Trabajo completo

Estudio de la validez de los supuestos en los modelos lineales mixtos mediante un análisis de residuos

RECIBIDO: 04/10/2011
ACEPTADO: 26/07/2012

Garcia, M. del C. • Rapelli, C.

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvard Oroño 1261. (2000) Rosario, Santa Fe, Argentina.
Teléfono: 0341-4802794-int.152. E-mail: mgarcia@fcecon.unr.edu.ar

RESUMEN: Los modelos mixtos se utilizan para analizar datos longitudinales. El uso apropiado de estos modelos requiere que los supuestos subyacentes en el análisis sean validados tan cuidadosamente como sea posible, incluyendo los supuestos distribucionales. Los residuos son de gran importancia para detectar violaciones de los supuestos involucrados en el modelo. En los modelos mixtos los residuos marginales, condicionales y los BLUP, que predicen los correspondientes errores y los efectos aleatorios respectivamente, se utilizan para verificar homocedasticidad de varianzas, presencia de outliers y normalidad de los errores. Los residuos con confusión mínima fueron propuestos recientemente para comprobar la normalidad del error condicional y su uso no está muy difundido entre los investigadores. En este trabajo se aborda el estudio de los supuestos de un modelo utilizando datos obtenidos de una investigación para determinar la seguridad cardiológica de una nueva droga. Se realizan gráficos probabilísticos normales con bandas de confianza simuladas ("envelopes") para comparar el comportamiento de los residuos con confusión mínima y los residuos condicionales clásicos en la evaluación del supuesto de normalidad del error condicional.

PALABRAS CLAVE: Estudios longitudinales, Modelos lineales mixtos, Diagnósticos, Residuos.

SUMMARY: *Study of the validity of linear mixed model assumptions through the analysis of residuals*
Mixed models are frequently used in the analysis of longitudinal data. The appropriate

use of these models requires checking the vali-dity of the underlying assumptions, including the distributional assumptions. Residuals are used to examine the validity of model assumptions. In mixed models, the marginal and conditional residuals and BLUP are used to verify homoscedasticity, the presence of outliers and the normality of errors. The least confounded residuals are useful to evaluate the normality of the conditional error. In this paper, we examine the assumptions of a model of evaluating the cardiology security of a drug.

KEYWORDS: longitudinal data, linear mixed models, diagnostics, residuals.

---

1. Introducción

Los modelos mixtos constituyen una valiosa herramienta para analizar datos compuestos por mediciones repetidas y en particular datos longitudinales. Para evaluar la validez de estos modelos se utilizan residuos, cuyo análisis se usa frecuentemente para comprobar el cumplimiento de los supuestos de los modelos estadísticos. En los modelos mixtos se distinguen los residuos marginales, condicionales y los BLUP, que predicen los correspondientes errores y los efectos aleatorios respectivamente. Cada residuo, estandarizado o no, se usa para comprobar algún supuesto. Los dos últimos residuos mencionados presentan confusión en el sentido que no sólo son función del error que predicen y pueden ser inadecuados para evaluar algún supuesto sobre los mismos (1). En los últimos años se propuso, para minimizar el efecto de la confusión, una transformación lineal de los residuos condicionados, surgiendo los residuos con confusión mínima que son residuos condicionales que contienen una mínima fracción de confusión (1).

En este trabajo se presentan técnicas de diagnóstico para realizar el chequeo de los supuestos del modelo. Para la evaluación del supuesto de normalidad del error condicional se realiza un gráfico probabilístico normal con bandas de confianza simuladas (“envelopes”) de los usuales residuos condicionales y de los residuos con confusión mínima, sugeridos recientemente y cuyo uso no está muy difundido entre los investigadores. Se aplican en un estudio para comprobar la seguridad cardiológica de una nueva droga.

2. Métodos estadísticos

En los estudios longitudinales las unidades experimentales se observan repetidamente en varias ocasiones. El modelo lineal mixto, adecuado para analizar ese tipo de datos, se puede expresar como,

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon} \qquad (2.1)$$

siendo,

$$\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_1', \ldots, \mathbf{Y}_N')$$

$$\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_N')$$

$$\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_N')$$

$$\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_N')$$

$$\mathbf{Z} = \oplus_{i=1}^{N} \mathbf{Z}$$

donde $\oplus$ representa la suma directa, $\mathbf{Y}_i$, i=1,... N, el vector (n_i x 1) de medidas repetidas del i-ésimo individuo, x_i y z_i son matrices conocidas de dimensión (n_i x p) y (n_i x k) respectivamente, $\mathbf{b}_i$ es un vector aleatorio de dimensión (k x 1), cuyas componentes se denominan efectos aleatorios, ε_i es un vector aleatorio (n_i x 1) de mediciones de error (intra-sujeto), β un vector (p x 1) de parámetros.

Los supuestos del modelo son, â

$$\begin{pmatrix}\mathbf{b}\\ \boldsymbol{\varepsilon}\end{pmatrix} \sim N_{kN+n}\left(\begin{pmatrix}\mathbf{0}_{kN}\\ \mathbf{0}_n\end{pmatrix},\begin{pmatrix}\sigma^2\mathbf{D} & \mathbf{0}_{kN\times n}\\ \mathbf{0}_{n\times kN} & \sigma^2\boldsymbol{\Sigma}\end{pmatrix}\right),$$

con, $n=\sum_{i=1}^{N} n_i$, $\mathbf{D}=\mathbf{I}_N\otimes\mathbf{G}$ y , $\boldsymbol{\Sigma}=\oplus_{i=1}^{N}\mathbf{R}_i$

$\otimes$ denota el producto Kronecker e I_N la matriz identidad de orden N y además,

$$\mathbf{b}_i \overset{iid}{\sim} N_k(\mathbf{0},\sigma^2\mathbf{G}) \quad y \quad \boldsymbol{\varepsilon}_i \overset{ind}{\sim} N_{n_i}(\mathbf{0},\sigma^2\mathbf{R}_i).$$

Bajo el modelo (2.1) la matriz de variancias y covariancias de Y, cuyos elementos se sintetizan en un vector θ de parámetros de covariancia, es $\mathrm{Var}(\mathbf{Y})=\mathbf{V}=\sigma^2(\mathbf{ZDZ}'+\boldsymbol{\Sigma})$. Cada elemento de la matriz V se simboliza V_i.

La estimación de los parámetros β y θ se realiza minimizando la función objetivo, menos dos veces el logaritmo de la función de verosimilitud (-2l), mediante el algoritmo de Newton-Raphson (2). Condicional a D y Σ el estimador máximo verosímil de β es,

$$\hat{\boldsymbol{\beta}}=\left(\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{V}}_i^{-1}(\boldsymbol{\theta})\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{V}}_i^{-1}(\boldsymbol{\theta})\mathbf{Y}_i\,,$$

y la predicción del vector de efectos aleatorios, denominado BLUP empírico (EBLUP),

.

$$\hat{\mathbf{b}}=\hat{\mathbf{D}}\mathbf{Z}_i'\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

Residuos para los modelos mixtos

El análisis de residuos se utiliza para chequear la validez de los supuestos subyacentes (3, 4). Puesto que los modelos mixtos tienen dos fuentes de variabilidad (intra y entre unidad) se definen tres tipos de residuos (1,5):

I. Residuos marginales,

$$\hat{\boldsymbol{\xi}}=\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{M}^{-1}\mathbf{QY}=[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{M}]\boldsymbol{\xi}\,,$$

que predicen los errores marginales,

$$\boldsymbol{\xi}=\mathbf{Y}-E(\mathbf{Y})=\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}=\mathbf{Zb}+\boldsymbol{\varepsilon}\,. \qquad (2.2)$$

II. Residuos condicionales

$$\hat{\boldsymbol{\varepsilon}}=\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{Z}\hat{\mathbf{b}}=\boldsymbol{\Sigma}\mathbf{QY}=\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\varepsilon}+\boldsymbol{\Sigma}\mathbf{QZb}\,,$$

que predicen los errores condicionales,

$$\boldsymbol{\varepsilon}=\mathbf{Y}-E(\mathbf{Y}/\mathbf{b})=\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zb}\,. \qquad (2.3)$$

III. BLUP, $\mathbf{Z}\hat{\mathbf{b}}=\mathbf{ZDZ}'\mathbf{QZb}+\mathbf{ZDZ}'\mathbf{Q}\boldsymbol{\varepsilon}$,

que predice los efectos aleatorios,

$$\mathbf{Zb}=E(\mathbf{Y}/\mathbf{b})-E(\mathbf{Y})\,, \qquad (2.4)$$

donde,

$$\mathbf{Q}=\mathbf{M}(\mathbf{I}-\mathbf{T}),\quad \mathbf{M}=\sigma^2\mathbf{V}^{-1}=(\mathbf{ZDZ}'+\boldsymbol{\Sigma})^{-1},$$

$$\mathbf{T}=(\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{M}.$$

De acuerdo con Hilden-Minton (1), un residuo se considera puro para un tipo de error específico si depende sólo de las componentes fijas y del error que se supone que predice. Los residuos que dependen de otros tipos de error se denominan "residuos confundidos". Los residuos $\hat{\varepsilon}$ y $\widehat{Zb}$ están confundidos con b y ε, respectivamente. Esto implica, por ejemplo, que $\hat{\varepsilon}$

puede no ser adecuado para chequear la normalidad de ε debido a que cuando b no se distribuye normalmente, $\hat{\varepsilon}$ puede no presentar un comportamiento normal aún cuando ε sea normal.

Cada tipo de residuo es útil para diagnosticar algún incumplimiento de los supuestos del modelo (2.1).

Los *residuos marginales* se usan para estudiar la linealidad de la respuesta con respecto a las variables explicativas. Para evaluar la validez de la estructura de covariancia intra individuo, $\mathbf{V}_i=\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'+\mathbf{R}_i$, Lesafre y Verbeke (6) consideran los residuos para la matriz de covariancia intra individuo (denominados residuos de covariancia) $\left\|\mathbf{I}_{ni}-\mathbf{R}_i\mathbf{R}_i'\right\|$, siendo $\mathbf{R}_i=\hat{\mathbf{V}}_i^{-1/2}\boldsymbol{\xi}_i$ y $\|\mathbf{A}\|$ la norma de Frobenius de la matriz $\mathbf{A}$.

Los *residuos condicionales* se utilizan para chequear la homocedasticidad y normalidad del error condicional e identificar posibles outliers. Pinhero y Bates (7) utilizan $\hat{\varepsilon}/\hat{\sigma}$, siendo $\hat{\sigma}$ una estimación de σ, en un gráfico probabilístico normal (Q-Q plot) para evaluar la homocedasticidad y normalidad del error condicional. Nobre y Singer (5) sugieren estandarizar los resduos condicionales,

$$\hat{\varepsilon}_k^* = \frac{\hat{\varepsilon}_k}{\hat{\sigma}/\sqrt{\hat{p}_{kk}}}$$

siendo $\hat{p}_{kk}$ un estimador de p_{kk} el k-ésimo elemento diagonal de la matriz $\boldsymbol{\Sigma Q \Sigma}$ componente de la Var $(\hat{\boldsymbol{\varepsilon}})=\sigma^2\ \boldsymbol{\Sigma Q \Sigma}$. Estos residuos se pueden utilizar para la identificación de observaciones atípicas.

Los *EBLUP* se utilizan para comprobar la normalidad de los efectos aleatorios y para identificar individuos atípicos. Pinhero y Bates (7) sugieren graficar los elementos de $\hat{\mathbf{b}}_i$ versus el número de sujeto para identificar sujetos "outliers" con respecto al j-ésimo elemento del BLUP respectivo. En cambio Waternaux et al. (9) utilizan con el mismo fin el gráfico de los $\mathbf{Z}_i\hat{\mathbf{b}}_i$ o la distancia $\hat{\boldsymbol{\xi}}=\hat{\mathbf{b}}_i'\mathrm{Var}(\hat{\mathbf{b}}_i-\mathbf{b}_i)\hat{\mathbf{b}}_i$. .

Residuos con confusión mínima

Una situación a la que permanentemente se enfrentan los investigadores es la evaluación del supuesto de normalidad. En este contexto los residuos con confusión mínima resultan de mayor utilidad que los residuos condicionales estandarizados.

Hilden-Minton (1) comentaron que la habilidad para chequear la normalidad de $\boldsymbol{\varepsilon}$, usando $\hat{\boldsymbol{\varepsilon}}$, decrece a medida que la $\mathrm{Var}[\boldsymbol{\Sigma Q Z' b}]=\sigma^2\boldsymbol{\Sigma Q Z D Z' Q \Sigma}$, variancia una componente de (2.2), aumenta en relación a la $\mathrm{Var}[\boldsymbol{\Sigma Q \varepsilon}]=\sigma^2\boldsymbol{\Sigma Q \Sigma Q \Sigma}$, variancia de una componente de (2.3); esto motiva la definición de la "fracción de confusión" para el k-ésimo residuo condicional como,

$$0 \le F_k = \frac{\mathbf{U}_k'\boldsymbol{\Sigma Q Z D Z' Q \Sigma}\mathbf{U}_k}{\mathbf{U}_k'\boldsymbol{\Sigma Q \Sigma}\mathbf{U}_k} = 1-\frac{\mathbf{U}_k'\boldsymbol{\Sigma Q \Sigma Q \Sigma}\mathbf{U}_k}{\mathbf{U}_k'\boldsymbol{\Sigma Q \Sigma}\mathbf{U}_k} \le 1, \quad (2.5)$$

donde $\mathbf{U}_K=(u_{ij})_{n\times v}=(\mathbf{U}_{K_1},\mathbf{U}_{K_2},\ldots,\mathbf{U}_{K_v})$
siendo $\mathbf{U}_\mathbf{k}$ la k-ésima columna de $\mathbf{I}_\mathbf{n}$.

Esta fracción (F_k) representa la proporción de la variabilidad de $\hat{\boldsymbol{\varepsilon}}_\mathbf{k}$ atribuida a la confusión con el BLUP. Hilden-Minton (1) proponen considerar una transformación lineal, $\mathbf{L}'\hat{\boldsymbol{\varepsilon}}$, tal que $\mathbf{L}'\hat{\boldsymbol{\varepsilon}}$ tenga confusión mínima. Esto se obtiene maximizando (2. 2) sujeto a la restricción que Var ($\mathbf{L}'\hat{\boldsymbol{\varepsilon}}$) es mayor que 0, (8). Una exposición completa se presenta en Hilden-Minton (1).

3. Aplicación

Para realizar el análisis de supuestos se utilizan datos obtenidos en un estudio farmacológico 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 suministrar diariamente diferentes dosis de la droga (Grupos 1 a 4) y al restante placebo (Grupo 0). A cada paciente se realiza un electrocardiograma en 16 oportunidades: antes de recibir la primera dosis (T=0), dos horas después de haber recibido la primera dosis y luego uno diariamente durante 14 días. Se registra una medida de interés, la longitud del intervalo QT_c (obtenido a partir del electrocardiograma), con el fin de comprobar si la droga prolonga la longitud del intervalo.

Se propuso un modelo lineal mixto con un efecto aleatorio y errores independientes con variancia homogénea.

$$Y_{ij} = \beta_{00}G_0 + \beta_{01}G_1 + \beta_{02}G_2 + \beta_{03}G_3 + \beta_{04}G_4 + b_{0i}$$

$$+ (\beta_{10}G_0 + \beta_{11}G_1 + \beta_{12}G_2 + \beta_{13}G_3 + \beta_{14}G_4)t_{ij} + \varepsilon_{ij}$$

$$\mathrm{Var}(\mathbf{b}) = \sigma^2 \mathbf{D} = \sigma_b^2$$

$$\mathrm{Var}(\mathbf{e}) = \sigma^2 \boldsymbol{\Sigma} = \sigma_\varepsilon^2 \mathbf{I}$$

siendo, las variables indicadoras de grupo, G_q = Grupo q, q=0,1,..,4.

Para la aplicación se utiliza el procedimiento “mixed” del software estadístico SAS (10, 11) y el software estadístico R (12). Los modelos estimados por grupo y las variancias de las componentes aleatorias resultan:

$$\text{Grupo 0} \quad Y_{ij} = 371.77 + 0.1729\, t_i$$

$$\text{Grupo 2} \quad Y_{ij} = 366.99 + 0.5804\, t_i$$

$$\text{Grupo 4} \quad Y_{ij} = 378.84 + 0.2411\, t_i$$

$$\text{Grupo 1} \quad Y_{ij} = 372.60 + 0.2199\, t_i$$

$$\text{Grupo 3} \quad Y_{ij} = 371.97 + 0.2002\, t_i$$

$$\hat{\sigma}_b^2 = 266.45$$

$$\hat{\sigma}_\varepsilon^2 = 210.39$$

Se realiza un análisis gráfico de residuos con el fin de visualizar la presencia de alejamientos sistemáticos del modelo.

El gráfico 3.1, de los residuos marginales versus la covariable, muestra que el modelo lineal resulta apropiado.

Gráfico 3.1. Residuos marginales versus el tiempo

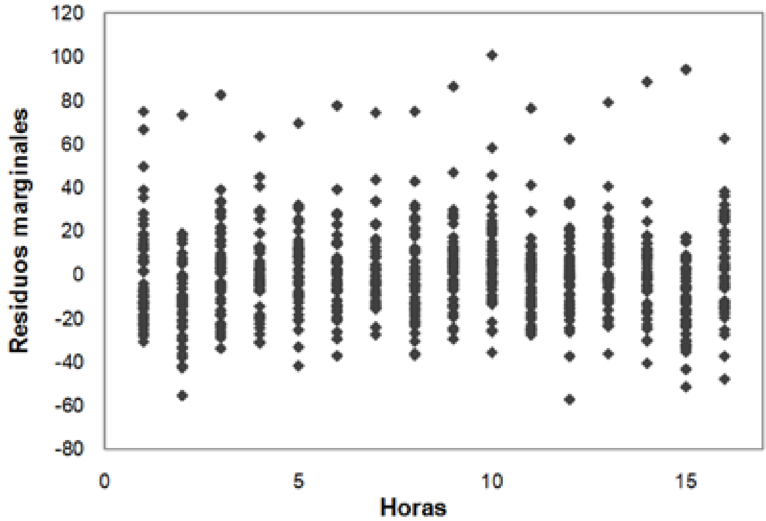
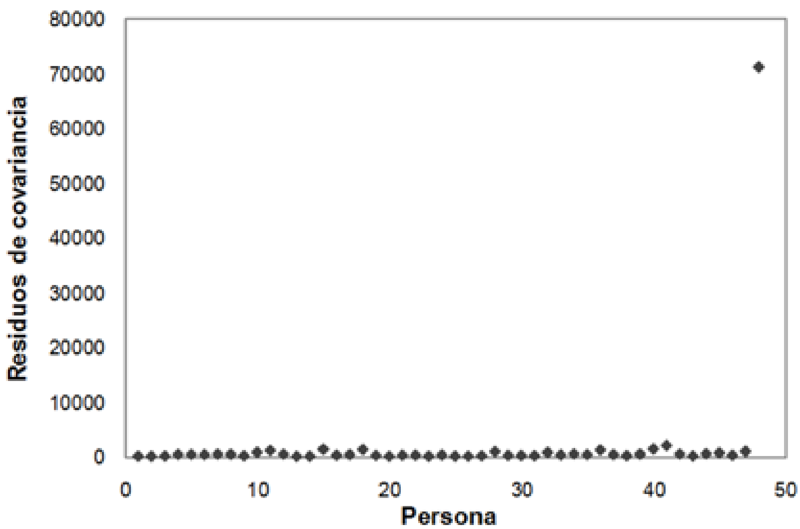


Gráfico 3.2. Residuos de covariancia intra individuo versus sujeto



El gráfico 3.2. permite comprobar si la matriz de covariancias intra sujeto adoptada en el modelo es adecuada para todos los sujetos. En general se considera que si los valores se alejan demasiado de cero la matriz no ajusta bien. Para el individuo 49 se visualiza un residuo de gran magnitud indicando que la matriz elegida no representa bien la covariancia de ese sujeto.

Los residuos condicionales estandarizados se utilizan en los siguientes gráficos para evaluar la presencia de individuos atípicos, gráfico 3.3, y comprobar la homocedasticidad del error condicional, gráfico 3.4.

No se observa la presencia de observaciones atípicas, es decir, observaciones con comportamiento diferente al resto de las observaciones.

Gráfico 3.3. Residuos condicionales estandarizados versus sujetos

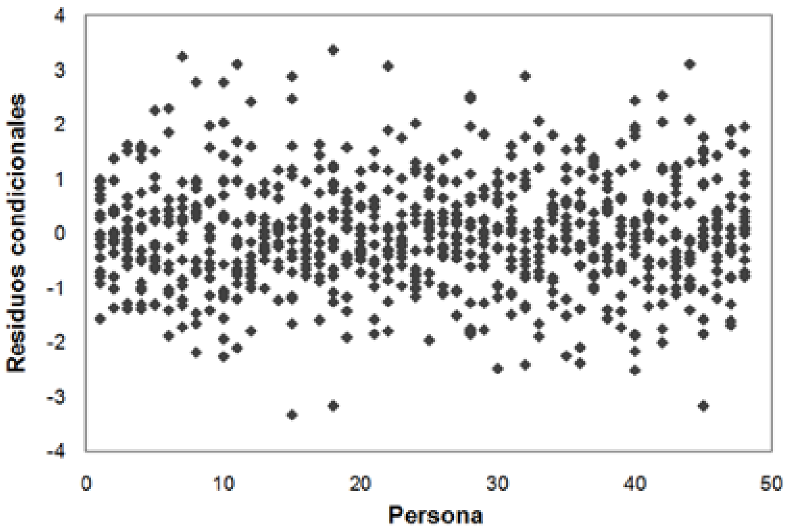
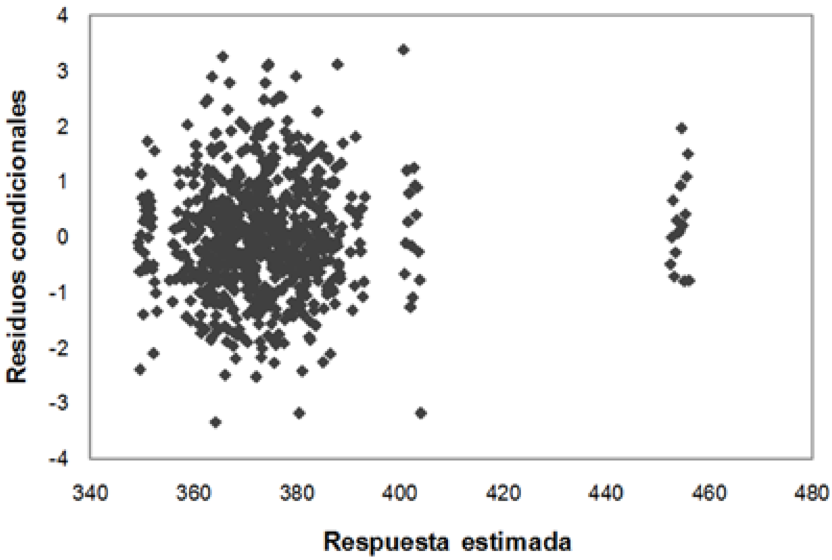


Gráfico 3.4. Residuos condicionales estandarizados versus la respuesta estimada

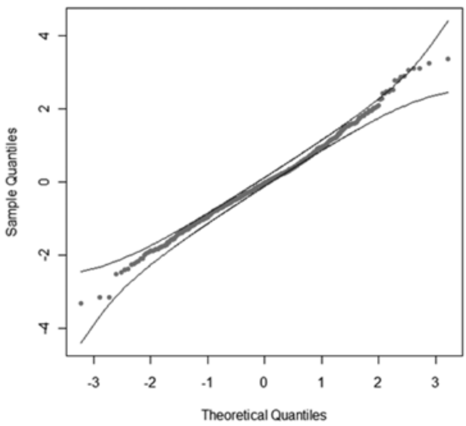


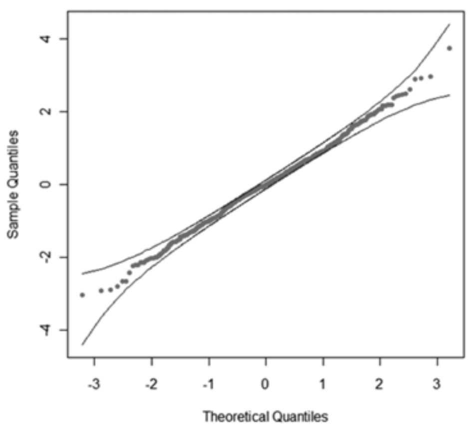
Si bien no se evidencia heterogeneidad de variancias aparece un conjunto de valores estimados diferentes al resto. Estos valores corresponden al último individuo.

Para evaluar la normalidad del error condicional se utiliza un gráfico probabilístico normal (QQ-plot), que se obtiene graficando los cuantiles de los residuos ordenados versus los cuantiles de una distribución normal estándar. Si los residuos realmente tienen una distribución normal se espera que el gráfico muestre una línea recta. Pero el problema que surge con esta línea es que pasa por los cuartiles inferior y superior y no dice nada acerca de las colas de la distribución. Entonces, se puede graficar una banda de confianza simulada, también llamada un sobre ("envelope") de confianza, del 95% para el QQ-plot, de manera que si los puntos caen fuera de la banda el supuesto de normalidad para los residuos no es válido. Para producir una banda de confianza por simulación se generan muestras aleatorias de una distribución normal de residuos estandarizados que tienen la estructura de covarianza idéntica a la de los datos. Este procedimiento se realiza usando tanto los residuos condicionales estandarizados como los residuos con mínima confusión.

Gráfico 3.5. Gráfico probabilístico normal con bandas de confianza simuladas de los residuos

(a) condicionales estandarizados **(b)** mínima confusión

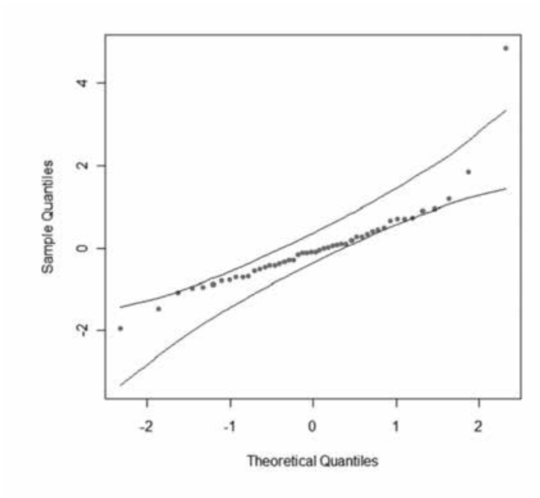




El grafico 3.1 (a) es el habitual gráfico probabilístico normal de los residuos estandarizados. Este gráfico sugiere no normalidad de los errores condicionales, mientras que en (b), donde se grafican los residuos condicionales con confusión mínima, no se identifican observaciones fuera del intervalo simulado, indicando que los residuos transformados muestran normalidad.

Para corroborar lo visualizado en el gráfico anterior se realiza un QQ-plot de la predicción de los efectos aleatorios.

Gráfico 3.6. Gráfico probabilístico normal con bandas de confianza simuladas de los efectos aleatorios



El punto que cae fuera de la banda pone en duda el cumplimiento del supuesto de normalidad de los efectos aleatorios.

Las conclusiones obtenidas de estos dos últimos gráficos se resumen como,

- Los residuos condicionales estandarizados no muestran normalidad.
- Los residuos con confusión mínima evidencian su utilidad para comprobar la normalidad del error condicional.
- Los efectos aleatorios no muestran normalidad.

Esto sugiere que los residuos estandarizados con mínima confusión se pueden emplear para evaluar el supuesto de normalidad del error condicional aun cuando los efectos aleatorios no sean normales.

4. Discusión

En este trabajo se considera un análisis de residuos para detectar alejamientos de los supuestos en un modelo lineal mixto.

Se pueden definir varios residuos, sin embargo, algunos de ellos están sujetos a confusión debido a la presencia de otras fuentes de variación. Específicamente, para evaluar la normalidad se utilizan los residuos condicionales estandarizados que presentan confusión en el sentido que no sólo son función del error que predicen y pueden no ser adecuados para evaluar ese supuesto. Para minimizar el efecto de la confusión se sugiere utilizar una transformación lineal de los residuos condicionados, los residuos con confusión mínima ("least confounded residuals").

En la aplicación el uso de residuos permitió comprobar el cumplimiento de los supuestos. Además, se mostró mediante un gráfico probabilístico normal con bandas de confianza ("envelopes") simuladas que el uso de los residuos condicionales puede no ser adecuado. En cambio, el uso de estos residuos transformados muestra que la confusion presente en el error condicional se debe tomar en cuenta.

Referencias bibliográficas

1. Hilden-Minton, J. A. (1995). Multilevel diagnostics for mixed and hierarchical linear models. PhD Thesis. University of California, Los Angeles.

2. Demidenko, E. (2004). "Mixed models: theory and applications". Wiley (New York.

3. Cook, R. D. and Weisberg, S. (1982). "Residuals and influence regression". Chapman & Hall (New York).

4. Residual plots for repeated measures. Statistics in Medicine 11, 115–124.

5. Nobre, J. S. and Singer, J. M. (2006) Residual Analysis for Linear Mixed Models. Biometrical Journal 49 (2007) 6, 863-875.

6. Lesaffre, E. and Verbeke, G. (1998). Local influen-ce in linear mixed models. Biometrics 54, 570–582.

7. Pinheiro, J. C. and Bates, D. M. (2000)."Mixed-effects in S and S-PLUS". Springer (New York).

8. Searle, S. R. (1982)."Matrix Algebra Useful for Statistics". Wiley (New York).

9. Waternaux, C., Laird, N. M., and Ware, J. H. (1989). Methods for analysis of longitudinal data: blood-lead concentrations and cognitive development. Journal of the American Statistical Association 84, 33–41.

10. SAS Institute Inc. 2004. SAS/STAT Software: Versión 9.1. Cary, NC: SAS Institute Inc.

11. Littell, R.C., Milliken, G.A., Stroup, W.W.y Wolfinger, R.D.(1996) "SAS System for Mixed Models". SAS Institute Inc. (Cary, North Carolina).

12. R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.