Psicothema 2010. Vol. 22, nº 2, pp. 323-333
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Selección de modelos anidados para datos longitudinales usando criterios de información y la estrategia de ajuste condicional

Guillermo Vallejo Seco, Jaime Arnau Gras*, Roser Bono Cabré*,
Paula Fernández García y Ellián Tuero Herrero
Universidad de Oviedo y ** Universidad de Barcelona

Un marco teórico potente resulta clave para especificar el modelo mixto que explica mejor la variabilidad de datos longitudinales. A falta de teoría, la mayoría de las investigaciones realizadas hasta la fecha, se ha centrado en ajustar la matriz de dispersión usando criterios de selección de modelos para elegir entre estructuras de covarianza no anidadas. En este trabajo, comparamos el desempeño del estadístico razón de verosimilitud (LRT) condicional y de varias versiones de los criterios de información para seleccionar estructuras de medias y/o de covarianzas anidadas, asumiendo conocido el verdadero proceso generador de datos. Los resultados numéricos indican que los criterios de información eficientes funcionaban mejor que sus homólogos consistentes cuando las matrices de dispersión usadas en la generación eran complejas y peor cuando eran simples. Globalmente, el desempeño del LRT condicional basado en el estimador de máxima verosimilitud completa (FML) era superior al resto de los criterios examinados. Sin embargo, el desempeño era inferior cuando se basaba en el estimador máxima verosimilitud restringida (REML). También encontramos que la estrategia sugerida en la literatura estadística de usar el estimador REML para seleccionar la estructura de covarianza y el estimador FML para seleccionar la estructura de medias debería ser evitada.

Nested model selection for longitudinal data using information criteria and the conditional adjustment strategy. Knowledge of the subject matter plays a vital role when attempting to choose the best possible linear mixed model to analyze longitudinal data. To date, in the absence of strong theory, much of the work has focused on modeling the covariance matrix by comparing non-nested models using selection criteria. In this paper, we compare the performance of conditional likelihood ratio test (LRT) and several versions of information criteria for selecting nested mean structures and/or nested covariance structures, assuming that the true data-generating processes are known. Simulation results indicate that the efficient criteria performed better than their consistent counterparts when covariance structures used in the data generation were complex, and worse when structures were simple. The conditional LRT under full maximum likelihood (FML) estimation was better overall than the other criteria in terms of selection performance. However, under restricted maximum likelihood (REML), estimation was inferior. We also find that the strategy suggested in the statistical literature of using REML for covariance structure selection, and FML for mean structure selection may be misleading.

Actualmente, cada vez son más las disciplinas que utilizan enfoques basados en la teoría del modelo lineal mixto para analizar datos que presentan una estructura jerarquizada (Vallejo, Arnau y Bono, 2008a). Además del gran desarrollo teórico que estos modelos han experimentado en las tres últimas décadas, el establecimiento definitivo de los mismos se ha visto favorecido por la incorporación de procedimientos analíticos específicos dentro de los principales paquetes estadísticos profesionales, incluyendo el módulo *Proc Mixed* en SAS, la función *lme* en S-PLUS/R o los comandos *Mixed y xtmixed* en SPSS y STATA, respectivamente. Los modelos lineales mixtos permiten analizar datos de corte longitudinal y transversal, aunque son especialmente útiles cuando se trabaja con datos temporales, ya que permiten ajustar y realizar inferencias acerca de la estructura de medias (como sucede con los clásicos enfoques univariante y multivariante de medidas repetidas) y modelar la estructura de covarianza (en términos de efectos aleatorios y error puro). Mediante este enfoque, más que asumir una matriz de dispersión demasiado parca (p. e., la matriz de simetría compuesta —CS— típica del enfoque univariante) o una complemente general (p. e., la matriz no estructurada —UN— típica del enfoque multivariante), se trata de buscar un equilibrio entre los criterios de flexibilidad y parsimonia o simplicidad científica (Ato y Vallejo, 2007). Al respecto, hay que advertir que si un investigador especifica un modelo excesivamente simple corre el riesgo de efectuar inferencias erróneas, debido a la subestimación de los errores estándar. Si, por el contrario, formula un modelo ex-

Fecha recepción: 12/7/2009 • Fecha aceptación: 26/10/2009
Correspondencia: Guillermo vallejo Seco
Facultad de Psicología
Universidad de Oviedo
33450 Oviedo (Spain)
e-mail: gvallejo@uniovi.es

cesivamente complejo corre el riesgo de efectuar inferencias ineficientes.

Con el fin de mejorar la calidad de las inferencias obtenidas con el enfoque del modelo mixto, resulta crucial modelar dos aspectos diferentes de los datos (Littell, Pendergast y Natarajan, 2000). Por un lado, los efectos fijos usados para describir el promedio de las respuestas en función del tiempo (en adelante, estructura de medias). Y, por otro lado, los efectos aleatorios usados para describir la variación entre las medidas repetidas dentro de los sujetos (en adelante, estructura de covarianza). Cuando se modela de forma efectiva la estructura de covarianza y también la de medias, dado que la forma de primera depende de la elección que se haga de la segunda (Fitzmaurice, Laird y Ware, 2004), se obtienen estimaciones más exactas (con menor sesgo) y precisas (con menor varianza) de los parámetros. Vallejo, Ato y Valdés (2008b) confirman la importancia de identificar el verdadero proceso generador de datos (PGD). En este estudio las tasas de error basadas en el verdadero PGD nunca excedían su valor nominal. Sin embargo, los errores estándar resultaban sesgados cuando se especificaba erróneamente el verdadero PGD.

La selección del modelo óptimo resulta central para interpretar adecuadamente los datos, no obstante, dicho objetivo es difícilmente alcanzable porque para una misma evidencia muestral existen múltiples modelos candidatos (Claeskens y Hjort, 2008). Para facilitar el modelado de la matriz de covarianza, SAS, probablemente el programa más popular y versátil de cuantos existen actualmente (Feng, Zhou, Zhang y Zhang, 2009), y otros programas estadísticos incorporan un completo menú de estructuras. Por ejemplo, *Proc Mixed* permite ajustar y comparar modelos de simetría compuesta, de esfericidad, autorregresivos, de media móvil, autorregresivos e integrados de media móvil, antedependientes y no estructurados (para detalles concretos, véase Zimmerman y Núñez-Antón, 2009). *Proc Mixed* también permite especificar estructuras de covarianza heterogéneas dentro y a través de los grupos, lo cual evita tener que aceptar la equicorrelación de las observaciones y la homogeneidad de las matrices de dispersión.

Existen diversos criterios para determinar la bondad de ajuste del modelo elegido durante el proceso de modelado. Para comparar modelos anidados (uno se puede obtener a partir de otro manipulando parámetros), el criterio más usado es el test de razón de verosimilitudes (LRT) con la desvianza obtenida a partir de la función de máxima verosimilitud completa (FML) o de máxima verosimilitud restringida/residual (REML), según se trate de elegir entre modelos con idéntica estructura de covarianza o de medias (Kreft y de Leeuw, 1998). También se suelen emplear herramientas estadísticas menos formales, tales como el Criterio de Información (IC) de Akaike (AIC), el AIC Corregido (AICC), el AIC Consistente (CAIC), el Criterio de Información Bayesiano (BIC) y el Criterio de Información Hannan-Quinn (HQIC), así como diversas versiones surgidas a partir de estos. Especialmente, los criterios AIC y BIC, por hallarse ambos implementados en la mayor parte de los programas que ajustan modelos mixtos; los programas específicos HLM y MLwiN constituyen una excepción a lo dicho. El origen de los IC es diferente, pero su estructura es similar; de hecho, difieren en el peso que asignan al factor de penalización (Lee y Ghosh, 2009). En mayor o menor medida, todos ellos penalizan el logaritmo de la función de verosimilitud por el número de parámetros, la mayor parte de las veces desde la formulación marginal del modelo (se ignoran explícitamente los efectos aleatorios a la hora modelar la variación de los datos multinivel), y eligen aquel modelo que minimiza el valor de los mismos. Vaida y Blanchard (2005) y Liang, Wu y Zou (2008) ofrecen detalles del comportamiento del criterio AIC usando una formulación jerárquica del modelo.

Otros criterios de selección, tales como el coeficiente de determinación ajustado (R^2_{adj}) el coeficiente de correlación de concordancia (CCC) o la suma de cuadrados residual de predicción (PRESS), han recibido escasa atención. No obstante, en uno de los pocos estudios que han examinado el desempeño de los criterios predictivos (basados en el ajuste de los valores predichos) usando la formulación marginal y jerárquica del modelo, Wang y Schaalje (2009) informan que los criterios (R^2_{adj}) CCC y PRESS no se comportaban mejor que los criterios AIC y BIC. La comparación se hacía entre dos modelos anidados con idéntica estructura de covarianza. Detalles técnicos de los criterios predictivos los proporcionan Orelien y Edwards (2007), Schabenberger (2004) y Vonesh, Chinchilli y Pu (1996).

En función de sus propiedades asintóticas los IC pueden ser clasificados en dos categorías: (a) criterios eficientes, tales como AIC o AICC y (b) criterios consistentes, tales como BIC, CAIC o HQIC. Se dice que un criterio es eficiente si la discrepancia entre el verdadero PGD y el modelo especificado para aproximarlo disminuye conforme aumenta el tamaño muestra. A su vez, se dice que un criterio es consistente si la probabilidad de elegir el modelo correcto aumenta conforme lo hace el tamaño de muestra. Los criterios eficientes parten de la hipótesis de que el verdadero PGD es dimensión infinita y seleccionan el mejor modelo de dimensión finita. En cambio, los consistentes parten de la hipótesis de que el verdadero PGD es dimensión finita y tienden a elegirlo siempre que el tamaño de muestra tienda a infinito. Cuando se apela al concepto de eficiencia asintótica no se asume que el verdadero PGD esté incluido dentro de la familia de modelos investigados. Sin embargo, cuando se apela al concepto de consistencia asintótica está implícita la hipótesis de que verdadero PGD pertenece a la clase de modelos considerados, lo cual puede ser falso.

Los análisis de contenido ponen de relieve que los IC más usados para elegir modelos con idéntica estructura de medias son el AIC y el BIC (Littell et al., 2000). El desempeño de estos criterios ha sido examinado por diversos autores, incluyendo Ferron, Dailey y Yi (2002), Gomez, Schaalje y Fellingham (2005), Keselman, Algina, Kowalchuk y Wolfinger (1998) y Vallejo et al. (2008b). Exceptuando el estudio de Ferron et al. (2002), donde el AIC identificó el verdadero PDG en el 79 % de las veces y el BIC en el 66%, los estudios restantes avalan críticamente la sugerencia efectuada por Littell et al. (2000) de modelar la estructura de covarianza con estos criterios, sobre todo, mediante el BIC. En el estudio de Keselman et al. (1999) el AIC seleccionó la estructura correcta en el 47% de las veces y el BIC en el 35%, en el estudio de Vallejo et al. (2008b) el AIC lo hizo en el 68% de las veces y el BIC el 48%, mientras que en el de Gomez et al. (2005) ambos criterios lo hicieron en el 22% de las veces. Auque el desempeño dependía de las condiciones manipuladas, en todos los estudios se puso de relieve que la selección mejoraba conforme aumentaba el tamaño de muestra y disminuía la complejidad de la matriz.

Un estudio más completo es el llevado a cabo por Gurka (2006). Este investigador examinó el desempeño de los criterios AIC, AICC, BIC y CAIC en términos de seleccionar el modelo de curva de crecimiento correcto bajo diversas condiciones, incluyendo diferentes formas de calcular los criterios y diferentes métodos de estimación de parámetros. Los IC fueron evaluados bajo tres escenarios dife-

rentes en función de su habilidad para: (a) seleccionar la estructura de medias correcta entre tres posibles modelos, dada una matriz CS; (b) seleccionar la estructura de covarianza correcta entre tres efectos aleatorios posibles con la misma estructura de medias; y (c) seleccionar el modelo correcto entre seis modelos que resultaban de combinar tres estructuras de medias con dos de covarianza. Los resultados obtenidos por Gurka muestran, entre otras cosas, que los IC basados en el método REML elegían el verdadero modelo de medias tan bien o mejor que los IC basados en el método FML; lo cual no deja de ser chocante, teniendo en cuenta que en la literatura estadística especializada (Molenberghs y Verbeke, 2001; Littell et al., 2006; Singer y Willet, 2003) se defiende ajustar dicha estructura vía FML exclusivamente. Gurka también halla que el desempeño de los criterios eficientes basados en el estimador REML mejoraba cuando se excluía del mismo el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ (en adelante $REML_2$); en cambio, el desempeño los criterios consistentes mejoraba cuando se mantenía dicho término ($REML_1$). Los resultados globales revelan que los criterios consistentes seleccionaban el verdadero PGD más del 89% de las veces, frente a los eficientes que lo hacían en torno al 81%.

Los hallazgos de Gurka (2006) afectan de lleno al proceso de selección de modelos, dado que ponen de relieve diversas inconsistencias existentes en la literatura estadística y en la documentación de los programas comerciales, tanto en lo referido a los métodos de estimación de parámetros como a las fórmulas usadas para calcular los IC. Conviene resaltar, no obstante, que sus estudios están basados en escenarios excesivamente simples, lo cual limita el alcance de sus resultados. Antes de proceder a generalizar los resultados de Gurka, sería muy clarificador investigar el desempeño de los IC, contemplando las mejoras analíticas referidas, cuando se manipula la distribución del término de error, la complejidad de las estructuras usadas para generar los datos y el número modelos incluidos en el proceso de selección.

Es razonable pensar que buena parte del elevado desempeño encontrado en el estudio de Gurka (p.e., todas las versiones de los IC examinados elegían la verdadera estructura de covarianza más del 90% de las veces) se explique por la forma de las matrices manipuladas, completamente generales *versus* excesivamente parcas, y por el reducido número de alternativas implicadas en el proceso de selección. En principio, asumido conocido el verdadero PGD, cabe esperar que el número de modelos incluidos en la comparación afecte más a los criterios eficientes que a los consistentes, ya que los primeros asumen que el verdadero PGD es de dimensión infinita; no obstante, es evidente que para ningún criterio será lo mismo seleccionar un modelo entre dos alternativas que entre dos docenas. Además, también cabe preguntarse ¿hasta qué punto resulta realista asumir que la varianza de las observaciones se mantiene constante y/o que la correlación no decrece a lo largo del tiempo cuando se estudian las curvas de crecimiento?

Por consiguiente, el presente trabajo tiene como objetivo determinar cuán efectivos son los criterios AIC, AICC, BIC, CAIC y HQIC para descubrir el verdadero PGD en una familia de modelos anidados. Estos criterios serán evaluados bajo estimación FML y $REML_1$/$REML_2$ cuando se manipulan diversas estructuras de medias y/o de covarianzas. Además, para proporcionar un punto de referencia para la comparación también utilizaremos el criterio de ajuste condicional LRT. En aras de dar respuesta al objetivo planteado, usaremos sendos diseños crossover en los que se violan separada y conjuntamente los supuestos de normalidad de los datos y de homogeneidad de las matrices de dispersión.

Definición de las herramientas usadas para seleccionar el mejor modelo mixto

Usar la metodología del modelo mixto en el contexto longitudinal, implica tener que elegir entre modelos alternativos para explicar la variabilidad observada en los datos del modo más sencillo posible. Aunque no existe unanimidad acerca de cual es la mejor forma de seleccionar el modelo óptimo, herramientas tales como los IC y el LRT son usadas frecuentemente.

Criterios de Información (IC)

En la Tabla 1 se definen las versiones de los IC investigados, tanto bajo estimación FML como bajo estimación $REML_1$/$REML_2$. También se indican las fórmulas empleadas por el módulo *Proc Mixed* del SAS (versión 9.2, 2008) y por el comando *Mixed* del SPSS (versión 17, 2008).

Test de razón de verosimilitudes (LRT)

Como ya ha sido indicado, el estadístico de bondad de ajuste más usado para comparar modelos anidados es el LRT. Este contraste puede obtenerse a partir de la expresión siguiente:

$$\Delta = -2[\hat{l}_{reducido(H_0)} - \hat{l}_{completo(H_1)}],$$

donde Δ es el estadístico desvianza y $\hat{l}_{reducido(H_0)}$ y $\hat{l}_{completo(H_1)}$ los máximos de la función FML o REML, según se trate de elegir entre modelos con idéntica estructura de covarianza o de medias, bajo la hipótesis nula y alternativa, respectivamente. El estadístico Δ se distribuye bajo H_0 según χ^2_v donde v indica la diferencia entre el número de parámetros estimados en el modelo completo y en el modelo reducido.

A pesar de la amplia utilización del LRT, su uso conlleva ciertas limitaciones que es preciso tener en cuenta. Por ejemplo, únicamente está definido para comparar modelos anidados y tan sólo permite comparar dos al mismo tiempo. Cuando el número de modelos anidados sea superior a dos, la aplicación del LRT requiere proceder jerárquicamente (para más detalles, véase Dayton, 2003). Por el contrario, es importante destacar que los IC son válidos para comparar y seleccionar modelos anidados y no anidados. Además, permiten la comparación simultánea de un conjunto de modelos.

Método de la simulación

Para evaluar el desempeño de los métodos descritos realizamos tres estudios de simulación. En el primero mantuvimos constante la estructura de covarianza y modelamos la estructura de medias. En el segundo supusimos conocida la estructura de medias y modelamos la de covarianza. En el tercero modelamos ambas estructuras a la vez. En cada una de ellos utilizamos un diseño crossover con dos tratamientos, dos secuencias y seis (en el tercer estudio también doce) periodos de una semana, en el que se violaban separada y conjuntamente los supuestos de normalidad y esfericidad multimuestral. Los participantes del primer grupo recibieron la secuencia de tratamiento AAABBB, mientras que los del segundo grupo recibieron la secuencia inversa para contrarrestar los posibles efectos residuales. En base a lo expuesto, se planteó el modelo de la forma

Tabla 1
Definición de los criterios de información usados en la selección del modelo mixto

$AIC = -2\log l_{FML} + 2(p+q)^{SAS,SPSS}$	$AIC_1 = -2\log l_{REML1} + 2q$
	$AIC_2 = -2\log l_{REML2} + 2q^{SAS,SPSS}$
$AICC_1 = -2\log l_{FML} + 2(p+q)\left(\frac{N}{N-p-q-1}\right)^{SAS,SPSS}$	$AICC_1 = -2\log l_{REML1} + 2q\left(\frac{N-p}{N-p-q-1}\right)$
$AICC_2 = -2\log l_{FML} + 2(p+q)\left(\frac{n}{n-p-q-1}\right)$	$AICC_2 = -2\log l_{REML1} + 2q\left(\frac{n}{n-q-1}\right)$
	$AICC_1 = -2\log l_{REML2} + 2q\left(\frac{N-p}{N-p-q-1}\right)^{SAS,SPSS}$
	$AICC_2 = -2\log l_{REML2} + 2q\left(\frac{n}{n-q-1}\right)$
$BIC_1 = -2\log l_{FML} + (p+q)\log(N)^{SPSS}$	$BIC_1 = -2\log l_{REML1} + q\log(N-p)$
$BIC_2 = -2\log l_{FML} + (p+q)\log(n)^{SAS}$	$BIC_2 = -2\log l_{REML1} + q\log(n)$
	$BIC_1 = -2\log l_{REML2} + q\log(N-p)^{SPSS}$
	$BIC_2 = -2\log l_{REML2} + q\log(n)^{SAS}$
$CAIC_1 = -2\log l_{FML} + (p+q)[\log(N)+1]^{SPSS}$	$CAIC_1 = -2\log l_{REML1} + q[\log(N-p)+1]$
$CAIC_2 = -2\log l_{FML} + (p+q)[\log(n)+1]^{SAS}$	$CAIC_2 = -2\log l_{REML1} + q[\log(n)+1]$
	$CAIC_1 = -2\log l_{REML2} + q[\log(N-p)+1]^{SPSS}$
	$CAIC_2 = -2\log l_{REML2} + q[\log(n)+1]^{SAS}$
$HQIC_1 = -2\log l_{FML} + 2(p+q)\log[\log(N)]$	$HQIC_1 = -2\log l_{REML1} + 2q\log[\log(N-p)]$
$HQIC_2 = -2\log l_{FML} + 2(p+q)\log[\log(n)]^{SAS}$	$HQIC_2 = -2\log l_{REML1} + 2q\log[\log(n)]$
	$HQIC_1 = -2\log l_{REML2} + 2q\log[\log(N-p)]$
	$HQIC_2 = -2\log l_{REML2} + 2q\log[\log(n)]^{SAS}$

Nota: p = número de parámetros del modelo de medias; q = número de parámetros de la estructura de covarianza; n = número total de sujetos; N = número total de observaciones; FML = estimador de máxima verosimilitud completa; $REML_1$/ $REML_2$ = estimadores de máxima verosimilitud residual con y sin el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$

$y_{ij}=\beta_{00}+\beta_{01}G_j+u_{0j}+\beta_{10}T_{ij}+\beta_{11}G_j\times T_{ij}+u_{1j}+\beta_{20}CT_{ij}+\beta_{21}G_j\times CT_{ij}+u_2 \in CT_{ij}+e_{ij}$, donde $G_j=0$ si el *i*-ésimo participante era asignado a la secuencia AAABBB durante un periodo de seis semanas y $G_j=1$ si era asignado a la secuencia BBBAAA; T_{ij} denota la semana en la cual se registraba la respuesta y CT_{ij} denota el cambio de tendencia lineal entre los tres primeros periodos y los tres últimos. Las variables $T_{ij} \in \{1, 2, 3, 4, 5, 6\}$ y $CT_{ij} \in \{1, 2, 3, 1, 2, 3\}$ fueron centradas con respecto a sus respectivas medias, concretamente $T_{ij}^{\bullet}=(T_{ij}-3.5)$ y $CT_{ij}^{\bullet}=(CT_{ij}-2)$.

Variables manipuladas en el primer estudio

En el primer estudio se evaluó el desempeño de los criterios de selección para elegir de un conjunto de modelos candidatos la ver-

Table 2
Modelos usados para ajustar la estructura de medias y valor de los parámetros de efectos fijos

$M_1^{©}$ $y_{ij} = \beta_{00} + u_{0j} + e_{ij}$

M_2 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + e_{ij}$

M_3 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$

M_4 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + e_{ij}$

M_5 $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$

$M_6^{©}$ $y_{ij} = \beta_{00} + \beta_{01}G_j + u_{0j} + \beta_{10}T_{ij}^{\bullet} + \beta_{11}G_j \times T_{ij}^{\bullet} + u_{1j}T_{ij}^{\bullet} + \beta_{20}CT_{ij}^{\bullet} + \beta_{21}G_j \times CT_{ij}^{\bullet} + u_{2j}CT_{ij}^{\bullet} + e_{ij}$

$\beta_{\text{completo}}^{©} = \left[\beta_{00} = 3.125\ \beta_{01} = 1.25\ \beta_{10} = -.20\ \beta_{11} = 0.45\ \beta_{20} = -.25\ \beta_{21} = 0.50\right]$

$\beta_{\text{reducido}}^{©} = \left[\beta_{00} = 3.125\ \beta_{01} = 0.00\ \beta_{10} = 0.00\ \beta_{11} = 0.00\ \beta_{20} = 0.00\ \beta_{21} = 0.00\right]$

$M_6^{©}, M_1^{©}$ = modelos completo y reducido usados para generar los datos

dadera estructura de medias. Dicha evaluación fue realizada bajo estimación FML y $REML_1$/$REML_2$ cuando se manipulaban las variables siguientes:

(a) *Tipo de modelo usado para generar los datos*. El ajuste de la estructura de medias implicaba seleccionar de un conjunto de seis modelos anidados el verdadero PGD. En la mitad de las condiciones manipuladas, dicho proceso requería ajustar un modelo completo y en la otra mitad uno reducido. En ambos casos, los efectos fijos del modelo fueron definidos combinando distintas matrices de diseño y distintos vectores de parámetros. Bajo la primera situación, además del intercepto, el modelo que generó los datos (M_6) incluía como efectos fijos los grupos de tratamiento (G), la tendencia lineal (T), el cambio de tendencia (CT), la interacción G × T y la interacción G × CT. Los cinco modelos restantes fueron especificados erróneamente eliminando de la matriz de diseño una o más covariadas. Por ejemplo, el M5 fue especificado erróneamente eliminando la covariada G × CT, mientras que el M1 lo fue eliminando las covariadas G, T, CT, G × T y G × CT, respectivamente. Bajo la segunda situación, el modelo usado para generar los datos fue el M_1. Los modelos restantes fueron especificados erróneamente siguiendo un proceso inverso al descrito, es decir, añadiendo variables a la matriz de diseño. En la Tabla 2 aparecen recogidos los modelos usados en el proceso de comparación, así como el valor de los parámetros de efectos fijos de los modelos que generaron los datos. El vector de coeficientes del modelo completo, ligeramente modificado, se corresponde con el de un experimento descrito por Hedeker y Gibbons (2006; páginas 122-126).

(b) *Tamaño de muestra total*. El desempeño fue investigado usando dos tamaños de muestra distintos: $n = 30$ y $n = 60$. Estos tamaños grupales fueron seleccionados por ser representativos de los encontrados frecuentemente en las investigaciones psicológicas. Dentro de cada tamaño de muestra, el valor del coeficiente de variación muestral Δ se fijó en 0.33, donde $\Delta = \frac{1}{\bar{n}}[S_j(n_j - \bar{n})^2 / J]^{1/2}$ siendo $\bar{n}$ el tamaño promedio de los grupos. Cuando el diseño está equilibrado, Δ=0 Para n=30 los tamaños grupales fueron: (10-20), (15-15) y (20-10), mientras que para $n = 60$ los tamaños grupales fueron: (20-40), (30-30) y (40-20).

(c) *Patrones de covarianza empleados para generar los datos*. Los patrones utilizados para generar los datos fueron tres, a saber: coeficientes aleatorios lineales (RCL), autorregresivo de primer orden heterogéneo [ARH(1)] y UN. El primer patrón es un ejemplo de modelo jerárquico que permite modelar un intercepto y una o más tendencias más para cada participante. Este patrón puede resultar muy útil para caracterizar los datos, dado que aúna flexibilidad y parquedad, de hecho, tan sólo requiere estimar parámetros. En este estudio q=3 el intercepto, la tendencia lineal y el cambio de tendencia. El segundo patrón permite que las varianzas sean heterogéneas y que las covarianzas decrezcan exponencialmente, pero asume que las observaciones se hallan igualmente espaciadas entre sí. Este modelo es típico de las series temporales cortas y precisa estimar (t+1) parámetros. Por su parte, el tercer patrón representa la estructura de covarianza que mejor se ajusta a los datos, además no exige que las observaciones se encuentren igualmente espaciadas. No obstante, requiere estimar t (t+1)/2 parámetros.

(d) *Desviación del supuesto de esfericidad*. Aunque el modelo mixto no asume que las varianzas de las diferencias entre pares de medidas repetidas sean iguales (supuesto de esfericidad), la investigación empírica ha puesto de relieve que las inferencias realizadas con este enfoque sí pueden verse afectadas por la falta de esfericidad (Vallejo, Fernández, Herrero y Conejo, 2004). Por este motivo, patrones de covarianza con valores de ε (índice de ausencia de esfericidad derivado por Box) de .47 y .70 fueron empleados para investigar sus efectos en el desempeño de los criterios de selección. Las estructuras de covarianza usadas están disponibles en la Web http://gip.uniovi.es/gdiyad/docume/Psicothema/.

(e) *Igualdad de las matrices de dispersión*. El desempeño de las herramientas de selección fue evaluado cuando las matrices de covarianza grupales eran homogéneas y también cuando eran heterogéneas. En el primer caso, los elementos de las dos matrices de dispersión fueron iguales entre sí (Σ_2=Σ_1) mientras que en el segundo caso, los elementos de una de las matrices fueron cinco veces mayores que los de la otra (Σ_2=5Σ_1).

(f) *Emparejamiento de las matrices de covarianza y el tamaño de los grupos*. La forma de relacionar el tamaño de los grupos y el tamaño de las matrices de dispersión pueden tener diferentes efectos en las pruebas estadísticas. Cuando el diseño está equilibrado, la relación entre el tamaño de las matrices de dispersión y el tamaño de los grupos es nula. Cuando el diseño está desequilibrado, la relación puede ser positiva o negativa. Una relación positiva implica que el grupo de menor tamaño se asocia con la matriz de dis-

persión menor, mientras que una relación negativa implica que el grupo de menor tamaño se asocia con la matriz de dispersión mayor.

(g) *Forma de la distribución de la variable de medida*. Aunque el enfoque del modelo mixto está basado en el cumplimiento del supuesto de normalidad, cuando se trabaja con datos reales es común que los índices de asimetría γ_1 y curtosis γ_2 se desvíen de cero (Micceri, 1989), lo cual puede inducirnos a interpretar incorrectamente los resultados. Para investigar el efecto que ejerce forma de la distribución en el desempeño de los criterios de selección, generamos datos desde distribuciones normales y no normales mediante las distribuciones *g* y *h* introducidas por Tukey (1977). Además de la distribución normal (g = h = 0; $\gamma_1 = \gamma_2 = 0$), también investigamos otras tres: (a) $g = 0$ y $h = .109$, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial doble o de Laplace ($\gamma_1 = 0$ & $\gamma_2 = 3$); (b) $g = .76$ y $h = -.098$, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial ($\gamma_1 = 2$ & $\gamma_2 = 6$) y (c) $g = 1$ y $h = 0$, una distribución que tiene el mismo grado de sesgo y de curtosis que la distribución lognormal ($\gamma_1 = 6.18$ & $\gamma_2 = 110.94$) Las distribuciones *g* y *h* fueron obtenidas utilizando la función RANNOR del SAS. Mediante ella generamos variables aleatorias normales estándar (Z_{ijk}) y transformamos cada una de ellas como $Z^{*}_{ijk} = g^{-1}[\exp(gZ_{ijk}) - 1]\exp(hZ^{2}_{ijk}/2)$ donde *g* y *h* son números reales que controlan el grado sesgo y de curtosis. Por último, para obtener una distribución con desviación estándar σ_{jk} cada una de las puntuaciones que conforman la variable dependiente fue creada utilizando el modelo lineal $Y_{ijk} = \sigma_{jk} \times (Z^{*}_{ijk} - \mu_{gh})$ donde $\mu_{gh} = \{\exp[g^2/(2-2h)] - 1\}/[g(1-h)^{1/2}]$ es la media de la de la distribución *g* y *h* (para detalles véase Kowalchuk y Headrick, 2009).

Variables manipuladas en el segundo estudio

En este estudio se evaluó el desempeño los criterios de selección para elegir de un conjunto de modelos candidatos la verdadera estructura de covarianza. Dicho ajuste implicaba seleccionar de un conjunto de seis patrones anidados el verdadero PGD. En la mitad de las condiciones manipuladas, se requería ajustar un modelo en el cual la varianza se mantenía constante a lo largo del tiempo y la covarianza decrecía exponencialmente (AR(1)) y en la otra mitad un modelo UN. Bajo la primera situación, además del modelo AR(1) usado para generar los datos, el conjunto de modelos candidatos incluía un modelo de independencia (IND), un modelo ARH(1), un modelo Toeplitz homogéneo (TOEP), un modelo TOEP heterogéneo (TOEPH) y un modelo UN. El modelo IND asume varianza constante y covarianza serial nula, mientras que los modelos TOEP y TOEPH generalizan, respectivamente, a los modelos AR(1) y ARH(1). Diversos investigadores ofrecen una descripción detallada de estos modelos, incluyendo Fitzmaurice et al. (2004), Littell et al. (2006) y Zimmerman y Núñez-Antón (2009). Repárese que las estructuras están anidadas unas dentro de otras, en el sentido que IND es un caso especial de AR(1), ésta lo es de TOPH, la cual a su vez lo es de TOEPH y ésta última lo es necesariamente de UN.

Bajo la segunda situación, el conjunto de modelos candidatos era idéntico al descrito, pero los datos fueron generados a partir del modelo UN. Además de los métodos de estimación y de los patrones de covarianza usados en la generación de los datos, también fueron manipuladas las variables tamaño de muestra, igualdad de las matrices de dispersión y forma de la distribución de la población. Las estructuras de covarianza usadas están disponibles en la citada Web.

Variables manipuladas en el tercer estudio

Para profundizar en el desempeño de las pruebas ajustamos simultáneamente la estructura de medias y la estructura de covarianza. Dicho ajuste implicaba seleccionar de un conjunto de nueve modelos candidatos el verdadero PGD. En la Tabla 3 aparecen recogidos los modelos utilizados en la comparación, así como el valor de los parámetros de efectos fijos usados para generar los datos. Examinando la Tabla 3 se aprecia que los modelos estaban anidados unos dentro de otros, en aquellos casos que el número de

Tabla 3
Conjunto de modelos de medias y de covarianza candidatos y valor de los parámetros de efectos fijos

M_1	$E(y_{ij}) = \beta_{00}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_2	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_3	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{AR}(1)]$
M_4	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_5	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_6	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ARH}(1)]$
M_7	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
$M^{©}_8$	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}CT^{\bullet}_{ij} + \beta_{21}G_j \times CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
M_9	$E(y_{ij}) = \beta_{00} + \beta_{01}G_j + \beta_{10}T^{\bullet}_{ij} + \beta_{11}G_j \times T^{\bullet}_{ij} + \beta_{20}T^{2\bullet}_{ij} + \beta_{30}CT^{\bullet}_{ij} + \beta_{31}G_j \times CT^{\bullet}_{ij}$	$Var(y_{ij}) = \mathbf{V}_i[\text{ANTE}(1)]$
$\beta' = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -0.50\ \beta_{11} = 0.50\ \beta_{20} = -0.50\ \beta_{21} = 0.50]$		
$\beta' = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -1.00\ \beta_{11} = 1.00\ \beta_{20} = -1.00\ \beta_{21} = 1.00]$		

Nota: $M_1 \subset M_2 \subset M_3 \subset M_4 \subset M_5 \subset M_6 \subset M_7 \subset M^{©}_8 \subset M_9$; $M^{©}_8$ = modelo usado para generar los datos

efectos fijos era idéntico, como sucedía con los modelos $M_3 - M_4$ y $M_6 - M_7$, las estructuras de covarianza diferían y se hallaban anidadas entre sí, en el sentido que AR(1) es un caso especial de ARH(1) la cual es a su vez un caso especial de ANTE(1). Para una exhaustiva descripción de esta última estructura, véase Zimmerman y Núñez-Antón (2009).

En este tercer estudio, además de los métodos de estimación, también fueron manipuladas las variables tamaño de muestra, número de medida repetidas (t = 6 y t = 12), valor de los parámetros de efectos fijos, igualdad de las matrices de dispersión y forma de la distribución. Los valores de los parámetros de covarianza de la matriz ANTE(1) están disponibles en laWeb citada.

Resultados del primer estudio

La tabla 4 recoge el porcentaje de veces que los 29 criterios examinados, 10 vía FML y 19 vía REML, seleccionaban el verdadero modelo de medias cuando la estructura usada para generar los

Tabla 4
Porcentaje de veces que los criterios elegían el modelo de medias verdadero cuando el patrón de covarianza conocido era ARH (1)

		Modelo Completo					Modelo Reducido				
ME	Criterio	Norm	Lapla	Expon	Logn	Media	Norm	Lapla	Expon	Lognor	Media
FML	$AIC^{(SAS, SPSS)}$	59.24	49.31	38.69	30.26	**44.37** [1]	68.73	71.66	65.08	49.23	**63.68**
FML	$AICC_1$	26.54	17.37	12.33	08.21	**16.11**	93.61	94.69	91.54	61.78	**85.40**
FML	$AICC_2^{(SAS, SPSS)}$	54.23	44.43	33.85	25.34	**39.46** [2]	75.76	76.60	70.13	52.28	**68.69**
FML	$HQIC_1$	41.96	32.65	22.44	16.72	**28.44**	88.29	90.43	86.18	66.17	**82.77**
FML	$HQIC_2^{(SAS)}$	50.70	38.37	30.27	21.50	**35.21** [3]	78.63	82.12	77.64	63.09	**50.37**
FML	$BIC_1^{(SPSS)}$	23.46	15.72	07.69	05.15	**13.00**	95.97	97.38	95.15	82.23	**92.68** [2]
FML	$BIC_2^{(SAS)}$	38.82	29.10	19.28	13.02	**25.06**	89.55	91.74	87.44	69.50	**84.56**
FML	$CAIC_1^{(SPSS)}$	17.28	10.61	05.01	02.49	**08.85**	97.94	98.73	96.79	85.37	**94.70** [1]
FML	$CAIC_2^{(SAS)}$	30.45	21.30	11.96	07.12	**17.69**	93.99	95.93	92.75	76.83	**89.87** [3]
FML	LRT	38.53	29.71	20.94	14.74	**25.98**	85.23	87.09	82.10	62.79	**79.30**
$REML_1$	AIC_1	99.99	99.97	96.71	87.36	**96.01**	91.87	93.08	90.81	78.09	**88.46**
$REML_2$	$AIC_2^{(SAS, SPSS)}$	85.30	88.60	84.75	74.54	**83.30**	61.52	65.74	58.98	38.70	**56.23**
$REML_1$	$AICC_1$	99.99	99.49	99.04	94.12	**98.16**	93.59	94.12	**92.51**	80.02	**90.06**
$REML_1$	$AICC_2$	75.00	75.01	75.26	75.38	**75.16**	98.50	98.46	97.78	90.12	**96.22**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	85.28	88.54	89.98	80.29	**86.02**	69.69	71.58	62.53	41.49	**61.32**
$REML_2$	$AICC_2$	63.52	65.61	68.92	70.84	**67.23**	85.85	85.61	58.07	47.73	**69.31**
$REML_1$	$HQIC_1$	99.99	99.99	99.88	98.50	**99.59**	97.87	96.41	96.69	87.40	**94.59**
$REML_1$	$HQIC_2$	99.99	99.98	99.27	94.67	**98.48**	95.25	96.55	93.98	82.78	**92.14**
$REML_2$	$HQIC_1$	85.33	88.86	92.49	94.06	**90.18**	88.42	88.95	76.04	51.95	**76.34**
$REML_2$	$HQIC_2^{(SAS)}$	85.42	88.74	89.49	84.50	**87.04**	79.42	80.81	67.84	45.45	**68.38**
$REML_1$	BIC_1	99.99	99.99	99.99	99.74	**99.93** [3]	99.68	98.68	98.39	93.25	**97.50** [2]
$REML_1$	BIC_2	99.99	99.99	99.93	98.15	**99.52**	98.26	98.61	97.10	88.41	**95.60**
$REML_2$	$BIC_1^{(SPSS)}$	87.24	89.59	92.69	94.28	**90.95**	90.15	88.29	73.49	49.84	**75.44**
$REML_2$	$BIC_2^{(SAS)}$	85.38	88.72	91.59	91.91	**89.40**	87.98	86.39	71.79	40.39	**71.64**
$REML_1$	$CAIC_1$	99.99	99.99	99.99	99.96	**99.99** [1]	**99.82**	99.79	99.01	93.90	**98.13** [1]
$REML_1$	$CAIC_2$	99.99	99.99	99.99	99.75	**99.93** [2]	95.46	99.40	98.32	91.76	**96.23** [3]
$REML_2$	$CAIC_1^{(SPSS)}$	86.83	89.72	92.77	94.42	**90.94**	90.16	88.51	73.40	49.79	**70.46**
$REML_2$	$CAIC_2^{(SAS)}$	85.40	88.73	92.40	93.20	**89.93**	89.50	86.97	72.15	48.47	**74.27**
$REML_2$	LRT	15.39	12.05	10.65	09.94	**12.01**	41.68	55.08	58.68	56.97	**53.10**

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables naturaleza del emparejamiento, tamaño de muestra, desviación de la esfericidad e igualdad de las matrices de dispersión; ME = método de estimación; FML = estimación por máxima verosimilitud completa; $REML_1$ = estimación por máxima verosimilitud residual incorporando el término aditivo; $REML_2$ = estimación por máxima verosimilitud residual eliminando el término aditivo

datos era ARH(1). El patrón de resultados correspondiente a las matrices RCL y UN no aparece recogido. Observando la Web citada se aprecia que dicho patrón era cualitativa y cuantitativamente similar al descrito. Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables igualdad de las matrices de dispersión y tamaño de muestra. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía de la forma de la distribución de la variable de medida, tipo de modelo a seleccionar y procedimiento de estimación utilizado. Aunque no aparece recogido en la tabla 4, las variables naturaleza del emparejamiento, desviación de la esfericidad e igualdad de las matrices de dispersión afectaron ligeramente al porcentaje de identificaciones correctas y el tamaño de muestra sustancialmente.

2. El desempeño de los IC era más elevado bajo estimación REML que bajo estimación FML. Promediando a través de las 2592 (144 × 18) condiciones manipuladas, el porcentaje de aciertos obtenidos vía REML fue del 86.5%, mientras que el obtenido vía FML promediando a través de 1296 (144 × 9) condiciones fue del 52.3%. Con el LRT sucedió lo contrario.

Tabla 5
Porcentaje de veces que los criterios elegían la estructura de covarianza verdadera bajo estimación FML y REML

		Patrón Autorregresivo [AR(1)]					Patrón General (UN)				
ME	Criterio	Norm	Lapla	Expon	Logn	Media	Norm	Lapla	Expon	Lognor	Media
FML	$AIC^{(SAS, SPSS)}$	77.91	53.77	12.60	04.21	**37.12**	56.37	59.18	86.06	92.92	**73.63** [2]
FML	$AICC_1$	72.22	67.06	33.82	17.00	**47.53**	25.79	26.13	35.19	39.19	**31.60**
FML	$AICC_2^{(SAS, SPSS)}$	84.33	64.80	17.80	06.74	**43.42**	32.61	34.07	63.59	77.57	**51.96** [3]
FML	$HQIC_1$	96.23	87.53	38.39	17.92	**60.02**	07.35	08.54	38.57	49.19	**25.92**
FML	$HQIC_2^{(SAS)}$	90.51	74.25	24.47	08.89	**49.53**	22.96	25.92	64.47	74.09	**46.86**
FML	$BIC_1^{(SPSS)}$	99.52	97.20	63.09	35.31	**73.78** [2]	00.05	00.13	08.02	14.25	**05.61**
FML	$BIC_2^{(SAS)}$	96.34	89.70	41.54	18.94	**61.63**	02.53	04.13	30.20	37.92	**18.70**
FML	$CAIC_1^{(SPSS)}$	99.89	98.73	72.42	43.71	**78.69** [1]	00.00	00.06	04.03	08.05	**03.04**
FML	$CAIC_2^{(SAS)}$	99.43	95.24	54.98	28.39	**69.51** [3]	00.18	00.76	14.01	20.17	**08.78**
FML	LRT	67.77	42.12	09.12	03.45	**30.61**	84.12	82.17	94.18	97.05	**89.38** [1]
$REML_1$	AIC_1	83.69	34.41	34.13	15.79	**42.00**	50.74	58.73	82.63	89.32	**70.36** [2]
$REML_2$	$AIC_2^{(SAS, SPSS)}$	83.69	34.41	34.13	15.79	**42.00**	50.74	58.78	82.62	89.31	**70.37** [3]
$REML_1$	$AICC_1$	86.44	39.60	45.58	25.24	**49.21**	32.85	38.22	65.83	76.65	**53.39**
$REML_1$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.14	41.29	44.16	**37.02**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	86.44	39.59	45.57	25.24	**49.21**	32.85	38.27	65.83	76.65	**53.40**
$REML_2$	$AICC_2$	70.05	40.30	51.32	37.53	**49.81**	30.50	32.13	41.29	44.36	**37.07**
$REML_1$	$HQIC_1$	97.93	68.99	77.26	54.75	**74.73**	04.97	09.48	36.21	53.47	**26.03**
$REML_1$	$HQIC_2$	93.68	52.42	58.38	33.53	**59.51**	18.76	25.73	61.52	75.25	**45.32**
$REML_2$	$HQIC_1$	97.93	68.66	77.22	54.73	**74.64**	04.47	08.14	35.88	53.28	**25.44**
$REML_2$	$HQIC_2^{(SAS)}$	93.68	52.42	58.38	33.53	**59.50**	18.76	25.69	61.48	75.23	**45.29**
$REML_1$	BIC_1	99.94	88.44	93.95	81.83	**91.04** [3]	00.00	00.04	06.87	17.28	**06.05**
$REML_1$	BIC_2	98.45	73.13	82.08	61.17	**78.73**	00.99	03.90	27.31	44.94	**19.29**
$REML_2$	$BIC_1^{(SPSS)}$	99.94	88.39	93.43	81.66	**90.86**	00.00	00.04	06.87	17.17	**06.02**
$REML_2$	$BIC_2^{(SAS)}$	98.45	73.21	82.08	61.16	**78.71**	00.99	03.89	27.31	44.91	**19.28**
$REML_1$	$CAIC_1$	99.99	92.51	96.20	87.41	**94.03** [1]	00.00	00.08	03.37	09.93	**03.35**
$REML_1$	$CAIC_2$	99.67	83.80	90.38	75.17	**87.50**	00.00	00.44	12.19	25.54	**09.55**
$REML_2$	$CAIC_1^{(SPSS)}$	98.98	92.49	96.18	87.40	**94.01** [2]	00.00	00.08	03.36	09.93	**03.35**
$REML_2$	$CAIC_2^{(SAS)}$	99.66	83.79	90.38	75.16	**87.25**	00.00	00.44	12.18	25.54	**09.55**
$REML_2$	LRT	78.11	25.50	17.60	07.87	**32.28**	81.45	82.78	92.57	95.46	**88.06** [1]

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables tamaño de muestra e igualdad de las matrices de dispersión

€

3. En promedio, las diferencias existentes entre los IC y el LRT bajo estimación FML fueron mínimas. En concreto, los IC seleccionaron correctamente el modelo completo en el 25.4% de las veces y el reducido en el 79.2%, mientras que el LRT lo hizo en el 26% y 79.3% de las veces. Sin embargo, bajo estimación REML los IC seleccionaron correctamente el modelo completo en el 91.2% de las veces y el reducido en el 81.8%, mientras que el LRT lo hizo en el 12% y 53.1% de las veces, respectivamente.

4. Con relación al desempeño de los IC bajo estimación REML, los criterios consistentes elegían el verdadero modelo de medias en el 89.4% de las veces y los eficientes el 80.7%. Además, el desempeño de ambas clases de criterios mejoraba cuando se incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$ en la ecuación. El porcentaje de aciertos de los criterios consistentes era del 98.1% bajo $REML_1$ y del 81.2% bajo $REML_2$. Por su parte, los criterios eficientes elegían el verdadero PGD el 90% de las veces bajo $REML_1$ y el 75.5% bajo $REML_2$.

Tabla 6
Porcentaje de veces que los criterios elegían correctamente el modelo de efectos fijos y aleatorios bajo estimación FML y REML

		t = 6		*t* = 12		
ME	Criterio	$\beta'_{(1)}$	$\beta'_{(2}$	$\beta'_{(1)}$	$\beta'_{(2)}$	Global
FML	$AIC^{(SAS, SPSS)}$	38.94[1]	67.01[2]	73.92	79.44	**64.8[2]**
FML	$AICC_1$	08.62	24.53	20.75	24.70	**19.7**
FML	$AICC_2^{(SAS, SPSS)}$	33.59[2]	64.31	74.74[3]	83.22[3]	**64.0**
FML	$HQIC_1$	20.74	55.14	69.62	89.10	**58.7**
FML	$HQIC_2^{(SAS)}$	30.64	64.23[3]	77.03[2]	84.56	**64.1[3]**
FML	$BIC_1^{(SSPSS)}$	05.54	25.84	33.32	70.72	**33.9**
FML	$BIC_2^{(SAS)}$	17.17	50.51	70.91	89.84[2]	**57.1**
FML	$CAIC_1^{(SPSS)}$	02.74	16.98	22.82	57.16	**24.9**
FML	$CAIC_2^{(SAS)}$	09.52	34.71	55.95	85.54	**46.4**
FML	LRT	31.89[3]	73.75[1]	87.93[1]	99.16[1]	**73.2[1]**
$REML_1$	AIC_1	81.51[1]	82.87	88.59	89.36	**85.6[1]**
$REML_2$	$AIC_2^{(SAS, SPSS)}$	74.21[2]	87.27[1]	92.68[1]	98.93[1]	**88.3[2]**
$REML_1$	$AICC_1$	77.34[3]	81.38	88.25	89.39	**84.1**
$REML_1$	$AICC_2$	42.99	54.64	44.14	44.48	**46.6**
$REML_2$	$AICC_1^{(SAS, SPSS)}$	63.58	85.08[2]	90.91[2]	98.91[2]	**84.6[3]**
$REML_2$	$AICC_2$	37.22	56.31	46.10	49.43	**47.3**
$REML_1$	$HQIC_1$	59.98	74.79	79.13	89.59	**75.9**
$REML_1$	$HQIC_2$	73.21	81.04	87.47	89.22	**82.7**
$REML_2$	$HQIC_2$	51.50	76.42	81.90	98.32	**77.1**
$REML_2$	$HQIC_2^{(SAS)}$	60.76	84.18[3]	91.03[3]	98.87[3]	**83.7**
$REML_1$	BIC_1	32.28	47.20	42.69	80.73	**50.7**
$REML_1$	BIC_2	54.90	73.73	80.09	87.12	**73.9**
$REML_2$	$BIC_1^{(SPSS)}$	27.13	47.99	42.52	87.77	**51.4**
$REML_2$	$BIC_2^{(SAS)}$	46.73	73.11	80.98	98.79	**74.9**
$REML_1$	$CAIC_1$	23.35	43.77	32.04	73.29	**43.1**
$REML_1$	$CAIC_2$	40.34	59.66	65.73	88.20	**63.5**
$REML_2$	$CAIC_1^{(SPSS)}$	20.16	35.62	32.11	79.41	**41.8**
$REML_2$	$CAIC_2^{(SAS)}$	34.51	58.58	63.99	97.62	**63.7**
$REML_2$	LRT	28.33	71.11	79.59	96.89	**68.9**

Nota: Los datos denotan el porcentaje promedio de elecciones correctas a través del tamaño de muestra, forma de la distribución e igualdad de las matrices de dispersión. $\beta'_{(1)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -0.50\ \beta_{11} = 0.50\ \beta_{21} = -0.50\ \beta_{21} = 0.50]$; $\beta'_{(2)} = [\beta_{00} = 1.00\ \beta_{01} = 1.25\ \beta_{10} = -1.00\ \beta_{11} = 1.00\ \beta_{10} = -1.00\ \beta_{21} = 1.00]$, *t* = número de periodos de observación. La última columna representa el porcentaje promedio a través de los tres experimentos.

Resultados del segundo estudio

En la tabla 5 aparece el porcentaje de veces que los criterios examinados elegían la estructura de covarianza verdadera, cuando los datos fueron generados desde sendas matrices AR(1) y UN. Los datos tabulados denotan el porcentaje promedio de elecciones correctas a través de las variables igualdad de las matrices y tamaño de muestra. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía del patrón de covarianza usado para generar los datos y de la forma de la distribución. La influencia del los procedimientos de estimación era menor. Aunque no se recoge en la Tabla 5, los detalles se encuentran en la Web antes citada, las variables igualdad de las matrices de dispersión y tamaño de muestra afectaban sustancialmente al porcentaje de identificaciones correctas cuando el modelo usado para generar los datos había sido el UN, pero escasamente cuando se usaba el AR(1).

2. El desempeño del LRT fue superior al de los IC, tanto bajo estimación FML como REML. En promedio el LRT seleccionó correctamente el verdadero PGD en el 60% de las veces (el 31% bajo ARH(1) y el 89 % bajo UN), mientras que los IC lo hicieron en el 47.2%.

3. Cuando el modelo utilizado para generar los datos fue el AR(1), el porcentaje de aciertos disminuía conforme los datos se desviaban de la normalidad. El fenómeno contrario se producía con los datos generados bajo el modelo UN.

4. Con independencia del procedimiento de estimación utilizado, los criterios consistentes se comportaban mejor que sus homólogos eficientes cuando el modelo utilizado para generar los datos fue el AR(1). El porcentaje de aciertos de los criterios consistentes fue del 73.2% y el de los eficientes del 44.9%. Por el contrario, la situación se invertía cuando el modelo utilizado para generar los datos fue el UN. Los criterios consistentes elegían el verdadero PGD en el 18.1% de las veces y los eficientes en el 52.7%. A diferencia de lo encontrado en el estudio anterior, el desempeño de ambas clases de criterios no mejoraba cuando el estimador REML incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$.

Resultados del tercer estudio

En la tabla 6 aparece tabulado el porcentaje de veces que los criterios examinados elegían correctamente la estructura de medias y de covarianza, tanto bajo estimación FML como REML. Los datos denotan el porcentaje promedio de elecciones correctas a través de las variables tamaño de muestra, igualdad de las matrices de dispersión y forma de la distribución. Globalmente, los resultados indican lo siguiente:

1. La ejecución de los criterios examinados dependía del método de estimación, valor de los parámetros y número de medidas repetidas. Aunque no aparece recogido en la Tabla 6, los detalles pueden consultarse en la Web, la variable tamaño de muestra afectaba sustancialmente a la selección del verdadero PGD y las variables forma de la distribución e igualdad de las matrices de dispersión moderadamente.

2. El desempeño de los IC fue mejor bajo estimación REML que bajo estimación FML. Promediando a través de las 1052 (64 × 18) condiciones manipuladas, el porcentaje de aciertos obtenidos vía REML fue del 69.7% (del cual el 65.2% corresponde a los consistentes y el 74.2% a los eficientes), mientras que el obtenido vía FML promediando a través de 576 (64 × 9) condiciones fue del 55.9% (del cual el 47.5% corresponde a los consistentes y el 64.3% a los eficientes). Por su parte, el LRT eligió el verdadero PGD en el 73.1% de las veces bajo estimación FML y en el 68.9% bajo estimación REML.

3. Cuando el método de estimación usado era FML y $t = 6$, las diferencias existentes entre los IC y el LRT no excedían los 2 puntos porcentuales. Sin embargo, bajo estimación REML las diferencias favorecían a los IC y eran superiores a los 20 puntos. Sorprendentemente, la situación se invertía cuando $t = 12$. En este caso las diferencias excedían los 10 puntos porcentuales.

4. Por último, hay que destacar que el desempeño de los IC mejoraba si el estimador REML incluía el término $(\log|\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i|)/2$, pero sólo cuando $t = 6$. También cabe resaltar que el desempeño de los criterios consistentes era superior cuando se usaban las fórmulas implementadas en el módulo *Proc Mixed* del SAS, en lugar de las implementadas en el comando *Mixed* del SPSS. Observando la Tabla 1 se puede comprobar que los algoritmos usados por SAS y SPSS para calcular los criterios eficientes son idénticos.

Resultados globales

El desempeño global fue del 56.5% usando FML y del 62.9% usando REML. Los IC seleccionaron el verdadero PGD el 48.1% de las veces vía FML y el 68.7% vía REML, mientras que el LRT lo hizo el 64.8% y 57.7% de las veces, respectivamente. Bajo el método de estimación FML, los criterios eficientes seleccionaron el verdadero PGD el 51.3% de las veces - 58.1% el AIC y 44.5% el AICC - y los criterios consistente el 47.1% - 46.4% el BIC, 42.9% el CAIC y 52.1% el HQIC. A su vez, bajo el método REML, los criterios eficientes seleccionaron el verdadero PGD el 69.6% de las veces -74.7% el AIC y 64.5% el AICC - y los criterios consistente el 68.1% de las veces - 67.2% el BIC, 63.9% el CAIC y 73.2% el HQIC.

Conclusiones, recomendaciones y limitaciones

Son muchas las conclusiones se pueden extraer del estudio actual, no obstante, conviene destacar las cinco siguientes:

• En primer lugar, los datos pusieron de relieve que ninguno de los procedimientos examinados elegía consistentemente el verdadero PGD; sin embargo, su ejecución mejoraba al aumentar el tamaño de muestra, el número de medidas repetidas y la magnitud de los parámetros.

• En segundo lugar, independientemente de la parte del modelo que se ajustase, los IC basados en el estimador REML seleccionaban el verdadero PGD un mayor número de veces que los IC basados en el estimador FML. Este resultado no sólo confirma y extiende los hallazgos de Gurka (2006), sino que cuestiona la recomendación recogida en la literatura estadística especializada de usar en exclusiva los IC con REML para comparar modelos con idéntica estructura de medias (Orelien y Edwards, 2007).

• En tercer lugar, el desempeño de los IC mejoraba si el estimador REML incluía el término constante, especialmente cuando el número de medidas repetidas era moderado. A diferencia de lo que sucedía en el trabajo de Gurka (2006), donde el estimador REML1 sólo mejoraba el desempeño de los criterios consistentes, en el trabajo actual también mejoraba la ejecución de los criterios eficientes. Este resultado coincide con el encontrado por Wang y Schaalje (2009) al comparar el desempeño de los criterios AIC y BIC con el de los criterios predictivos R^2_{adj} CCC y PRESS.

• En cuarto lugar, el desempeño de los IC era superior cuando dichos criterios se calculaban usando el número total de sujetos (nivel 2), en vez del número total de observaciones (nivel 1). Este hallazgo, además de corroborar los resultados de Gurka (2006), también sirve de soporte empírico a la estrategia seguida en el módulo *Proc*

Mixed del SAS (2008), como opuesta a la seguida en el comando *Mixed* del SPSS (2008), de calcular los criterios consistentes usando el número de participantes en el nivel 2 del modelo jerárquico.

• En quinto lugar, a pesar de que los criterios AIC (78%), AICC (76.5%) y HQIC(76.8%) basados en el estimador REML1 elegían el verdadero PGD el mayor número de veces, cuando se requería ajustar la estructura de covarianza y el modelo completo el desempeño del LRT era tan bueno o mejor que el de los criterios reseñados.

Para concluir queremos efectuar una recomendación, una advertencia y una sugerencia. Globalmente, los criterios eficientes trabajaban mejor que los criterios consistentes cuando la estructura de covarianza era compleja y, viceversa, cuando era sencilla. Los criterios consistentes tendían a seleccionar modelos más parcos, generalmente de carácter estacionario, que los criterios eficientes. Ahora bien, en los estudios longitudinales de carácter aplicado suele ser habitual que la varianza de las observaciones sea heterogénea y que la correlación entre las mismas decrezca a lo largo del tiempo, de ahí que nos decantemos por el empleo de los criterios eficientes, en particular del AIC basado en el estimador $REML_1$. A nuestro juicio es el que cumple mejor el objetivo de encontrar un equilibrio entre un modelo complejo y otro parco. Hecha esta recomendación, debemos advertir que los resultados son limitados a las condiciones examinadas, si bien conjeturamos que pueden ser generalizadas a un rango más amplio de condiciones; por ejemplo, a situaciones donde los modelos no se hallen anidados unos dentro de otros. Finalmente, en la investigación realizada el verdadero PGD siempre pertenecía a la familia de modelos investigados. Sin embargo, cuando se trabaja con datos reales desconocemos si el verdadero PGD pertenece a la clase de modelos considerados. Por este motivo, sería deseable realizar una investigación donde el objetivo fuese comparar los IC en términos de seleccionar el modelo más próximo al verdadero PGD, dado que éste no se haya incluido en el conjunto de modelos presentes en la comparación.

Agradecimientos

Este trabajo ha sido financiado mediante sendos Proyectos de Investigación concedidos por el Ministerio de Ciencia e Innovación. (Ref.: PSI-2008-03624/PSI2009-11136/PSIC).

Referencias

Ato, M., & Vallejo, G. (2007). *Diseños experimentales en Psicología*. Madrid: Pirámide.

Claeskens, G., y Hjort, N.L. (2008). *Model Selection and Model Averaging*. New York: Cambridge University Press.

Dayton, C.M. (2003). Model comparisons using information measures. *Journal of Modern Applied Statistical Methods, 2*, 281-292.

Feng, R., Zhou, G., Zhang, M., y Zhang, H. (2009). Analysis of Twin Data Using SAS. *Biometrics*, Epub 2008 July 21 [PMID: 18647295].

Ferron, J., Dailey, R., y Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*, 379-403.

Fitzmaurice, G.M., Laird, N.M., y Ware, J.H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley.

Gómez, V.E., Schaalje, G.B., y Fellingham, G.W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation, 34*, 377-392.

Gurka, M.J., y Edwards, L.J. (2008). Mixed models. En C.R. Rao, J.P. Miller y D.C. Rao (Eds.): *Handbook of Statistics, vol. 27, Epidemiological and Medical Statistics* (pp. 253-280). New York: Elsevier Science.

Gurka, M.J. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60*, 19-26.

Kowalchuk, R.K., y Headrick, T.C. (2009). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*. DOI:10.1348/000711009X42 3067.

Hedeker, D., y Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: John Wiley.

Keselman, H.J., Algina, J., Kowalchuk, R.K., y Wolfinger, R.D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics- Simulation and Computation, 27*, 591-604.

Kreft, I.G., y de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.

Lee, H., y Ghosh, S.K. (2009). Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation, 79*, 93-106.

Liang, H., Wu, H., y Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika, 95*, 773- 778.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., y Schabenberger, O. (2006). SAS *System for Mixed Models*. 2nd edition. Cary, NC: SAS Institute Inc.

Littell, R.C., Pendergast, J., y Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793-1819.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin, 92*, 778-785.

Molenberghs, G., y Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling, 1*, 235-269.

Orelien, J.G. y Edwards, L.J. (2007). Fixed-effect variable selection in linear mixed models using R^2 statistics. *Computational Statistics & Data Analysis, 52*, 1896-1907.

SAS Institute Inc. (2008). *SAS/STAT® Software: Version 9.2*. SAS Institute Inc., Cary, NC.

Schabenberger, O. (2004). Mixed model influence diagnostics. *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., Paper 189-29.

Singer, D.J., y Willet, J.B. (2003). *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

SPSS for Windows (2008). Version 17, SPSS Inc., IL: Chicago.

Tukey, J.W. (1977). Modern techniques in data analysis. NSF-sponsored regional research conference at Southern Massachusetts University (North Dartmouth, MA).

Vaida, F., y Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92*, 351-370.

Vallejo, G., Arnau, J., y Bono, R. (2008a). Construcción de modelos jerárquicos en contextos aplicados. *Psicothema, 20*, 830-838.

Vallejo, G., Ato, M., y Valdés, T. (2008b). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology, 4*, 10-21.

Vallejo, G., Fernández, P., Herrero, J., y Conejo, N. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema, 16*, 498-508.

Vonesh, E.F., Chinchilli, V.M., y Pu, K. (1996). Goodness-of-Fit in generalized nonlinear mixed-effects models. *Biometrics, 52*, 575-587.

Wang, J., y Schaalje, G.B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics - Simulation and Computation, 38*, 788-801.

Zimmerman, D.L., y Núñez-Antón, V. (2009). *Antedependence Models for Longitudinal Data*. London: Chapman & Hall/CRC.