si los supuestos iniciales se cumplen. Si el modelo está correctamente especificado los estimadores de los parámetros tienen propiedades deseables y se obtienen inferencias válidas.
Este proyecto está centrado en el propósito de presentar y discutir varios métodos para evaluar y validar el ajuste de modelos lineales y no lineales mixtos para datos longitudinales y aplicarlos a datos biológicos, sociales y económicos.

RESULTADOS OBTENIDOS

En esta etapa de la investigación se utilizó para los modelos lineales mixtos un procedimiento de diagnóstico, el análisis de influencia loca,l que permite identificar las unidades atípicas y evaluar su efecto sobre las distintas componentes del modelo. Para los modelos no lineales mixtos se compararon dos enfoques de modelación y procedimientos de estimación. Se obtuvieron dos trabajos.

ANÁLISIS DE INFLUENCIA LOCAL PARA LA EVALUACIÓN DE UN MODELO LINEAL MIXTO

Los modelos mixtos se estiman por métodos basados en la función de máxima verosimilitud que son sensibles a unidades atípicas., que pueden tener una gran influencia sobre los resultados del análisis.
La sensibilidad de un modelo se estudia por medio de medidas que expresan la estabilidad del mismo bajo perturbaciones. El desplazamiento de la verosimilitud (LD) es la medida de influencia más útil que permite evaluar los efectos de perturbaciones sobre los parámetros estimados. Existen varios enfoques para comprobar la influencia en un análisis estadístico. Uno de ellos, los diagnósticos de omisión de casos, que encuadran dentro del análisis de influencia global, evalúan el efecto de una unidad sacándola del conjunto de datos. Otro enfoque, la influencia local, estudia el comportamiento local del desplazamiento de la verosimilitud. Sea el vector $\mathbf{Y}_i$ de n_i respuestas de la i-ésima unidad, i=1,...,N, y una matriz $\mathbf{X}_i$ de p predictores y sus coeficientes $\boldsymbol{\beta}$, el modelo lineal mixto se expresa

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\,\mathbf{b}_i + \mathbf{e}_i$$

$$\mathbf{e}_i \overset{iid}{\sim} N_{ni}(\mathbf{0};\mathbf{R}_i) \qquad \mathbf{R}_i = \sigma^2\,\mathbf{I} \qquad \mathbf{b}_i \overset{iid}{\sim} N_k(\mathbf{0},\mathbf{D})$$

La influencia local estudia el comportamiento local de la función LD(ω) alrededor de un valor de perturbación de interés**.** Lesaffre y Verbeke (1997) mostraron que este método es útil para la detección de casos influyentes en datos longitudinales y presentaron un esquema de perturbación alternativo en el contexto de los modelos lineales mixtos, que permite investigar cuánto están afectados los estimadores por cambios en los pesos de las contribuciones individuales al logaritmo de la verosimilitud. Esto produce la denominada influencia local del i-ésimo sujeto (C_i) sobre la estimación de los parámetros, que mide el efecto de dar un peso levemente distinto al individuo i, mientras que los pesos de los otros individuos permanecen constantes.
La medida de influencia C_i se expresa en función de componentes interpretables, dependen de
1. las covariables, $\mathbf{X}_i$, para la estructura media
2. un vector de residuos para la estructura media del modelo.
3. las covariables $\mathbf{Z}_i$ para la estructura de covariancia
4. una matriz de residuos para la estructura de covariancias para el sujeto i y
5. una medida de la variabilidad de las mediciones del sujeto i.

Las dos primeras permiten detectar influencia sobre los efectos fijos y las restantes sobre las componentes de variancia.
El cálculo de las medidas de influencia y sus componentes interpretables se realiza utilizando una macro de SAS.

APLICACION

El método se utiliza para evaluar la seguridad cardiológica de una droga en 48 pacientes que fueron asignados a cinco tratamientos, cuatro de ellos consisten en tomar diariamente diferentes dosis de la droga (Grupos 1 a 4) y el otro, placebo (grupo 0). A cada paciente se realiza un electrocardiograma en 7 oportunidades: antes de recibir la primera dosis, dos horas después de haber recibido la primera dosis, luego uno diariamente durante4 días y, por último, uno 2 días después de haber finalizado el tratamiento. Se registra una medida de interés, la longitud del intervalo QTc, con el fin de comprobar si la droga prolonga la longitud del intervalo.
Para este estudio se propuso un modelo lineal mixto con un efecto aleatorio.
A partir de los resultados obtenidos se puede concluir que la principal ventaja de C_i es que permite encontrar para los individuos atípicos una explicación de las causas por las cuales resultan influyentes. Debido a que el vector de residuos para la estructura media y la matriz de residuos para la covariancias de los sujetos 18 y 40 muestran valores grandes indican que los perfiles observados y la estructuras de covariancias para ambos sujetos están pobremente predichos por los correspondientes elementos del modelo postulado.

MODELOS MARGINALES Y CONDICIONALES PARA CURVAS DE CRECIMIENTO

Los estudios de crecimiento, en los cuales se miden repetidamente a diferentes individuos, consideran que la forma en que la respuesta media cambia en el tiempo es no lineal.
Para datos longitudinales, la correlación entre las mediciones repetidas se puede modelar explícitamente, postulando una matriz de covariancias para las mismas, o implícitamente, mediante la introducción de efectos aleatorios al modelo. La primera forma de modelar conduce a los modelos marginales (o promedio poblacional), que centran el interés en los parámetros que describen las medias marginales tratando a los de la estructura de covariancias como parámetros de ruido. Los modelos condicionales, denominados modelos mixtos o específicos del sujeto, incluyen efectos aleatorios para explicar la correlación. Si bien en los modelos lineales ambos enfoques son equivalentes, para los no lineales no sucede lo mismo.
Siendo f una función no lineal conocida y $\mathbf{g}(.,.,.)$ posiblemente no lineal, se expresan los modelos

Condicional

$$\mathbf{Y}_i = \boldsymbol{f}(\mathbf{X}_i,\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i$$
$$\boldsymbol{\beta}_i = \mathbf{g}(\mathbf{a}_i,\boldsymbol{\beta},\mathbf{b}_i)$$
$$\mathbf{b}_i \sim N_v(\mathbf{0},\boldsymbol{\Psi}) \qquad \boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0},\boldsymbol{\Lambda}_i(\boldsymbol{\gamma}))$$

Marginal

$$\mathbf{Y}_i = \boldsymbol{f}(\mathbf{X}_i,\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i$$
$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0},\boldsymbol{\Lambda}_i(\boldsymbol{\gamma}))$$

Debido a que los efectos aleatorios no son observables, la estimación máximo verosímil y la inferencia para estos modelos se basa sobre la densidad marginal de las respuestas, calculada como,

$$\ell(\boldsymbol{\beta},\boldsymbol{\Psi},\boldsymbol{\gamma}/\mathbf{Y}) = \int p(\mathbf{Y}/\mathbf{b},\boldsymbol{\beta},\boldsymbol{\gamma})\; p(\mathbf{b}/\boldsymbol{\Psi})\; d\mathbf{b}$$

Como la función f es no lineal en los efectos aleatorios, la integral generalmente no tiene una solución analítica simple, por lo que algunos autores aproximan el integrando con expansiones en serie de Taylor de primer orden alrededor de los estimadores de los efectos aleatorios ($\hat{\mathbf{b}}$), mientras que los otros lo hacen alrededor de $\mathbf{b}_i=\mathbf{0}$. .

APLICACION

Se evalúan los pesos de ratones hembras y machos pertenecientes a dos cohortes. Cada una de las cohortes está constituida por 50 ratones, 25 hembras y 25 machos, a los cuales se les midió el peso en 11 ocasiones. La expresión del modelo usado es

$$Y_{ij} = \beta_{0i}\exp(-\beta_{1i}\exp(-\beta_{2i}t_j)) + \varepsilon_{ij}$$
$$\beta_{0i} = \beta_0 + \beta_{01}S_i + b_{0i}$$
$$\beta_{1i} = \beta_1 + \beta_{11}S_i$$
$$\beta_{2i} = \beta_2 + b_{2i}$$

Tabla 1 Estimación efectos fijos y errores estándares modelo modelo condicional mediante dos métodos de linealización

MODELO CONDICIONAL						
Efectos fijos	Método de primer orden SS			Método de primer orden PA		
	Estimación	Desv.Est.	t	Estimación	Desv.Est.	t
β_0	35.6398	0.3625	98.31*	35.5966	0.3601	98.85*
β_{01}	6.3816	0.4804	13.28*	6.4067	0.4778	13,41*
β_1	3.0404	0.0127	240.11*	3.0352	0.0126	241,05*
β_{11}	0.1741	0.0176	9.89*	0.1751	0.0175	10,02*
β_2	0.0567	0.0004	135.43*	0.0565	0.0004	133.97*

Tabla 2 Estimación efectos fijos y errores estándares modelo marginal .Estructuras simetría compuesta y arbitraria heterogénea

MODELO MARGINAL						
Efectos fijos	Estructura CSH			Estructura UN		
	Estimación	Desv.Est.	t	Estimación	Desv.Est.	T
β_0	34.6189	0.2709	127.78*	34.9147	0.3493	100.24*
β_{01}	6.3015	0.3828	16.46*	5.5624	0.4707	11.82*
β_1	3.0596	0.0117	260.53*	3.0340	0.0137	220.73*
β_{11}	0.1828	0.0170	10.74*	0.1322	0.0187	7.07*
β_2	0.0560	0.0003	187.34*	0.0590	0.0003	170.84*

Quizás, la diferencia más importante entre los dos enfoques se refiere al objetivo que se persiga con el estudio. Si el interés está sólo en la estimación de los efectos fijos ambos enfoques producen resultados bastante parecidos, dependiendo de la elección de los efectos aleatorios y de la covariancia intra individuo. Sin embargo, si el interés está en predecir los perfiles individuales se debería usar el enfoque condicional. A partir de los resultados obtenidos se considera que el modelo básico que se debe utilizar es el condicional ya que conduce a un modelo marginal específico. Es preferible trabajar con un modelo condicional y el modelo marginal derivado de él. el método condicional de primer orden se debe preferir para estimar los parámetros del modelo