

# Análisis de datos longitudinales y de curvas de crecimiento. Enfoque clásico y propuestas actuales

Jaume Arnau y Nekane Balluerka  
Universidadde Barcelona y \* Universidad del País Vasco

En el presente trabajo se examinan los principales modelos de análisis aplicados a datos longitudinales y de curvas de crecimiento. Se parte de un enfoque clásico que abarca los modelos de análisis univariado y multivariado de la variancia de medidas repetidas, para proseguir con el *MANOVA* generalizado (*GMANOVA*) y con el modelo multinivel de dos estadíos, que constituye la base para el desarrollo del modelo general lineal mixto. Por último, se aborda el modelo multinivel para datos longitudinales y de curvas de crecimiento. Se describen las características y los supuestos esenciales de cada modelo, tratando de poner de manifiesto las razones por las que han ido emergiendo los sucesivos modelos y de establecer las principales similitudes y diferencias existentes entre ellos. Se concluye que los modelos lineales jerárquicos para datos longitudinales constituyen una excelente alternativa para realizar investigaciones aplicadas en el ámbito de las ciencias sociales y del comportamiento.

*Longitudinal and growth trajectory data analysis. Traditional approach and current proposals.* In this article, the main models for analysing longitudinal and growth trajectory data are examined. We start from a traditional approach that include univariate and multivariate analysis of variance for repeated measures, to continue with generalized MANOVA and with two stage-multilevel model, which constitutes the basis for the development of the general linear mixed model. Lastly, multilevel models for longitudinal and growth trajectory data are studied. The main characteristics as well as assumptions of each model are described, trying to make it clear the reasons why the successive models have appeared and to establish the main similarities and differences between them. It is concluded that hierarchical linear models for longitudinal data constitute an excellent strategy to do applied research in the realm of social and behavioural sciences.

En oposición a los datos transeccionales, los datos longitudinales son observaciones registradas en los mismos individuos a través del tiempo. Así, los diseños longitudinales de medidas repetidas constan de uno o más grupos de sujetos medido(s) en una o más variables a lo largo de dos o más puntos temporales. A diferencia de lo que ocurre en el contexto experimental, donde las medidas repetidas corresponden a los tratamientos, en las situaciones no-experimentales tales medidas representan las ocasiones temporales en las que se realizan las mediciones. Las curvas de crecimiento constituyen un ejemplo típico de este tipo de datos. En ellas, el investigador registra una serie de medidas en sucesivos intervalos temporales de amplitud constante, en una o más muestras de sujetos, con el fin de examinar el proceso de desarrollo de cada individuo y las posibles diferencias existentes en dicho proceso entre distintas muestras de individuos.

Desde una perspectiva clásica, los datos longitudinales han sido analizados mediante el modelo del análisis de la variancia univariado (*ANOVA*) o multivariado (*MANOVA*) de medidas repetidas.

Sin embargo, recientemente, han emergido una serie de modelos de análisis que superan, en múltiples aspectos, a los modelos clásicos. Todos ellos se subsumen bajo un modelo más amplio, conocido como modelo general lineal mixto. En el presente trabajo, exponemos los principales modelos de análisis que se han utilizado tradicionalmente para examinar los datos de diseños longitudinales y abordamos los modelos que, a nuestro juicio, están adquiriendo gran relevancia y se perfilan como pilares básicos de la investigación longitudinal en el futuro. Así, comenzamos describiendo las características y los supuestos esenciales de los modelos *ANOVA* y *MANOVA* clásicos, para proseguir con el modelo *MANOVA* generalizado, y finalizamos analizando distintos modelos multinivel o jerárquicos y su relación con el modelo general lineal mixto. De esta forma, pretendemos poner de manifiesto la evolución que se ha producido en las estrategias de análisis asociadas a los datos longitudinales y las importantes posibilidades que brindan los nuevos modelos de análisis para realizar investigaciones de carácter aplicado en el ámbito de las ciencias sociales y del comportamiento.

Modelos de análisis clásicos en el ámbito del diseño longitudinal

## *Análisis univariado de la variancia (ANOVA)*

En el análisis univariado de la variancia, las medidas repetidas se analizan siguiendo el enfoque del diseño multimuestra de me-

didadas repetidas, conocido también como modelo mixto (Lindquist, 1953). En dicho modelo, los sujetos son considerados como un factor aleatorio y las ocasiones y los grupos como factores fijos. Básicamente, esta estrategia estima las medias de las observaciones de los sujetos y las compara a través de los distintos grupos, o a lo sumo, examina si se ajustan a polinomios conocidos. De esta forma, constituye un procedimiento adecuado cuando el interés del investigador radica en analizar el efecto global que ejerce el tiempo o determinadas variables de crecimiento, tales como la edad, sobre el desarrollo y su posible interacción con factores entre-grupos (o entre-sujetos).

La aplicación del ANOVA a esta clase de datos requiere el cumplimiento de una serie de supuestos (Boik, 1981; Huynh, 1978). En primer lugar, las observaciones de los diferentes sujetos han de ser independientes entre sí. En segundo lugar, en cada grupo, las puntuaciones de las diferentes ocasiones o variables han de seguir una distribución normal multivariante. En tercer lugar, las matrices de variancia-covariancia de las medidas de cada grupo han de ser iguales entre sí (supuesto de homogeneidad). Por último, la matriz de variancia-covariancia intragrupo ha de satisfacer el supuesto de esfericidad, es decir, las variancias de las diferencias entre los distintos pares de medidas repetidas deben ser homogéneas (Huynh y Feldt, 1970; Rouanet y LePine, 1970). La ausencia de homogeneidad de las matrices de variancia-covariancia intragrupo puede incrementar la distancia entre el nivel de significación real y el nominal, afectando a la potencia de la prueba estadística. De la misma forma, el incumplimiento del supuesto de esfericidad lleva a un sesgo positivo en el estadístico F (Box, 1954). En tal caso, para que la prueba estadística sea válida, debe utilizarse algún procedimiento que permita ajustar los grados de libertad de la F. Así, cabe realizar el ajuste a partir del límite inferior de  $e$  o mediante el factor de corrección de Greenhouse y Geisser (1959) o la versión adaptada de Huynh y Feldt (1976). A su vez, el procedimiento de aproximación general mejorada (IGA) de Huynh (1978), la prueba de Welch-James propuesta por Johansen (1980), la aproximación multivariante o el enfoque del modelo mixto también permiten corregir el sesgo derivado de la violación del supuesto de esfericidad (las principales alternativas de análisis ante el incumplimiento de este supuesto pueden consultarse en Keselman, Algina y Kowalchuk, 2001, 2002, y en Keselman, Algina, Kowalchuk y Wolfinger, 1999).

*Análisis multivariante de la variancia (MANOVA)*

El análisis multivariante de la variancia, aplicado a datos longitudinales, asume que las medidas son múltiples variables dependientes que están correlacionadas en los mismos sujetos. Como señaló Finn (1969), cuando se dispone de medidas repetidas, el MANOVA constituye una buena alternativa al análisis univariado. Normalmente, los modelos multivariados se centran en el componente entre-grupos (o entre-sujetos) del análisis, tratando de explicar la variancia total de las variables dependientes en función de las diferencias existentes entre los miembros de los distintos grupos. Las pruebas de significación se llevan a cabo transformando las variables originales en contrastes de interés. Mediante esta transformación de las medidas repetidas es posible verificar si se produce algún tipo de cambio lineal, cuadrático, cúbico o de otro orden en función del tiempo pero, como destacan Wu, Clopper y Wooldridge (1999), el factor intrasujeto, en sí mismo, es eliminado del análisis.

Los modelos MANOVA enfatizan la parte fija del modelo. Se estima un modelo que explica la estructura de las medias, considerando las covariancias intra-grupo/sujeto como errores aleatorios, es decir, como la parte de la variancia de las variables dependientes que no puede ser explicada por la pertenencia a un determinado grupo. En su aplicación a las medidas repetidas, la parte fija del modelo es expandida. A fin de ajustar curvas de crecimiento polinómicas de un determinado grado, el conjunto de variables explicativas, entre las que la primera solo representa la pertenencia a un grupo concreto, se amplía con variables intra-sujeto que corresponden a los diferentes puntos temporales, tales como la edad.

El MANOVA se define, en notación matricial, por la expresión

$$Y = XB + E \quad (1)$$

En la ecuación (1),  $Y$  es la matriz ( $n \times p$ ) de observaciones ( $n$  observaciones y  $p$  medidas repetidas por sujeto a intervalos temporales fijos),  $X$  es la matriz ( $n \times k$ ) del diseño, la cual toma los valores 1 y 0 para representar la pertenencia o no pertenencia del sujeto a un determinado grupo,  $B$  es la matriz ( $k \times p$ ) de parámetros y  $E$  es la matriz ( $n \times p$ ) que incluye las fuentes de variación aleatorias. El modelo de la ecuación (1) asume que los errores de la  $i$ -ésima fila de  $E$  siguen una distribución normal multivariada,  $N_p(0, \Sigma_p)$ . Para su correcta aplicación, el MANOVA requiere que las respuestas de los sujetos sean independientes entre sí, que la distribución de las múltiples variables dependientes sea normal multivariada, y que el conjunto de datos sea completo, a saber, sin valores ausentes (Stevens, 1966).

El modelo de respuesta media o valor esperado de los valores observados viene dado por

$$E(Y) = XB \quad (2)$$

A fin de comprobar si las medias de  $p$  ocasiones son iguales entre sí, el enfoque multivariado utiliza una prueba derivada del estadístico  $T^2$  de Hotelling (1951). Mediante el estadístico  $T^2$  se pone a prueba la siguiente hipótesis de nulidad:

$$H_0: \mu_1 = \dots = \mu_p \quad (3)$$

Que es equivalente a:

$$H_0: \mu_1 - \mu_2 = \dots = \mu_{p-1} - \mu_p \quad (4)$$

El cálculo del estadístico multivariante,  $T^2$ , se obtiene del vector de medias de diferencias,  $Y_d$ , y de la matriz de variancias y covariancias de las diferencias entre las puntuaciones,  $S_d$ . De forma que,

$$T^2 = n Y_d' S_d^{-1} Y_d \quad (5)$$

El valor de  $T^2$  puede transformarse en la distribución conocida F, con  $(n-1)$  y  $(n-1)(p-1)$  grados de libertad.

Es importante señalar que el MANOVA clásico define la estructura entre-sujetos de los valores esperados de las observaciones sin tomar en consideración las relaciones entre las variables o medidas repetidas. Sin embargo, desde la perspectiva longitudinal, el principal interés radica en modelar los perfiles de las respuestas medias, es decir, en obtener información sobre el efecto del tiempo. Mediante el modelo de análisis de la variancia multivariante generalizado es posible obtener ese tipo de información.

## Análisis multivariante de la variancia generalizado (GMANOVA)

Para modelar los perfiles de las respuestas medias, en el modelo *MANOVA*, es necesario aplicar restricciones a las observaciones en función de los intervalos temporales. Una forma simple de conseguir este objetivo consiste en asumir que la matriz de parámetros (matriz **B**) de la ecuación que representa el modelo de respuesta media (ecuación 2), se deriva del modelo

$$\mathbf{B} = \mathbf{HT} \quad (6)$$

De esta forma, el nuevo modelo multivariante de la variancia se expresa, en notación matricial, por

$$\mathbf{Y} = \mathbf{XHT} + \mathbf{E} \quad (7)$$

donde **X** representa, como en el *MANOVA* clásico, la matriz ( $n \times k$ ) del diseño, **H** es la matriz ( $k \times q$ ) de parámetros y **T** corresponde a la matriz ( $q \times p$ ) que describe el perfil de los valores esperados de cada sujeto, es decir, el patrón de cambio de las observaciones a lo largo del tiempo. Esta matriz se denomina matriz intra-individuos o intra-grupos. La ecuación (7) representa el modelo *GMANOVA* o modelo de curva de crecimiento. Este modelo, propuesto por Pott-hoff y Roy (1964), surgió del intento de integrar los métodos alternativos al *ANOVA* mixto, es decir, los modelos *MANOVA* clásicos, dentro de un modelo más amplio: el modelo lineal general multivariado, el cual constituye una buena alternativa al modelo lineal clásico. A diferencia del *MANOVA* clásico, el *GMANOVA* incluye el ajuste polinómico de la curva como función del tiempo, permitiendo describir los perfiles con coeficientes aleatorios y generar la estructura de correlación entre las observaciones repetidas.

El valor esperado de los valores observados (matriz de observaciones) viene dado por

$$E(\mathbf{Y}) = \mathbf{XHT} \quad (8)$$

donde los elementos de **H** son los coeficientes polinómicos que representan el efecto del tiempo. La matriz **T**, con estructura Vandermonde, define el modelo de cambio que se postula en el diseño longitudinal. Con ello, se constata que el *MANOVA* es una generalización de los principios del *ANOVA* (Bartlett, 1947; Tukey, 1949). El modelo de curva de crecimiento resulta muy útil en el ámbito longitudinal, dado que trata de explicar la variación intraindividual en función del desarrollo natural o proceso de maduración. Además, permite trabajar con datos correlacionados y, en algunos casos, brinda la posibilidad de modelar estructuras de datos de ocasiones de medida fijas no balanceadas (Jennrich y Schluchter, 1986).

## Los modelos multinivel: una apuesta de futuro

Los modelos multinivel (Goldstein, 1995; Hox, 1996, 2002; Plewis y Hurry, 1998; Snijders y Bosker, 1999), conocidos también, en la literatura estadística, como modelos lineales jerárquicos (Bryk y Raudenbush, 1992; Raudenbush y Bryk, 2002), modelos de efectos o coeficientes aleatorios (De Leeuw y Kreft, 1986; Longford, 1993) o modelos de componentes de la variancia (Longford, 1989), son una extensión de los modelos de efectos mixtos descritos por Rao (1965) para las curvas de crecimiento y por Laird y Ware (1982) para el análisis de datos longitudinales.

Timm y Mieczkowski (1997) señalan que el crecimiento puede representarse mediante un modelo lineal jerárquico de dos estadíos, donde los parámetros poblacionales, los efectos individuales y la variación intra-sujeto se definen en el primer estadío y la variación entre-sujetos es modelada en el segundo estadío. En el modelo multinivel de dos estadíos para datos de medidas repetidas, se considera que tales datos configuran una estructura jerárquica a dos niveles: las observaciones son las unidades del primer nivel y los sujetos son las unidades del segundo nivel (Cnaan, Laird y Slasor, 1997; Van der Leeden, Vrijburg y De Leeuw, 1996). Se parte del supuesto de que la distribución de probabilidad de las medidas repetidas tiene la misma forma para cada sujeto, de modo que los parámetros de esta distribución varían a través de los sujetos. En el primer estadío se define una regresión lineal para las observaciones registradas en cada individuo. En el segundo estadío, los coeficientes de regresión o los parámetros de las curvas de crecimiento individuales modelados en el primer estadío son considerados como variables dependientes aleatorias. El propósito del segundo estadío radica en analizar la distribución de estos parámetros o efectos aleatorios en la población. A diferencia de los modelos *ANOVA* y *MANOVA* clásicos, el modelo multinivel para medidas repetidas no enfatiza la variación entre-sujetos, sino que persigue, como principal objetivo, modelar las curvas de crecimiento individuales y analizar, posteriormente, las posibles diferencias entre los individuos en los parámetros que describen los patrones de crecimiento. De esta forma, permite definir y examinar la variación intra y entre-sujetos. De hecho, aunque el predictor de regresión lineal para la respuesta media no presente ninguna peculiaridad, existen dos tipos de términos de error: los errores intra-sujeto y los errores entre-sujetos. Todas las observaciones del mismo sujeto tienen el mismo error entre-sujetos, pero los errores intra-sujeto difieren y pueden estar correlacionados. En relación con este hecho, otra diferencia relevante que presenta el modelo de dos estadíos respecto al modelo multivariado general es que requiere definir un patrón específico para la matriz de covariancia de las observaciones.

Es importante señalar que el modelo general lineal mixto constituye la integración de las dos ecuaciones correspondientes a las dos etapas del modelo de dos estadíos. Además, como veremos posteriormente, una extensión directa del modelo jerárquico de dos estadíos es su aplicación a datos longitudinales y de curvas de crecimiento. No obstante, antes de abordar el modelo general lineal mixto y el modelo multinivel para datos longitudinales, describiremos el modelo general de dos estadíos partiendo de una situación de naturaleza jerárquica en la que no se toman medidas repetidas.

## Modelo multinivel de dos estadíos

Este modelo puede conceptualizarse como un sistema jerárquico de ecuaciones de regresión. Cuando los datos son de naturaleza jerárquica, el modelo de regresión clásico induce a estimaciones sesgadas de los errores estándar e incrementa la probabilidad de cometer un error de Tipo I (Goldstein, 1995). A diferencia de la regresión clásica, la regresión multinivel asume, de forma más realista, que los coeficientes de regresión son estocásticos (es decir, aleatorios). Como señalan Cnaan, Laird y Slasor (1997), es un modelo muy flexible que permite examinar la variación que se produce a nivel individual (micro) y a nivel grupal (macro), y en el que las diferencias inter-grupos no se conciben como una mera diferencia entre las medias grupales, sino que se modelan como distribuciones distintas dentro de cada grupo.

A fin de ilustrar el modelo, describimos en primer lugar su estructura fuera del contexto longitudinal y dejamos, para más adelante, su estudio dentro del marco de datos de medidas repetidas. Supongamos que tenemos  $N$  individuos (nivel-1) anidados en  $J$  grupos (nivel-2). En un primer estadio (nivel-1 o modelo intra-grupo) se define una ecuación de regresión para cada grupo, siendo las unidades de este nivel los individuos. Si el modelo consta de una variable dependiente continua  $y$ , y de una única variable predictora o covariable del nivel-1,  $X$ , también de naturaleza continua, la ecuación del nivel-1 se expresa por

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \quad e_{ij} \equiv N(0, \sigma_e^2) \quad (9)$$

donde  $y_{ij}$  es la variable de respuesta del  $i$ -ésimo sujeto (nivel-1) dentro del  $j$ -ésimo grupo (nivel-2),  $\beta_{0j}$  es el intercepto del  $j$ -ésimo grupo del nivel-2,  $\beta_{1j}$  es el coeficiente de regresión asociado a la variable predictora  $X_{ij}$  del modelo a nivel-1 (por ejemplo, la edad de los sujetos), y  $e_{ij}$  es el error aleatorio asociado al  $i$ -ésimo sujeto. Se asume que los errores al nivel individual, en cada grupo, son independientes y tienen una distribución normal con media cero y variancia constante.

En notación matricial, el modelo del primer estadio adopta la siguiente expresión:

$$y_j = \mathbf{X}_j \beta_j + e_j \quad (10)$$

donde  $e_j$  es un vector de errores aleatorios y se distribuye como  $N(0, \mathbf{R})$ . Así, para cada grupo, existe un modelo a nivel-1 (intra-grupo) de la regresión con las mismas variables predictoras, pero con distintos coeficientes  $\beta_{0j}$  y  $\beta_{1j}$ .

En el segundo nivel (modelo entre-grupos), los coeficientes de regresión de los grupos actúan de variables dependientes que han de ser explicadas por variables predictoras del nivel-2. En el caso de tener una sola variable predictora  $Z$  (por ejemplo, un determinado tratamiento terapéutico), las ecuaciones del nivel-2 se expresan como sigue,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \quad u_{0j} \equiv N(0, \tau_{00}) \quad (11)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \quad u_{1j} \equiv N(0, \tau_{11}) \quad (12)$$

$$\text{Cov}(u_{0j}, u_{1j}) = \tau_{10}$$

donde  $\beta_{0j}$  y  $\beta_{1j}$  son los parámetros para el intercepto y la pendiente del  $j$ -ésimo grupo (nivel-2), que actúan de variables dependientes en las ecuaciones (11) y (12). A su vez,  $\gamma_{00}$  y  $\gamma_{10}$  son las correspondientes medias poblacionales de estos parámetros, y  $\gamma_{01}$  y  $\gamma_{11}$  los correspondientes coeficientes de regresión sobre la variable predictora  $Z$  asociados, también, a estos parámetros. Por último,  $u_{0j}$  y  $u_{1j}$  son las desviaciones aleatorias, tanto respecto del intercepto como de la pendiente de la parte fija, en una y otra ecuación del nivel-2. Se asume que ambos errores tienen una distribución normal con media cero y variancias  $\tau_{00}$  y  $\tau_{11}$ .

En notación matricial, el modelo del segundo estadio viene dado por la expresión:

$$\beta_j = \mathbf{Z}_j \gamma + \mathbf{u}_j \quad (13)$$

donde  $\beta_j$  es el vector de coeficientes de la regresión,  $\mathbf{Z}_j$  es la matriz del diseño con coeficientes fijos conocidos,  $\gamma$  es el vector de parámetros fijos y  $\mathbf{u}_j$  es el vector de componentes de error aleatorios, con media cero y variancia  $\mathbf{G}$ .

Los modelos del nivel-1 (intra-grupo) y del nivel-2 (entre-grupos), que acabamos de describir, adquieren mayor sentido cuando se integran en un solo modelo. De hecho, el análisis multinivel combina ambos modelos en un modelo más general conocido como modelo general lineal mixto (Harville, 1977). Cabe señalar que el trabajo de Kreft, De Leeuw y Van der Leeden (1994) describe exhaustivamente los principales paquetes estadísticos que permiten llevar a cabo análisis multinivel, a excepción del SAS, cuya exposición puede consultarse, entre otros, en el artículo de Singer (1998).

Modelo general lineal mixto

Sustituyendo las ecuaciones (11) y (12) en la ecuación (9) obtenemos el modelo siguiente:

$$y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \gamma_{10} X_{ij} + \gamma_{11} Z_j X_{ij} + u_{0j} + u_{1j} X_{ij} + e_{ij} \quad (14)$$

En el modelo de regresión multinivel representado en la ecuación (14), que constituye un caso particular del modelo general lineal mixto, cabe distinguir dos partes: una parte fija que no varía a través de los grupos y una parte aleatoria susceptible de variar de grupo a grupo. La parte fija está formada por los coeficientes de regresión  $\gamma$  y por sus variables asociadas:  $[\gamma_{00} + \gamma_{01} Z_j + \gamma_{10} X_{ij} + \gamma_{11} Z_j X_{ij}]$ . La parte aleatoria está compuesta por los términos de error del nivel 1 y del nivel 2:  $[u_{0j} + u_{1j} X_{ij} + e_{ij}]$ . Dentro de este contexto, los coeficientes  $\gamma$  se conocen por efectos fijos y los parámetros aleatorios incluyen los componentes de variancia/covariancia de los términos de error de ambos niveles. Así, se tiene, la variancia del intercepto  $\beta_{0j}$  ( $\tau_{00}$ ) y de la pendiente  $\beta_{1j}$  ( $\tau_{11}$ ) no explicadas por  $X_{ij}$  y  $Z_j$ , la covariancia entre el intercepto y la pendiente tras tener en cuenta los efectos de  $X_{ij}$  y  $Z_j$  ( $\tau_{10}$ ) y la variancia residual del nivel individual ( $\sigma_e^2$ ). Cabe señalar que el término de interacción  $\gamma_{11} Z_j X_{ij}$  se conoce como interacción transnivel, debido a que incluye variables predictoras de distintos niveles.

En notación matricial, la ecuación (14) se expresa mediante los siguientes términos:

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij} \mathbf{Z}_j \gamma + \mathbf{X}_{ij} \mathbf{u}_{ij} + e_{ij} \quad (15)$$

Si partimos del modelo completo expresado en notación matricial, ecuación (15), la parte fija,  $\mathbf{X}_{ij} \mathbf{Z}_j \gamma$ , define los valores esperados de las observaciones, mientras que la parte aleatoria,  $\mathbf{X}_{ij} \mathbf{u}_{ij} + e_{ij}$ , establece la estructura de covariancia de los datos, asumiéndose que los elementos de  $e_{ij}$  y de  $\mathbf{u}_{ij}$  tienen distribuciones independientes entre sí. De este modo, el modelo de respuesta media de los valores observados (matriz de observaciones) es:

$$E(\mathbf{Y}) = \mathbf{X} \mathbf{H} \quad (16)$$

y las covariancias entre los elementos de  $y_{ij}$  no explicadas por la parte fija del modelo (matriz de variancia de la parte aleatoria) se expresan por

$$\text{Var}(\mathbf{Y}) = \mathbf{Z} \mathbf{G} \mathbf{Z}' + \mathbf{R} = \mathbf{Z} \mathbf{G} \mathbf{Z}' + \mathbf{N}^2 \mathbf{I} \quad (17)$$

En su formulación final, el modelo general lineal mixto adopta la siguiente expresión:

$$\mathbf{Y}_j = \mathbf{X}_j \gamma + \mathbf{Z}_j \mathbf{u}_j + e_j \quad (18)$$

donde  $\mathbf{Y}_j$  es el vector de respuestas para la  $j$ -ésima unidad o grupo,  $\mathbf{X}_j$  y  $\mathbf{Z}_j$  son las matrices del diseño conocidas. La matriz  $\mathbf{Z}_j$  puede constar de variable continuas o de variables dummy, como las que incluye  $X_j$ ,  $\gamma$  es el vector de parámetros de efectos fijos,  $\mathbf{u}_j$  es el vector de parámetros de efectos aleatorios, con distribución  $N(0, \mathbf{G})$ , y  $\mathbf{e}_j$  es un vector de errores residuales con distribución  $N(0, \mathbf{R})$ . El modelo de la ecuación (18) se denomina, con frecuencia, modelo lineal jerárquico o modelo lineal mixto anidado (Goldstein, 1986; Longford, 1987). El nombre de modelo mixto se debe a que incluye parámetros de efectos fijos (los elementos de  $\gamma$ ) y parámetros de efectos aleatorios o componentes de variancia ( $\sigma_e^2$  al nivel 1 y los elementos de  $\mathbf{G}$  al nivel 2). Es importante señalar que la variancia de  $\mathbf{Y}$  (ecuación 17) se define especificando la matriz del diseño de efectos aleatorios,  $\mathbf{Z}$ , y las estructuras de covariancia para las matrices  $\mathbf{G}$  y  $\mathbf{R}$ .

Llegados a este punto, cabe destacar que los coeficientes  $\beta_j$  intragrupo, o coeficientes de regresión aleatorios, no forman parte de este modelo combinado. Así, en lugar de estimar una gran cantidad de parámetros  $\beta$ , se asume que éstos siguen una distribución normal en todos los grupos y se estiman los parámetros de dicha distribución:  $\gamma$  y  $\mathbf{G}$ . Ello convierte al modelo general lineal mixto en un modelo que aporta mayor cantidad de información que los modelos clásicos basados en el análisis de todo el conjunto de datos ignorando los grupos, y mucho más parsimonioso que los modelos que realizan análisis aislados para cada grupo, como pueden ser los modelos de coeficientes variables.

Para finalizar con la descripción del modelo general lineal mixto, y antes de abordar los modelos multinivel para datos longitudinales, citaremos las principales asunciones de este modelo cuando se dispone de datos continuos que siguen una distribución normal (Bryk y Raudenbush, 1992). Así, para el nivel-1, se asume que los términos aleatorios  $e_{ij}$  son independientes y siguen una distribución normal con media cero y variancia constante  $\sigma_e^2$ . Al nivel-2, se asume que los parámetros  $\beta_{0j}$  y  $\beta_{1j}$  tienen una distribución normal multivariada con medias  $\gamma_{00}$  y  $\gamma_{10}$  y variancias  $\tau_{00}$  y  $\tau_{11}$ , respectivamente. A su vez, se parte del supuesto de que los errores del nivel-1 y del nivel-2 no están correlacionados. Por último, se presupone que las variables predictoras de cada nivel son independientes de los errores de ese mismo nivel.

#### Modelo multinivel para datos longitudinales

Como se ha señalado anteriormente, los datos longitudinales o de curvas de crecimiento pueden representarse mediante una estructura jerárquica de dos niveles, en la que las observaciones directas o medidas repetidas (nivel-1) se hallan anidadas en los sujetos (nivel-2) (Bock, 1989; Bryk y Raudenbush, 1987, 1992; Goldstein, 1989, 1995; Raudenbush, 1989). Por otra parte, dado que las curvas de crecimiento representan un proceso de desarrollo que se produce en función del tiempo, una forma adecuada de modelarlas radica en describir los valores esperados de las observaciones como funciones polinómicas del tiempo (Van der Leeden, 1998a).

Como destacan Wu, Clopper y Wooldridge (1999), el modelo lineal jerárquico enfoca el análisis de los datos longitudinales de forma totalmente distinta a los métodos tradicionales. Así, en un primer nivel (modelo intra-sujeto) se modelan los patrones de crecimiento de cada individuo en función de un único conjunto de parámetros. Habitualmente, la variable de respuesta se modela en función de la edad y de un componente de error residual intra-su-

jeto. El modelado puede seguir una función polinómica de orden  $p$ , expresada mediante la siguiente ecuación:

$$y_{it} = \beta_{0i} + \beta_{1i} T_{it} + \beta_{2i} T_{it}^2 + \dots + \beta_{pi} T_{it}^p + e_{it} \quad (19)$$

donde  $y_{it}$  representa la medida de la variable dependiente para el sujeto  $i$  en la ocasión  $t$ . Los  $\beta_{pi}$ 's son los coeficientes de una función polinómica de grado  $p$  y los  $e_{it}$ 's son los errores aleatorios. Se asume que los residuales  $e_{it}$  tienen una distribución normal e independiente con media cero y variancia constante. Las  $T_{it}$ 's son los vectores polinómicos o variables explicativas, como por ejemplo la edad. Nótese que los parámetros  $\beta_{pi}$  y la variancia residual  $N_{e_i}^2$  son específicos de la persona  $i$ . Por lo tanto, con el modelo del nivel-1, se describen las medidas repetidas de un proceso de crecimiento o de desarrollo en términos, por ejemplo, de la variable edad como una función polinómica de un determinado grado.

A un segundo nivel (modelo entre-sujetos), se asume que los parámetros del proceso de crecimiento varían aleatoriamente entre los individuos. Estos parámetros son considerados como variables dependientes aleatorias que pueden variar dependiendo de características del nivel-2 o de los sujetos. En este nivel, los parámetros de crecimiento se modelan en función de los promedios poblacionales (parámetros poblacionales) y de las desviaciones que presentan los sujetos respecto de dichos promedios (residuales). La complejidad de este modelo aumenta cuando se incluyen covariables con el fin de explicar la variación observada entre los sujetos en los parámetros de crecimiento. El correspondiente modelo del nivel-2 para datos de crecimiento es dado por

$$\beta_{pi} = Z_i \gamma_p + u_{pi} \quad (20)$$

donde  $Z_i$  es una matriz que contiene las variables entre sujetos. Cuando sólo se modela la variación aleatoria de los parámetros individuales de crecimiento,  $Z_i$  toma la forma de una matriz identidad. Por último,  $\gamma_p$  representa el efecto de  $Z_p$  en el parámetro de crecimiento  $p$  y  $u_{pi}$  es un vector de errores aleatorios.

En notación matricial, el modelo completo a dos niveles para datos de crecimiento es:

$$\mathbf{y}_i = \mathbf{T}_i \beta_i + \mathbf{e}_i \quad (21)$$

$$\beta_i = \mathbf{Z}_i \gamma + \mathbf{u}_i \quad (22)$$

En la ecuación (21) o modelo intra-sujeto,  $\mathbf{y}_i$  es el vector de respuestas ( $t \times 1$ ) que contiene las medidas repetidas para el sujeto  $i$ ,  $\mathbf{T}_i$  es una matriz ( $t \times p$ ) de variables conocidas (por ejemplo edad en años, número de meses, etc.) y de las transformaciones (cuadrática, cúbica, etc.) de tales variables,  $\beta_i$  es un vector ( $p \times 1$ ) de parámetros individuales que especifican la forma de la curva de crecimiento para el sujeto  $i$ , y  $\mathbf{e}_i$  es un vector ( $t \times 1$ ) de errores aleatorios con distribución normal multivariada y una determinada matriz de covariancia. El modelo entre-sujetos, la ecuación (22), tiene también una formulación general (véase la ecuación 13), aunque su interpretación es diferente. De acuerdo con este modelo,  $\mathbf{Z}_i$  es una matriz del diseño entre-sujetos ( $p \times q$ ) con elementos conocidos y fijos,  $\gamma$  es un vector de coeficientes fijos ( $q \times 1$ ) y  $\mathbf{u}_i$  es un vector de errores aleatorios ( $p \times 1$ ). Sustituyendo la ecuación (22) en la (21), se deriva el modelo mixto completo para el diseño de medidas repetidas y viene dado por

$$\mathbf{Y}_i = \mathbf{TZ}_i \gamma + \mathbf{T}\mathbf{u}_i + \mathbf{e}_i \quad (23)$$

Se asume que los elementos de  $e_i$  tienen una distribución independiente y normal con variancia constante,  $e_i \equiv N(0, \sigma_e^2 \mathbf{I})$ , que los términos aleatorios del nivel-2, en  $u_i$ , siguen una distribución normal,  $u_i \equiv N(0, \mathbf{G})$  y que los términos de error del nivel-1 ( $e_i$ ) se distribuyen independientemente de los del nivel-2 ( $u_i$ ). En el modelo completo, el término  $[\mathbf{TZ}_i\gamma]$  constituye la parte fija, mientras que el término  $[\mathbf{T}u_i + e_i]$  constituye la parte aleatoria. Los efectos fijos definen los valores esperados de las observaciones y los efectos aleatorios son variancias y covariancias. La covariancia entre los elementos de  $\mathbf{Y}_i$  o los datos longitudinales consta de la parte entre-sujetos e intra-sujeto, de forma que

$$\text{Var}(\mathbf{Y}_i) = \text{Var}(\mathbf{T}u_i + e_i) = \mathbf{TGT}' + \sigma_e^2 \mathbf{I} \quad (24)$$

Cabe señalar que los supuestos relativos a los errores, especialmente al término  $e_i$ , llevan a estructuras de covariancia muy simples al nivel individual del modelo (errores constantes y no correlacionados a través de los puntos temporales). No obstante, cuando se dispone de muchos puntos temporales por sujeto, los residuales presentan, frecuentemente, algún patrón de autocorrelación serial. Los modelos multinivel brindan la posibilidad de seleccionar una matriz de covariancia que represente adecuadamente la dependencia entre las observaciones. Así, el modelado de la estructura de covariancia intra-sujeto adquiere una relevancia esencial, dado que la precisión en la estimación de los parámetros de regresión depende, en gran medida, de la adecuada elección de dicha estructura (Littell, Pendergast y Natarajan, 2000; Park y Lee, 2002).

#### Discusión

En el presente artículo se han examinado los principales modelos de análisis aplicados a datos longitudinales y de curvas de crecimiento, partiendo desde un enfoque clásico hasta llegar a los modelos lineales jerárquicos longitudinales. Se han abordado las características y los supuestos esenciales de cada modelo, tratando de poner de manifiesto las razones por las que han ido emergiendo los sucesivos modelos y de establecer las principales similitudes y diferencias existentes entre ellos. Dado que se han abordado los modelos con una variable dependiente continua, remitimos al lector interesado en la versión logística del modelo multinivel para variables dependientes categóricas a la excelente obra de Diggle, Liang y Zeger (1998).

A modo de conclusión, proporcionamos una serie de pautas que pueden resultar útiles para decantarse por uno u otro modelo, partiendo del análisis de las principales ventajas y limitaciones que presentan los modelos de más reciente aparición en psicología, es decir, los modelos multinivel, respecto a los modelos *MANOVA* clásicos de medidas repetidas. Para ello, tomamos como referencia básica los trabajos de Bryk y Raudenbush (1992) y de Van der Leeden (1998b).

La primera ventaja que presentan los modelos multinivel respecto a los *MANOVA* es que toman en consideración dos importantes características del crecimiento: a) que las respuestas de un mismo sujeto en diferentes ocasiones están correlacionadas entre sí, y b) que las covariancias entre las observaciones, es decir, las variaciones entre-sujetos, son función del tiempo o de la edad. Las covariancias entre las observaciones dependen tanto del modelo intra-sujeto como de las covariancias entre los parámetros del mo-

delo entre-sujetos. A diferencia del *MANOVA* de medidas repetidas, el modelo lineal jerárquico permite especificar distintas estructuras de covariancia para modelar los términos de error  $e_{ij}$  en el modelo intra-sujeto y proporciona métodos para examinar los posibles determinantes de dichas estructuras.

En segundo lugar, el modelo multinivel longitudinal tiene en cuenta explícitamente la trayectoria de crecimiento individual en el modelo intra-sujeto. Por el contrario, en los *MANOVA* de medidas repetidas, la variación individual no se modela directamente, sino que se refleja únicamente en la interacción entre las medidas repetidas y los sujetos.

En tercer lugar, el modelo jerárquico es más flexible que el *MANOVA* de medidas repetidas respecto a los requisitos que deben cumplir los datos (ser balanceados) y a los supuestos del modelo (por ejemplo, en cuanto a la matriz de covariancia intra-sujeto), debido a que se considera que las medidas repetidas están anidadas dentro del sujeto y no constituyen, como en el *MANOVA*, un conjunto fijo para todos los sujetos. Así, en el modelo multinivel, es posible formular diversas curvas de crecimiento así como incorporar covariables que varían a través del tiempo en el modelo intra-sujeto. Los efectos de tales covariables pueden ser fijos, de variación no-aleatoria o aleatorios. De la misma forma, la variable tiempo o edad no tiene por qué incluir una serie de puntos fijos, sino que puede ser continua. Por otra parte, los modelos multinivel tienen capacidad para modelar estructuras de datos de ocasiones de medida fijas no balanceadas, a saber, estructuras en las que tanto la cantidad de observaciones por sujeto como la amplitud de los intervalos temporales pueden ser variables y en las que algunos (o todos) los sujetos presentan valores perdidos. Hemos de señalar que este tipo de estructuras también pueden modelarse mediante algunos modelos *MANOVA* de medidas repetidas generalizados tales, como por ejemplo, el propuesto por Jennrich y Schluchter (1986).

Por último, cabe destacar que incluso cuando todos los requisitos de los datos y todos los supuestos necesarios para aplicar el *MANOVA* de medidas repetidas se cumplen, el análisis jerárquico proporciona los mismos estimadores puntuales, que el *MANOVA*, para los efectos fijos.

Sin embargo, los modelos multinivel también adolecen de ciertas deficiencias. Así, cuando se dispone de muestras pequeñas, los métodos máximo verosímiles, utilizados habitualmente para la estimación en este tipo de modelos, proporcionan estimaciones negativamente sesgadas de los componentes de variancia (Searle, Casella y McCulloch, 1992), y los errores estándar estimados pueden ser diferentes de los reales. En el mismo sentido, los resultados de las pruebas de hipótesis para los componentes de variancia y para los efectos aleatorios individuales deben interpretarse con cautela cuando el número de unidades del nivel-2 es pequeño ( $J < 30$ ) o los datos son excesivamente no balanceados (Sullivan, Dukes y Losina, 1999). Por otra parte, a pesar de que también se pueden plantear modelos multinivel longitudinales multivariados (Goldstein, 1995), el modelado no resulta tan sencillo como el de los modelos univariados abordados en el presente artículo.

De cualquier forma, los modelos lineales jerárquicos han marcado un hito en el modelado de datos longitudinales y se perfilan como una alternativa que brinda excelentes posibilidades para realizar investigaciones no-experimentales en el ámbito de la psicología, donde su desarrollo dista mucho del alcanzado en otras ciencias de la salud como, por ejemplo, la epidemiología.

## Referencias

- Bartlett, M.S. (1947). The general canonical correlation distribution. *Annals of Mathematical Statistics*, 18, 1-17.
- Bock, R.D. (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.
- Boik, R.J. (1981). A priori tests in repeated measures designs: effects of nonsphericity. *Psychometrika*, 46, 241-255.
- Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 25, 484-498.
- Bryk, A.S. y Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101, 147-158.
- Bryk, A.S. y Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- Cnaan, A., Laird, N.M. y Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16, 2.349-2.380.
- De Leeuw, J. y Krefit, I.G.G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, 11, 57-85.
- Diggle, P.J., Liang, K.Y. y Zeger, A.L. (1998). *Analysis of longitudinal data*. Oxford: Oxford University Press.
- Finn, J.D. (1969). Multivariate analysis of repeated measures data. *Multivariate Behavioral Research*, 4, 391-413.
- Goldstein, H. (1986). Multilevel mixed linear models analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- Goldstein, H. (1989). Models for multilevel response variables with an application to growth curves. En R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 107-125). San Diego, CA: Academic Press.
- Goldstein, H. (1995). *Multilevel Statistical Models*. New York: John Wiley and Sons.
- Greenhouse, S.W. y Geisser, S. (1959). On methods in analysis of profile data. *Psychometrika*, 24, 95-112.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.
- Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion. *Proceedings of the second Berkeley Symposium of Mathematical Statistics and Probability*. Berkeley: University of California Press.
- Hox, J.J. (1996). *Applied multilevel analysis*. New York: Springer-Verlag.
- Hox, J.J. (2002). *Multilevel analysis. Techniques and applications*. Hillsdale, NJ: Erlbaum.
- Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175.
- Huynh, H. y Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, 65, 1.582-1.589.
- Huynh, H. y Feldt, L.S. (1976). Estimation of the box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics*, 1, 69-82.
- Jennrich, R. y Schluchter, M.D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, 42, 805-820.
- Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67, 85-92.
- Keselman, H.J., Algina, J. y Kowalchuk, R.K. (2001). The analysis of repeated measures design: A review. *British Journal of Mathematical and Statistical Psychology*, 54, 1-20.
- Keselman, H.J., Algina, J. y Kowalchuk, R.K. (2002). A comparison of data analysis strategies for testing omnibus effects in higher-order repeated measures designs. *Multivariate Behavioral Research*, 37(3), 331, 357.
- Keselman, H.J., Algina, J., Kowalchuk, R.K. y Wolfinger, R.D. (1999). A comparison of recent approaches to the analysis of repeated measurement. *British Journal of Mathematical and Statistical Psychology*, 52, 63-78.
- Krefit, I.G.G., De Leeuw, J. y Van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *The American Statistician*, 48(4), 324-335.
- Laird, N.M. y Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Lindquist, E.F. (1953). *Design and analysis of experiments in psychology and education*. Boston MA: Houghton Mifflin.
- Littell, R.C., Pendergast, J.P. y Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19, 1.793-1.819.
- Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, 817-827.
- Longford, N.T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. En R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 297-310). San Diego, CA: Academic Press.
- Longford, N.T. (1993). *Random coefficient models*. Oxford, GB: Clarendon Press.
- Park, T. y Lee, J. (2002). Covariance models for nested repeated measures data: analysis of ovarian steroid secretion data. *Statistics in Medicine*, 21, 143-164.
- Plewis, I. y Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Education Research and Evaluation*, 4, 13-26.
- Potthoff, R.F. y Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-323.
- Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52, 447-458.
- Raudenbush, S.W. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research*, 13, 721-740.
- Raudenbush, S.W. y Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2<sup>nd</sup> Ed.). Thousand: Sage Publications.
- Rouanet, H. y LePine, D. (1970). Comparison between treatments in a repeated measurements design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 23, 147-163.
- Searle, S.R., Casella, G. y McCulloch, C.E. (1992). *Variance components*. New York: Wiley.
- Singer, J.D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24(4), 323-355.
- Snijders, T.A.B. y Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.
- Sullivan, L.M., Dukes, K.A. y Losina, E. (1999). An introduction to hierarchical linear modeling. *Statistics in Medicine*, 18, 855-888.
- Timm, N.H. y Mieczkowski, T.A. (1997). *Univariate and multivariate general linear models: theory and applications using SAS software*. Cary, NC: SAS Institute Inc.
- Tukey, J.W. (1949). One degree of freedom for nonadditivity. *Biometrics*, 5, 232-242.
- Van der Leeden, R. (1998a). Multilevel analysis of repeated measures data. *Quality and Quantity*, 32, 15-29.
- Van der Leeden, R. (1998b). Multilevel analysis of longitudinal data. En C.J.H. Bijleveld y L.T. Van der Kamp (Eds.), *Longitudinal data analysis. Designs, models and methods* (pp. 269-316). London: Sage publications.
- Van der Leeden, R., Vrijburg, K.E. y De Leeuw, J. (1986). A review of two different approaches for the analysis of growth data using longitudinal mixed linear models: comparing hierarchical linear regression (ML3, HLM) and repeated measures designs with structured covariance matrices (BMDP5V). *Computational Statistics and Data Analysis*, 21, 583-605.
- Wu, Y.B., Clopper, R.R. y Wooldridge, P.J. (1999). A comparison of traditional approaches to hierarchical linear modelling when analysing longitudinal data. *Research in Nursing and Health*, 22, 421-432.