Garcia, María del Carmen
Rapelli, Cecilia
Castellana, Noelia
Koegel, Liliana
Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

DIAGNÓSTICOS DE INFLUENCIA PARA LOS MODELOS LINEALES MIXTOS

Resumen
Los modelos lineales mixtos son apropiados para la modelación de datos longitudinales. La estimación de los parámetros de estos modelos se realiza utilizando métodos basados en la función de verosimilitud que son sensibles a unidades atípicas. El análisis de influencia tiene por objetivo determinar las unidades y/o las observaciones que producen excesiva influencia en los parámetros estimados, de manera que permitan desarrollar un modelo más adecuado. El estudio se realiza introduciendo cambios en las componentes del modelo y evaluando si se producen cambios importantes en los resultados. Uno de los enfoques para evaluar la influencia es el diagnóstico de omisión de casos, que encuadra dentro del análisis de influencia global, y evalúa el efecto de una unidad eliminándola del conjunto de datos. Otro enfoque, la influencia local, investiga el efecto que produce sobre la estimación de los parámetros la introducción de pequeñas perturbaciones en las componentes del modelo. Su uso permite encontrar las causas por las cuales las unidades atípicas resultan influyentes. Un método diagnóstico de reciente aparición, los gráficos de las sumas de cuadrados de los residuos estudentizados, permite detectar unidades atípicas sin omitirlas. En este trabajo se utilizan en forma comparativa esos enfoques, aplicándolos a datos provenientes de un estudio clínico realizado para evaluar la seguridad cardiológica de una nueva droga.

Palabras claves: Datos longitudinales. Modelos lineales mixtos. Análisis de influencia

A**bstrac** Mixed linear models are suitable for modeling longitudinal data. The parameter estimation of these models is performed using methods that are based on the likelihood function which are sensitive to unusual units. The influence analysis aims to detect observations/units that may produce excessive influence in the parameters estimates, in order to develop a more suitable model. The analysis is performed by introducing changes to the model components and assessing whether significant changes in the results are produced. One approach to assess the influence is the deletion case diagnosis that evaluates the effect of a unit, removing it from the dataset. This technique is considered as a global influence analysis. Another approach, the local influence, investigates the effect of introducing small perturbations in the model components on the parameter estimation. Its usage allows determining the causes for which atypical units are influential. A new diagnostic method based on studentized residual sum of squares plots allows the detection of discordant units without omitting them. In this paper, these approaches are compared considering data from a clinical trial which was designed to evaluate the cardiac safety of a new drug.

Keywords*:* Longitudinal data. Mixed linear models. Influence analysis

1. Introducción

Los conjuntos de datos longitudinales pueden contener unidades con una magnitud inusual Los modelos mixtos constituyen una valiosa herramienta para analizar datos longitudinales. La estimación de los parámetros de los modelos mixtos se realiza mediante el uso de métodos basados en la función de verosimilitud, que son sensibles a estas unidades atípicas. Los analistas deben ser cuidadosos ante la presencia de estos datos discordantes, pues pueden tener una influencia grande sobre los resultados del análisis. Un estudio de los mismos puede llevar a concluir que tales casos son completamente apropiados y deben ser retenidos en el análisis o puede sugerir la necesidad de obtener datos adicionales o que el modelo no sea adecuado. Una investigación de los casos influyentes es sólo posible una vez que ellos se hayan identificado.

Este trabajo presenta una comparación de métodos utilizados para comprobar la influencia de casos atípicos en el contexto de los modelos lineales mixtos. Para estudiar su comportamiento se utilizan datos provenientes de un estudio clínico desarrollado para evaluar la seguridad cardiológica de una droga.

2. Modelos lineales mixtos

En los estudios longitudinales las unidades (individuos o casos) se observan repetidamente en varias ocasiones. Los modelos lineales mixtos que contienen efectos fijos y aleatorios se utilizan para el análisis de este tipo de datos.

En estos modelos la respuesta media se expresa como combinación de características poblacionales, que son compartidas por todas las unidades y efectos específicos de la unidad que son propios de la misma. Los primeros se denominan efectos fijos, mientras que los últimos aleatorios. El modelo lineal mixto se expresa como,

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i, \quad i=1,\ldots,N, \tag{2.1}$$

donde, $\mathbf{Y}_i$ es un vector (n_i x 1) que contiene las respuestas de la i-ésima unidad, $\mathbf{X}_i$ es una matriz (n_i x p) para los efectos fijos, $\boldsymbol{\beta}$ es un vector (p x 1) de parámetros de efectos fijos, $\mathbf{Z}_i$ es una matriz (n_i x k) "diseño" para los efectos aleatorios, $\mathbf{b}_i$ es un vector de efectos aleatorios (k x 1) y $\mathbf{e}_i$ es un vector (n_i x p) de errores dentro de cada unidad.

Se asume que los vectores $\mathbf{e}_i$ y $\mathbf{b}_i$ son independientes y con distribución,

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

donde, $\mathbf{R_i}$ y $\mathbf{D}$ denotan matrices de covariancias de respectiva dimensión (n_i x n_i) y (k x k).

La estimación de los parámetros se realiza minimizando la función objetivo, menos dos veces el logaritmo de la función de verosimilitud (-2ℓ), mediante el algoritmo de Newton-Raphson. Los estimadores de los efectos fijos y aleatorios son, respectivamente, $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X_i}' \hat{\mathbf{V}}_\mathbf{i}^{-1}(\boldsymbol{\theta}) \mathbf{X_i} \right)^{-1} \sum_{i=1}^{N} \mathbf{X_i}' \hat{\mathbf{V}}_\mathbf{i}^{-1}(\boldsymbol{\theta}) \mathbf{Y_i}$ y $\hat{\mathbf{b}}_\mathbf{i} = \hat{\mathbf{D}} \mathbf{Z_i}' \mathbf{V_i}^{-1}(\boldsymbol{\theta})(\mathbf{Y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}})$, siendo $\mathrm{Var}(\mathbf{Y_i}) = \mathbf{Z_i}\ \mathbf{D}\ \mathbf{Z'_i} + \mathbf{R_i} = \mathbf{V_i}(\boldsymbol{\theta})$ y $\boldsymbol{\theta}$ un vector que contiene a los parámetros de covariancia.

Los métodos de estimación basados en la función de verosimilitud son sensibles a unidades atípicas y su presencia puede tener una gran influencia sobre los resultados del análisis.

3. Análisis de influencia

La evaluación cualitativa y cuantitativa de la influencia de unidades sobre el análisis se denomina análisis de influencia. Este análisis tiene por objetivo determinar las unidades y/o las observaciones que producen excesiva influencia en los parámetros estimados, de manera que permitan desarrollar un modelo más adecuado. El estudio se realiza introduciendo cambios en los datos o en las componentes del modelo y evaluando si se producen cambios importantes en los resultados.

La mayoría de los métodos actualmente disponibles para detectar unidades y observaciones discordantes son generalizaciones de los enfoques para datos univariados, basados en la omisión de los mismos. La influencia global considera los cambios en los parámetros poblacionales, no tomando en cuenta los parámetros específicos de la unidad. Si todas las observaciones de la unidad se excluyen es improbable que el efecto de esa unidad sobre sus parámetros específicos sea visto. Una dificultad que surge con estos enfoques es determinar cuando los cambios son suficientemente grandes como para realizar posteriores investigaciones, reformulación del modelo o eliminación de los datos.

La influencia local es otro método para detectar unidades influyentes midiendo los cambios en la función de verosimilitud con el cambio de pesos en las unidades. Resulta útil para investigar las causas de las desviaciones, pero se debe usar con precaución pues pueden fallar en la detección.

Recientemente se propuso un nuevo método que no elimina la unidad, denominado gráfico de la suma de cuadrados de los residuos estudentizados (TRSS) (Mun y Lindstrom, 2013). Como una aplicación de los gráficos TRSS, estos autores sugieren, también, un método de

eliminación de observaciones que detecta observaciones discordantes. El método propuesto proporciona una mayor información mediante la utilización de residuos modificados y evalúa eficazmente el efecto de unidades y observaciones discordantes en la estimación de parámetros que incluyen componentes de la variancia.

3.1. Influencia global

Una forma de verificar la influencia de un grupo de observaciones es omitir el grupo y observar los cambios en los estimadores. Si se producen grandes cambios el grupo es influyente.

Para los modelos lineales mixtos, estimados por máxima verosimilitud (ML) o máxima verosimilitud restringida (REML), una medida general es la distancia de verosimilitud (Cook y Weisberg, 1982), también llamada desplazamiento de la verosimilitud (Beckman, Nachtseim y Cook, 1987). Para construir esta estadística se estiman los parámetros con el conjunto de datos completos, representados por el vector $\hat{\boldsymbol{\psi}}$, y con el reducido ($\hat{\boldsymbol{\psi}}_{(U)}$), obteniéndose las distancias de verosimilitud y verosimilitud reducida, respectivamente, como

$$\begin{aligned} \mathrm{LD}_{(U)} &= 2[\ell(\hat{\boldsymbol{\psi}}) - \ell(\hat{\boldsymbol{\psi}}_{(U)})] \\ \mathrm{RLD}_{(U)} &= 2[\ell_R(\hat{\boldsymbol{\psi}}) - \ell_R(\hat{\boldsymbol{\psi}}_{(U)})] \end{aligned}.$$

Esta distancia suministra la magnitud del cambio que se produce en el logaritmo de la verosimilitud (ℓ) cuando es evaluada en los estimadores de los parámetros del conjunto de datos reducidos. Es decir, proporciona la magnitud por la cual la verosimilitud de los datos completo cambiaría si se utilizara un estimador basado sobre menos datos.

La distancia de verosimilitud es una medida global que expresa la influencia conjunta de las observaciones en el conjunto U sobre todos los parámetros en $\boldsymbol{\psi}$. Si esta medida sugiere que existen unidades influyentes se tendrían que determinar, a posteriori, los elementos del modelo que resultan influenciados.

Una forma de medir el impacto sobre el vector de los parámetros de efectos fijos y covariancia estimados es calcular una estadística a partir de las diferencias entre los estimadores de los parámetros con los datos completos y reducidos. Una estadística de este tipo es la distancia de Cook (D de Cook) que se expresa como,

$$\mathrm{D}(\boldsymbol{\beta}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)' \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_U)}{\mathrm{rg}(\mathbf{X})} \quad , \quad \mathrm{D}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_U)' \hat{\boldsymbol{\Gamma}}^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_U),$$

donde, el subíndice U denota el vector de estimadores después de eliminar las

observaciones en el conjunto U, $\text{Var}(\hat{\boldsymbol{\beta}})^{-1}$ la inversa de esa matriz de covariancias y $\boldsymbol{\Gamma}$ la matriz de covariancias asintótica de $\hat{\boldsymbol{\theta}}$. Cuanto más grande sea esta estadística mayor es la influencia. Los efectos sobre la precisión de los estimadores se separan del efecto de los estimadores puntuales. Los casos que tienen valores chicos de la distancia de Cook, por ejemplo, pueden afectar las pruebas de hipótesis e intervalos de confianza si su influencia sobre la precisión de los estimadores es grande.

Las estadísticas que se usan para evaluar el cambio en la precisión involucran el determinante de las matrices de covariancias y se puede calcular tanto para los efectos fijos como para los parámetros de covariancia ($\boldsymbol{\theta}$),

$$\text{COVRATIO}(\boldsymbol{\beta}) = \frac{\left|\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_U)\right|}{\left|\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\right|} \qquad \text{COVRATIO}(\boldsymbol{\theta}) = \frac{\left|\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_U)\right|}{\left|\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})\right|}.$$

Como la COVRATIO relaciona los determinantes de las matrices de covariancias de los estimadores de los parámetros de los modelos reducidos y completos el valor uno (1) indica que el caso no es influyente. Valores más grandes que uno indican mayor precisión en el conjunto completo.

3.2. Influencia local

Otro método usado para detectar observaciones influyentes, la influencia local, mide los cambios en la función de log verosimilitud asignando diferentes pesos a las unidades y resulta útil para investigar las fuentes de las desviaciones.

Este método, al igual que el anterior, utiliza la distancia de verosimilitud para encontrar casos influyentes. Cook (1986) propone estudiar el comportamiento local del desplazamiento de la verosimilitud usando la curvatura normal C_i. Lesaffre y Verbeke (1998) derivan C_i en la dirección de un vector que contiene un uno (1) en la posición i-ésima y 0 en las otras posiciones, correspondiendo a una perturbación del modelo postulado, llamada influencia local total del individuo i. Valores grandes de C_i indican que la observación es influyente. Sin embargo, éstas no indican las razones por las cuales algunos individuos son más influyentes que otros y por lo tanto limitan el valor diagnóstico.

Para remediar este problema, descompusieron C_i en componentes interpretables, $\mathbf{C}_i(\boldsymbol{\beta})$ y $\mathbf{C}_i(\mathbf{D},\sigma)$, que permiten encontrar una explicación parcial para el carácter influyente de un individuo. La primera mide la influencia sobre los efectos fijos y la otra sobre los parámetros de covariancia. Además, mostraron, a partir de la independencia asintótica de los efectos

fijos y componentes de variancia en los modelos lineales mixtos, que asintóticamente $C_i = C_i(\boldsymbol{\beta}) + C_i(\mathbf{D}, \sigma)$. Esto significa que la influencia local para los efectos fijos es independiente de la influencia local para las componentes de variancia y que su suma iguala a la influencia local total, es decir, para todos los parámetros simultáneamente. La medida C_i contiene cinco componentes interpretables que son funciones de los elementos del modelo, $\|\boldsymbol{X}_i \boldsymbol{X}_i'\|$, $\|\boldsymbol{R}_i\|^2$, $\|\boldsymbol{Z}_i \boldsymbol{Z}_i'\|^2$, $\|\mathbf{I} - \boldsymbol{R}_i \boldsymbol{R}_i'\|^2$ y $\|\mathbf{V}_i^{-1}\|^2$ siendo, $\boldsymbol{R}_i = \mathbf{V}_i^{-1/2}\, \mathbf{r}_i$, $\mathbf{r}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$, $\boldsymbol{X}_i = \mathbf{V}_i^{-1/2}\, \mathbf{X}_i$, $\boldsymbol{Z}_i = \mathbf{V}_i^{-1/2}\, \mathbf{Z}_i$, $\|\boldsymbol{X}_i \boldsymbol{X}_i'\|$ la longitud de las covariables estandarizadas para los efectos fijos y $\|\boldsymbol{R}_i\|^2$ la longitud al cuadrado de los residuos.

Para muestras grandes, $C_i(\boldsymbol{\beta})$ se puede descomponer usando solo las dos primeras componentes de las cinco mencionadas y el resto corresponden a $C_i(\mathbf{D}, \sigma)$.

Cuando C_i es grande debido a que $C_i(\boldsymbol{\beta})$ es grande la influencia de ese individuo se puede atribuir a que alguna o ambas partes sea grande. En ese caso el i-ésimo individuo no está bien ajustado o predicho por el modelo y tiene un vector de covariables grande. De manera similar grandes valores de $\|\boldsymbol{Z}_i \boldsymbol{Z}_i'\|^2$ y/o $\|\mathbf{I} - \boldsymbol{R}_i \boldsymbol{R}_i'\|^2$ implican $C_i(\mathbf{D}, \sigma)$ grande. El término $\|\mathbf{I} - \boldsymbol{R}_i \boldsymbol{R}_i'\|^2$ es cero si $\mathbf{V}_i$ es igual a $\mathbf{r}_i \mathbf{r}_i'$, que es un estimador de var($\mathbf{Y}_i$) solo si la media está correctamente modelada como $\mathbf{X}_i\boldsymbol{\beta}$. Entonces a este término se lo puede considerar como un residuo que mide cuan bien la estructura de covariancia de los datos es modelada por $\mathbf{V}_i(\boldsymbol{\theta}) = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \sigma^2 \mathbf{I}_{ni}$. El valor $\|\mathbf{V}_i^{-1}\|^2$ grande indica que el sujeto i tiene poca variabilidad.

Se procede a identificar unidades influyentes realizando gráficos de las componentes C_i, $C_i(\boldsymbol{\beta})$ y $C_i(\mathbf{D}, \sigma)$ vs el número de la unidad y comparando, cuando N es grande, con los valores de referencia $2\sum C_i / N$, $2\sum C_i(\boldsymbol{\beta}) / N$ y $2\sum C_i(\mathbf{D}, \sigma) / N$.

3.3. Gráficos de las sumas de cuadrados de los residuos estudentizados

Los métodos anteriores se focalizan sólo en los cambios de los coeficientes, no teniendo en cuenta la trayectoria de los casos atípicos.

Esta nueva herramienta diagnóstica (Mun y Lindstrom, 2013) introduce una leve modificación a la expresión de los residuos (residuos modificados) y construye la suma de cuadrados de estos residuos. Se consideran dos tipos de desviaciones que se pueden examinar simultáneamente. Considerando el concepto que una unidad influyente está alejada de su media, se define un valor para medir la desviación entre la media específica de una unidad y la media poblacional y otro que considera la distancia entre una trayectoria

individual y su media específica. La primera se denomina desviación tipo L (posición) y la segunda tipo S (forma).

Un residuo, definido como la diferencia entre la respuesta y la respuesta media estimada, se descompone como la suma de $e_{i.1}$ y $e_{i.0}$, siendo,

$e_{i.1}$ =(respuesta – respuesta media de la unidad) = $\mathbf{Y}_i - (\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{b}}_i)$

$e_{i.0}$ =(respuesta media de la unidad – respuesta media) = $\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{b}}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$

Los vectores $e_{i.0}$ y $e_{i.1}$ se pueden expresar como combinaciones lineales de los vectores

$$\boldsymbol{\eta}_i = \left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)', \mathbf{b}_i', \boldsymbol{\varepsilon}_i' \right]' \quad \text{y} \quad \mathbf{K}_i = \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i' \mathbf{V}_i^{-1}$$

$$\mathbf{e}_{i.0} = \mathbf{K}_i \left[-\mathbf{X}_i \ \mathbf{Z}_i \ \mathbf{I}_i \right] \boldsymbol{\eta}_i \quad \text{y} \quad \mathbf{e}_{i.1} = \left(\mathbf{I}_i - \mathbf{K}_i \right) \left[-\mathbf{X}_i \ \mathbf{Z}_i \ \mathbf{I}_i \right] \boldsymbol{\eta}_i .$$

Los vectores $\boldsymbol{\eta}_i$ están normalmente distribuidos con media cero y matriz de covariancias $\mathbf{T_i}$,

$$\text{Cov}(\boldsymbol{\eta}_i) = \begin{pmatrix} \text{Cov}(\hat{\boldsymbol{\beta}}) & \mathbf{H}_i \mathbf{Z}_i \mathbf{D} & \sigma^2 \mathbf{H}_i \\ \left(\mathbf{H}_i \mathbf{Z}_i \mathbf{D} \right)' & \mathbf{D} & \mathbf{0} \\ \sigma^2 \mathbf{H}_i' & \mathbf{0} & \sigma^2 \mathbf{I}_i \end{pmatrix} = \mathbf{T}_i, \quad \mathbf{H}_i = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}_i' \mathbf{V}_i^{-1}.$$

Estos residuos se usan para calcular las sumas de cuadrados de los residuos, denominadas $\text{RSS}_{i.0} = \mathbf{e}_{i.0}' \mathbf{e}_{i.0}$ y $\text{RSS}_{i.1} = \mathbf{e}_{i.1}' \mathbf{e}_{i.1}$, que contienen información sobre desviaciones tipo L y S, respectivamente.

Las sumas de cuadrados dependen de las unidades de medida y el número de mediciones por unidad por lo que resulta conveniente estandarizarlas,

$$\text{TRSS}_{i.0}^* = \frac{\text{RSS}_{i.0} - \text{E}(\text{RSS}_{i.0})}{\sqrt{\text{Var}(\text{RSS}_{i.0})}} \quad \text{y} \quad \text{TRSS}_{i.1}^* = \frac{\text{RSS}_{i.1} - \text{E}(\text{RSS}_{i.1})}{\sqrt{\text{Var}(\text{RSS}_{i.1})}} .$$

Las desviaciones positivas son más preocupantes que las negativas. Las sumas de cuadrados más chicas que su esperanza indican simplemente que el modelo ajusta mejor para esa unidad que para otras. Por lo cual sólo grandes valores positivos son de interés, definiendo $\text{TRSS}_{i.o} = \text{máx}\left\{0, \text{TRSS}_{i.o}^*\right\}$ y $\text{TRSS}_{i.1} = \text{máx}\left\{0, \text{TRSS}_{i.1}^*\right\}$.

El gráfico TRSS es un diagrama de dispersión entre $\text{TRSS}_{i.o}$ y $\text{TRSS}_{i.1}$ que muestra unidades discordantes y sus tipos de desviación:

- Cuando los valores de $TRSS_{i.o}$ son grandes y los de $TRSS_{i.1}$ chicos indica que el sujeto i está lejos de la media marginal (desviación tipo L)
- Cuando los valores de $TRSS_{i.1}$ son grandes y los de $TRSS_{i.o}$ chicos sugieren que la unidad puede tener diferentes trayectorias que las otras (desviación tipo S) o diferente estructura de correlación que las otras.

El gráfico TRSS muestra ambos tipos de desviaciones simultáneamente y permiten investigar una unidad sin eliminarla, pues la medida TRSS es una medida de no omisión de casos. Si existen varias unidades discordantes también permite detectarlas visualmente.

Se debe poner atención a las unidades que se presentan aisladas en los gráficos TRSS y unidades con valores de $TRSS_{i.o}$ y/o $TRSS_{i.1}$ más grandes que 2 ó 3.

Las unidades discordantes y sus tipos de desviación se determinan por la distancia al origen y la dirección en el gráfico TRSS. Es útil tener líneas de referencia para evaluar si una unidad se puede considerar o no atípica. Se establecieron tres posibles líneas a partir de una normal bivariada truncada, con dos niveles de cobertura (95 y 99%), usando

- la densidad de probabilidad más alta (HPD) que encuentra un valor de corte y un elipsoide en el primer cuadrante para encontrar la probabilidad de cobertura nominal,
- el cuantil local (LQ) establece un ángulo en el origen y encuentra los cuantiles de los datos al nivel nominal en término de la distancia al origen y
- regresión por cuantiles rotado (RQR) rota puntos de una bivariada truncada por 45º y obtiene (conseguir, lograr, alcanzar) una línea de regresión por cuantiles no paramétrica a un nivel nominal dado. Esta línea de cuantiles se gira de nuevo y forma la línea de referencia RQR.

4. Resultados

La metodología descrita se aplica a un conjunto de datos obtenidos en un estudio clínico desarrollado para evaluar la seguridad cardiológica de una droga. En el estudio participaron 48 pacientes los cuales fueron asignados a cinco tratamientos, cuatro de ellos consistían en tomar diariamente diferentes dosis de la droga (Grupos 1 a 4) y al otro se le suministró placebo (grupo 0). A cada paciente se realizó un electrocardiograma en 7 oportunidades: antes de recibir la primera dosis, dos horas después de haber recibido la primera dosis, luego uno diariamente durante 4 días y, por último, uno 2 días después de haber finalizado el tratamiento. Se registró una medida cardiológica de interés, la longitud del intervalo QT_c, con el fin de comprobar si la droga prolonga la longitud del intervalo.

Se propuso el siguiente modelo lineal mixto con un efecto aleatorio,

$$Y_{ij} = \beta_{00}G_0+\beta_{01}G_1+\beta_{02}G_2+\beta_{03}G_3+\beta_{04}G_4+b_{0i}+(\beta_{10}G_0+\beta_{11}G_1+\beta_{12}G_2+\beta_{13}G_3+\beta_{14}G_4)t_{ij}+e_{ij}$$

$$\text{Var}(\mathbf{b}) = \mathbf{D} = \text{Var}(b_{0i}) = \sigma_1^2$$

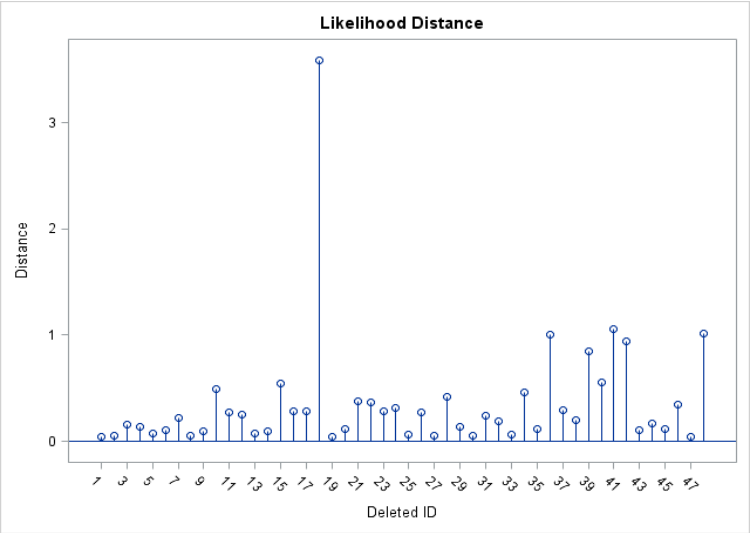
$$\text{Var}(\mathbf{e}) = \mathbf{R} = \sigma^2\,\mathbf{I}$$

El cálculo de las medidas de influencia y sus componentes interpretables se realiza utilizando una macro de SAS, el procedimiento “mixed” del software estadístico SAS y el paquete TRSS de R para los gráficos TRSS.

Es útil comenzar detectando los posibles casos atípicos utilizando una medida resumen. De esta forma se conocen los casos que podrían ser particularmente influyentes sobre algunos aspectos del análisis. Si no se identifica ninguno el procedimiento termina.

El enfoque de la influencia global, que se basa en la eliminación de una unidad, comienza detectando los casos atípicos mediante la distancia de verosimilitud (LD) (gráfico 1).

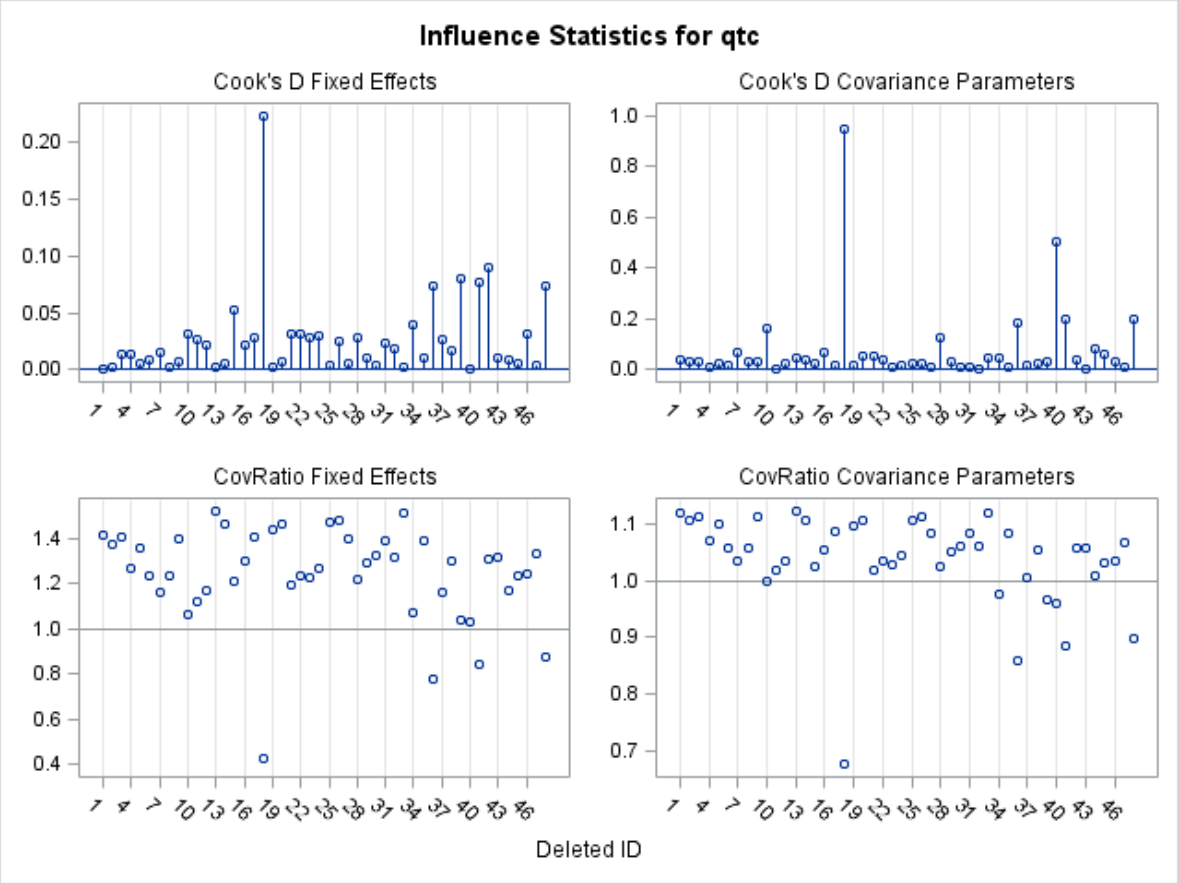
Gráfico 1 Diagnóstico de influencia general



Esa distancia para el paciente 18 tiene una gran magnitud y lo muestra como potencialmente influyente, seguido en menor medida, y en orden decreciente, por los individuos 41, 36, 48, 42, 39 y 40. Para los casos identificados se cuantifica el impacto que tienen sobre alguna componente del modelo.

El gráfico siguiente presenta los diagnósticos para medir la influencia, los gráficos de la izquierda suministran información sobre los efectos fijos y los de la derecha sobre los estimadores de los parámetros de covariancia.

Gráfico 2 Diagnósticos de influencia para los efectos fijos y de covariancia



El gráfico de la distancia de Cook muestra que el paciente con mayor efecto sobre los efectos fijos es el 18. Los individuos 39, 41, 42, 36 y 48 tienen una influencia fuerte sobre los efectos fijos.

Los pacientes 18, 36, 41 y 48 tienen valores de COVRATIO menores que 1 indicando que su eliminación del conjunto de datos produciría un aumento en la precisión estimada de los estimadores de los efectos fijos.

El valor de D de Cook para los parámetros de covariancia del paciente 18 es demasiado grande comparado con el de los otros pacientes. El caso 40 parece algo más influyente sobre los parámetros de covariancia que las unidades 36 y 48.

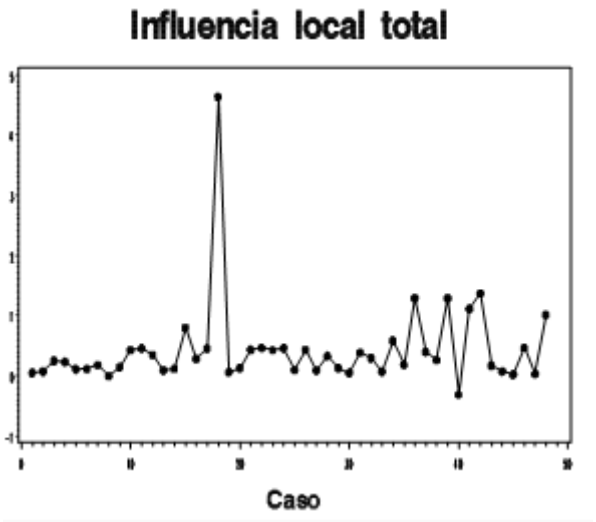
Los valores de COVRATIO muestran que en ausencia de las observaciones de los individuos 18, 36, 39, 40, 41 y 48 los parámetros de covariancia se podrían estimar con mayor precisión.

El procedimiento para realizar un análisis de influencia local consiste, primero, en detectar aquellos individuos que tienen un gran impacto sobre los parámetros estimados, a través de

C_i, y luego determinar las componentes del modelo que están más afectadas por los casos influyentes (la estructura media, la estructura de covariancias o ambas). Por último establecer las causas de la influencia para obtener una idea de las razones por las cuales ese caso es atípico.

Los casos con un valor grande de C_i se consideran que influencian la estimación del vector completo de parámetros.

Gráfico 3 Medidas de la influencia local total correspondientes a los individuos del estudio



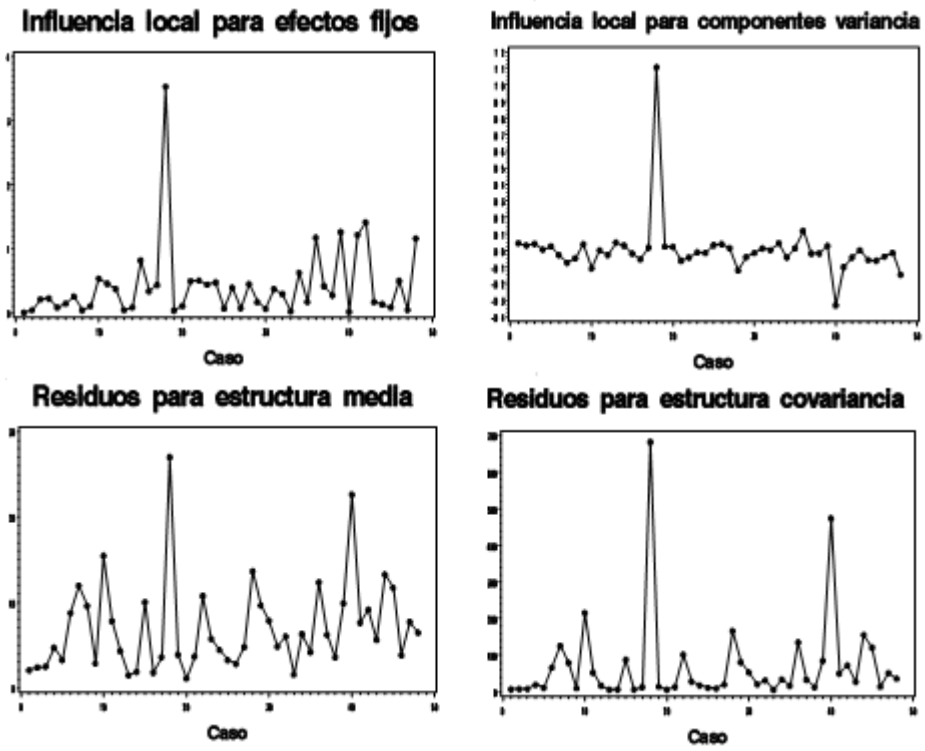
Los valores de C_i (gráfico 3) muestran que los casos 18, 36, 39, 40, 41, 42 y 48 poseen valores de la influencia local diferente al resto. Esto implica que son posiblemente influyentes sobre la estimación del vector de parámetros completo (ψ).

Los siguientes gráficos muestran en forma separada las medidas representadas en el gráfico 3.

Los sujetos 18, 36 y 40 son altamente influyentes tanto para los efectos fijos (gráfico 4 a) como para las componentes de variancia (gráfico 4 b). Los pacientes 39, 41, 42 y 48 son influyentes sólo para la estimación de los efectos fijos, ya que el gráfico 4b no los muestra con grandes valores de $C_i(\mathbf{D}, \boldsymbol{\sigma}^2)$. Las componentes residuales para la estructura media más altas corresponden a los pacientes 18 y 40, sugiriendo que sus perfiles medios no están bien predichos o representados por la estructura media del modelo utilizado. De la misma manera, los residuos más grandes para la estructura de covariancia corresponden a esos mismos sujetos. La matriz de covariancia de los mismos no está bien descrita por la covariancia del modelo.

Gráfico 4 Influencia local para los efectos fijos y componente de variancia

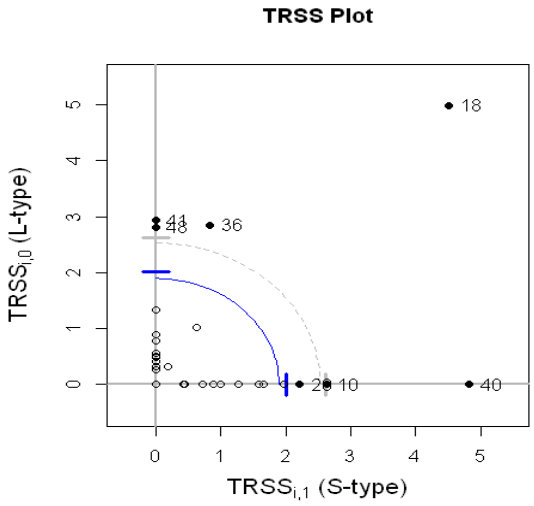
(a) (b)



(c) (d)

El gráfico TRSS siguiente muestra las unidades discordantes y las líneas de referencia

Gráfico 5 Gráfico de las sumas de cuadrados de los residuos estudentizados



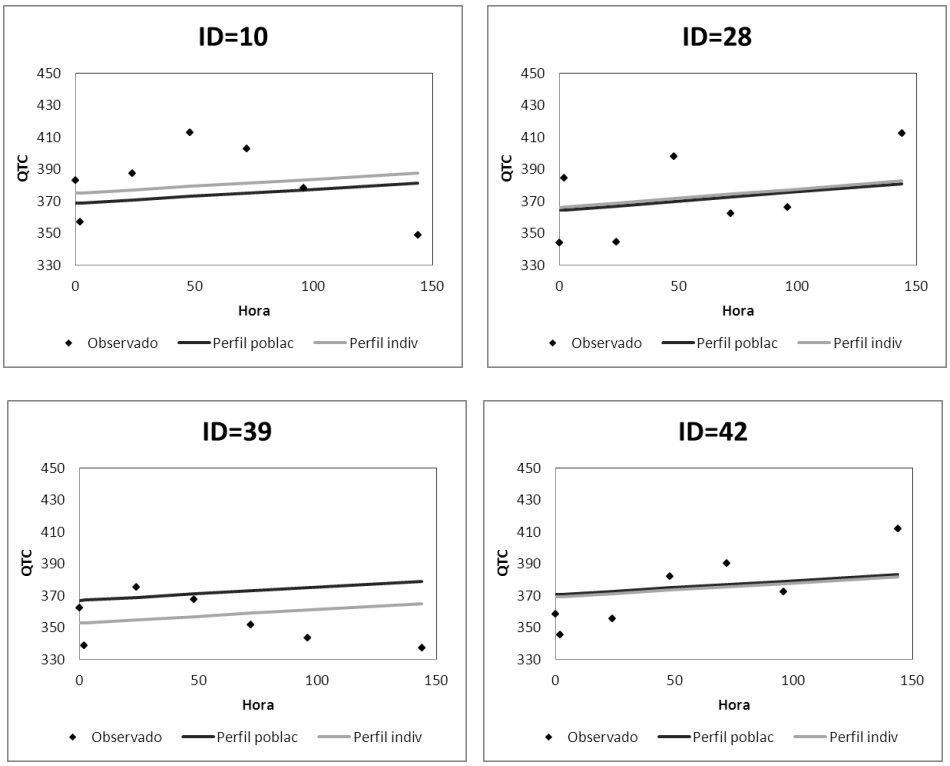
Se observa que las unidades 18, 36, 40, 41 y 48 superan las líneas de referencia del 99%

(línea punteada) y 95% (línea sólida), mientras que las unidades 10 y 28 son detectadas como influyentes mediante la línea del 95%.

Las unidades 10 y 28 son sólo identificadas por el gráfico TRSS, 39 y 42 por los enfoques global y local, mientras que las restantes por los tres métodos.

Para intentar explicar los motivos de las discrepancias entre los métodos **s**e presentan los gráficos de los perfiles individual y promedio por grupo y los valores observados para las unidades 10, 28, 39 y 42.

Gráfico 7 Perfiles individuales y promedio por grupo para cuatro pacientes influyentes



Las unidades 10 y 28 tienen $TRSS_{i1}$ grande y $TRSS_{i0}$ chica (Gráfico 6) sugiriendo que la unidad tiene trayectoria o estructura de correlación diferente que las otras.

La unidad 10 tiene trayectoria diferente y posiblemente no captada por el modelo. Tiene influencia sobre las componentes de variancia pero no sobre los efectos fijos.

El perfil individual de la unidad 28 es similar al perfil promedio del grupo. No tiene efectos sobre la estimación de los efectos fijos, por lo cual no la detectan los métodos global y local.

Los perfiles observados de las unidades 39 y 42 se desvían de la trayectoria lineal, pero como las desviaciones son más pequeñas que las observadas en la unidad 10 no son

captadas por los gráficos TRSS. La influencia de estas unidades es sólo sobre los efectos fijos.

5. Consideraciones finales

En este trabajo se presentan varios enfoques para detectar unidades que tienen una magnitud distinta al resto y el efecto que producen sobre los estimadores de los parámetros del modelo.

La idea general de los métodos de influencia global y local es introducir cambios en las componentes del modelo y evaluar si se producen cambios importantes en los resultados. El procedimiento comienza detectando los casos atípicos mediante la distancia de verosimilitud. Posteriormente, se descomponen los hallazgos iniciales para determinar si realmente esos casos afectan el proceso de estimación. Si esta medida general sugiere que existen unidades influyentes se tienen que determinar, a posteriori, los elementos del modelo que son influenciados.

Los gráficos TRSS, que fueron propuestos recientemente, no eliminan las unidades ni alteran el modelo para identificar las unidades discordantes. El método proporciona una mayor información sobre las mediciones repetidas mediante la utilización de residuos modificados y evalúa eficazmente el efecto de unidades y observaciones discordantes en la estimación de parámetros que incluyen componentes de la variancia.

Considerar unidades como influyentes no implica eliminarlas del conjunto o cambiar el modelo, pues, si los puntos afectan los efectos fijos sin ejercer demasiada influencia sobre la precisión de los parámetros de covariancia, su presencia no alterará ni las pruebas de hipótesis ni los intervalos de confianza para los parámetros de efectos fijos.

Los diagnósticos de los efectos fijos utilizan una matriz de covariancias especificada para los datos, así su influencia sobre las componentes de variancia se deberían examinar primero.

En la aplicación se muestra que:

- Influencia global y local: los diagnósticos ayudan a detectar pacientes atípicos mediante la inspección de la distancia de verosimilitud. Sin embargo, algunas unidades que se presentaron con valores altos de la distancia de verosimilitud restringida resultan tener mayor efecto sobre los efectos fijos y otras casi sin impacto sobre los efectos fijos se muestran principalmente influyentes sobre los estimadores puntuales de covariancia.

- Gráficos TRSS: detectan en general las mismas unidades que los métodos anteriores. Sin embargo, ayudan a identificar unidades con trayectorias o posiblemente con estructuras de correlación diferentes al resto.

Referencias Bibliográficas

Banerjee M, Frees EW. (1997) Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*; 92:999–1005.

Banerjee M. (1998) Cook's distance in linear longitudinal models. *Communications in Statistics: Theory and Methods*; 27:2973–2983.

Beckman, R. J., Nachtsheim, C.J. and Cook, R. D. (1987) "Diagnostics for mixed-model analysis of variance*". Technometrics* 29, 413-426.

Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons: New York, NY, 1980.

Christensen, R., Pearson, L.M. and Johnson, W. (1992) Case-deletion diagnostics for mixed models. *Technometrics* 34, 38-45.

Cook RD. (1977) Detection of influential observation in linear regression. *Technometrics*; 19,15–18.

Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall.

Cook, R.D. (1986) Assessment of local influence *Journal of the Royal Statistical Society*, Series B 48,133-169.

De Gruttola, V., Ware, J.H., and Louise, T.A. (1987). Influence analysis of generalized least squares estimators. *Journal of the American Statistical Associations* 82,911-917.

Garcia, M. del C., Koegel, L., Rapelli, C. (2008) Diagnósticos para los modelos lineales mixtos. Un análisis comparativo de dos enfoques para evaluar la influencia. Libro "II Jornada de Ciencia y Tecnología. Divulgación de la Producción Científica y Tecnológica de la UNR". 169-173.

Garcia, M. del C., Méndez, F. (2007) Métodos diagnósticos para evaluar la influencia en el contexto de los modelos lineales mixtos. XXXV Coloquio Argentino de Estadística. Pág web http://www.s-a-e.org.ar/ultimos13coloquios.htm

Kim C, and Storer BE. (1996) Reference values for cook's distance. *Communications in Statistics: Simulation and Computation*, 25:691–708.

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models *Biometrics* 54, 570-582.

Littell, R.C.. Milliken, G.A., Stroup, W.W.; Wolfinger, R.D. (1996) *SAS® System for Mixed Models.* Cary, NC: SAS Institute Inc.

Mun, J. and Lindstrom, M. (2013) Diagnostics for repeated measurements in linear mixed effects models. *Statistics in Medicine*, 32 1361–1375

Pan J, Fang K. (1996) Influential observation in the growth curve model with unstructured covariance matrix. *Computational Statistics & Data Analysis*; 22:71–87.

Pinheiro JC, Bates DM. (2000) *Mixed-effects Models in S and S-Plus*. Springer-Verlag Inc: New York, NY.

Tan FES, Ouwens MJNM, Berger MPF. (2001) Detection of influential observations in longitudinal mixed effects regression models. *The Statistician*, 50:271–284.