Introducción a los modelos mixtos (SIA 3011003)

Profesor Juan Carlos Salazar-Uribe
jcsalaza@unal.edu.co

Es razonable pensar, que así como se puede formular un MLM con intersectos aleatorios o uno con intersectos y pendientes aleatorias (que son polinomios en la variable temporal t_{ij} o x_{ij} de grado cero y uno respectivamente), se pueden formular polinomios de mayor grado en la componente aleatoria del modelo. Por ejemplo, se podría pensar en un modelo cuadrático con interacción y coeficientes aleatorios de grado 2. Suponga que se tiene las covariables *AGE* y *GENDER* y una respuesta de interés Y. Un MLM individual, cuadrático, con interacción y con coeficientes aleatorios de grado 2 sería:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 AGE_{ij} + \beta_2 GENDER_i + \beta_3 AGE_{ij} \times GENDER_i + \\ &\quad b_{0i} + b_{1i} AGE_{ij} + b_{2i} AGE_{ij}^2 + \\ &\quad e_{ij} \end{aligned}$$

Crecimiento craneofacial en sujetos colombianos. Por ejemplo, considere estos 4 pacientes y un polinomio cuadrático mixto (datos reales):

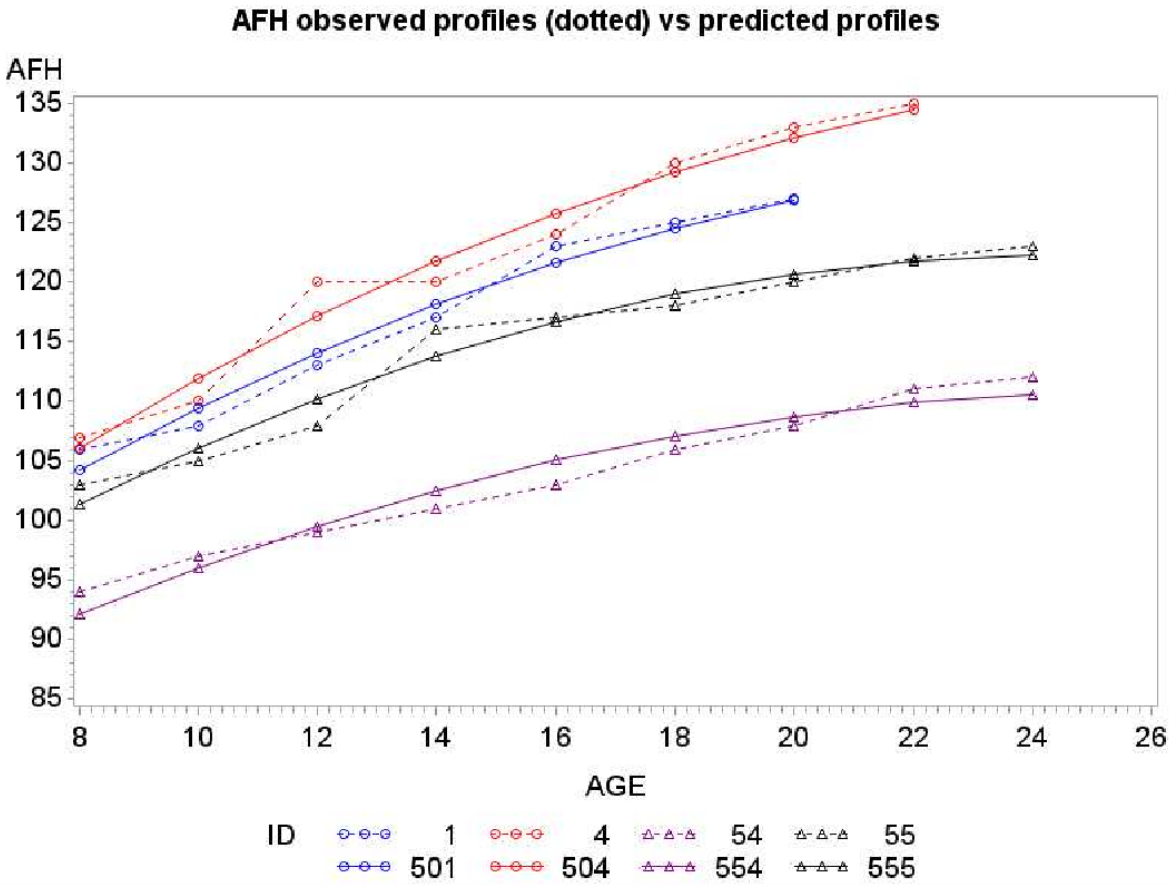

Figura 1: Datos de crecimiento facial

Sí se pueden formular modelos polinomiales de cualquier orden, pero hay que tener en cuenta que:

1) En general, modelos de interceptos o interceptos y pendientes aleatorias son en la práctica de mucha utilidad y permiten, sin sobreparametrizar, individualizar el MLM.
2) Polinomios de orden más allá de dos, generalmente en la práctica, pueden ser dificiles de justificar y ajustar y pueden llegar a generar inestabilidades numéricas.
3) Podrían no ajustarse bien a la realidad del fenómeno, es decir podrían no ser plausibles desde un punto de vista bioógico, económico o clínico.

Finalmente, no siempre es necesario especificar en la componente aleatoria polinomios, pues cualquier variable presente en los datos se puede declarar como aleatoria. Eso si, se debe poder justificar de manera clara y plausible su declaración como variable aleatoria antes de agregarla a la componente de efectos aletorios del modelo. POr ejemplo, en un estudio multi-centro, se podría declarar como aleatorio el hospital o el centro de salud al pensarlo como que ellos constituyen una m.a. de una gran población de hospitales o centros de salud. Otro aspecto importante de los MLM, es que se pueden ajustar aún si se tienen datos cross-seccionales (es decir, solo una observación por sujeto) y aún seguir declarando variables como aleatorias, si esto último se puede justificar y es razonable.

Siempre hay que tener en cuenta que un modelo estadístico es una aproximación a la realidad.

1. No hay un modelo "correcto" (olvide el 'santo grial' en modelamiento).
2. Un modelo es una herramienta para enfrentar una pregunta científica.
3. Un modelo útil combina los datos con información previa para abordar la pregunta o el asunto de interés.

Modelo lineal mixto como un modelo multinivel [1]. Los datos jerárquicos son comunes en muchos campos, desde productos farmacéuticos hasta agricultura y sociología. A medida que los datos y sus fuentes crecen, es factible que la información se observe en unidades anidadas a diferentes niveles, lo que requiere el enfoque de modelamiento multinivel.

[1] Pinheiro, J. C. and Chao, E. C. (2006), "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, 15, 58–81.
Rabe-Hesketh, S. and Skrondal, A. (2006), "Multilevel Modelling of Complex Survey Data," *Journal of the Royal Statistical Society*, Series A, 169, 805–827.

¿Qué es un modelo multinivel?. Los modelos multinivel (también conocidos como modelos lineales jerárquicos, modelos lineales de efectos mixtos, modelos mixtos, modelos de datos anidados, modelos de coeficientes aleatorios) son modelos estadísticos de parámetros que varían a más de un nivel. Los modelos multinivel son particularmente apropiados para diseños de investigación donde los datos de los participantes están organizados en más de un nivel; es decir, datos anidados. Las unidades de análisis suelen ser individuos (en un nivel inferior) que están anidados dentro de unidades contextuales/agregadas (en un nivel superior)

Piense, por ejemplo en procesos biológicos, psicológicos y sociales que influyen en la salud se producen en muchos niveles. Estos niveles, que se piensa influyen en un desenlace de salud o respuesta de interés (health outcome o resoponse) en orden del más pequeño al más grande, podrían ser:

1. Célula
2. Organo
3. Persona
4. Familia
5. Vecindario
6. Ciudad
7. Sociedad

y así, sucesivamente.

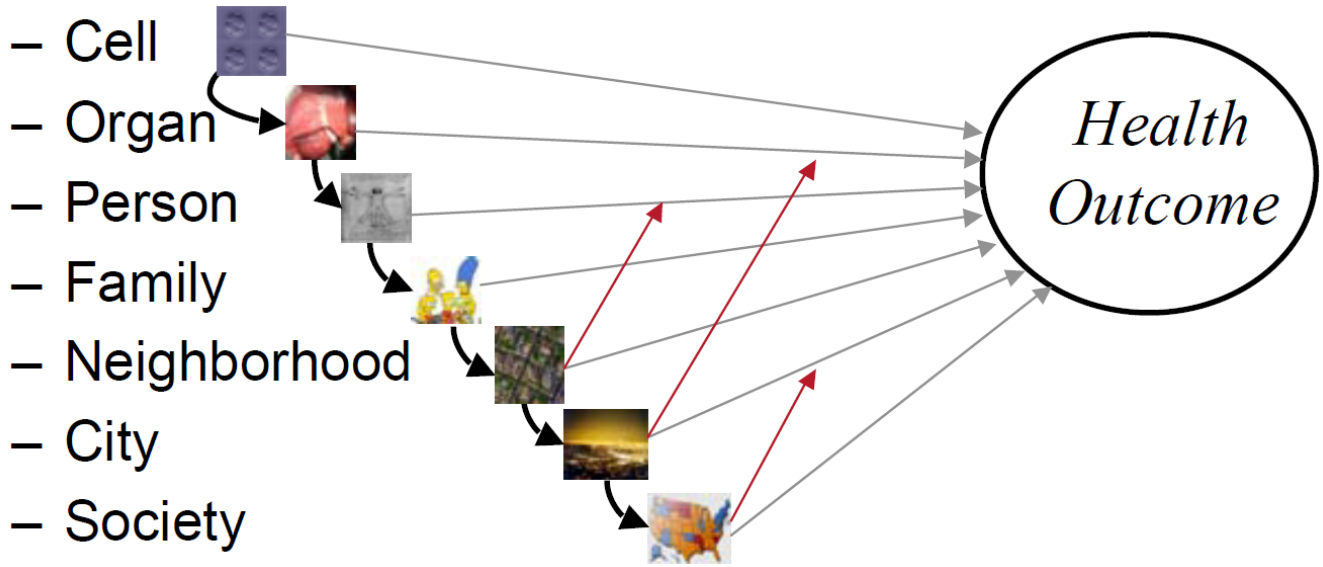

Figura 2: Health outcome: Alcohol Abuse. Red arrows represent interactions.

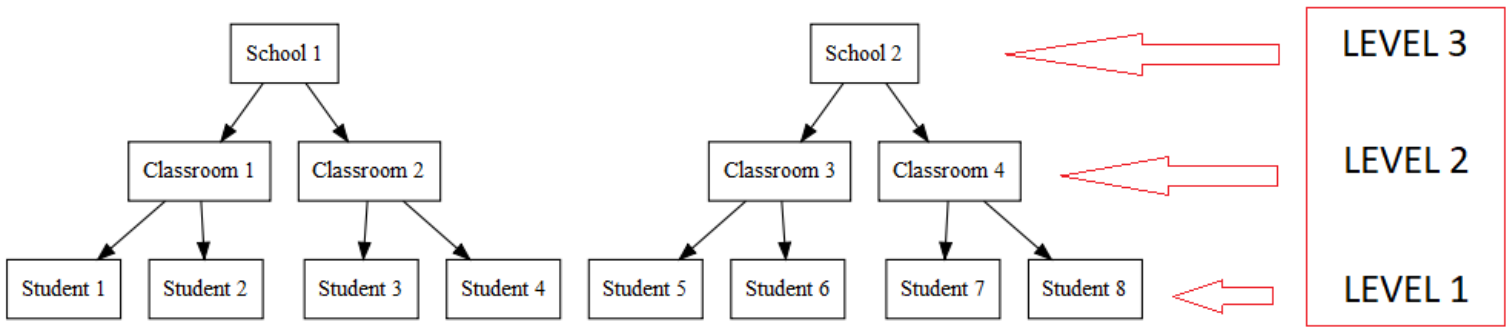

Figura 3: Ejemplo de una estructura educativa

Un análisis estadístico de la posible influencia de estos factores, debe considerar cada uno de estos niveles (por supuesto, si es posible recolectar información de todos y cada de ellos en cantidades adecuadas) y sus interacciones.

Ejemplo: abuso de alcohol

1. Célula: Neuroquímica
2. Órgano: Capacidad de metabolizar el etanol
3. Persona: susceptibilidad genética a la adicción
4. Familia: Abuso de alcohol en el hogar
5. Barrio: Disponibilidad de bares
6. Sociedad: Regulación; organizaciones;normas sociales

Los datos jerárquicos requieren técnicas analíticas especializadas que tengan en cuenta la interacción entre la información a diferentes niveles. Estas técnicas se conocen generalmente como modelos mixtos, que permiten desagregar la variabilidad de acuerdo a los distintos niveles involucrados en el análisis.

Como punto de partida, considere el siguiente ejemplo de la industria farmacéutica, notando el caracter jerárquico de los datos y preste especial atención a los múltiples niveles de análisis que se requieren. Brown y Prescott (1999)[2] discuten un ensayo aleatorizado multi-centro de hipertensión en el que los pacientes de cada centro se aleatorizan para reciben uno de tres medicamentos y luego se les hace seguimiento durante cuatro visitas.

[2] Brown, H. and Prescott, R. (1999), *Applied Mixed Models in Medicine*, New York: John Wiley & Sons.

Se registra la presión arterial diastólica (PAD) antes del tratamiento y en cada una de las cuatro visitas. Este estudio multi-centro tiene una estructura jerárquica de tres niveles:

1. Las visitas son las unidades que están al nivel-1.
2. Los pacientes son los grupos o clusters que están al nivel-2.
3. Los centros son los grupos o clusters que están al nivel-3.

Las visitas están anidadas dentro de los pacientes, que a su vez se anidan dentro de los centros.

Las unidades en los niveles que son más altos que el nivel 1 a veces se denominan clústeres. El tiempo de visita es una covariable de nivel 1. PAD basal y tratamiento varían solo de un paciente a otro y, por lo tanto, son covariables de nivel 2. No se miden covariables de nivel 3 en el centros.

¿Cuál es la justificación para distinguir las mediciones de la PAD según el paciente y el centro? pacientes en el mismo centro tienden a ser más similares entre sí que con los pacientes de otro centro. La razón para la similitud dentro del centro podría ser la cercanía de las residencias de los pacientes o la práctica médica compartida en el centro. Además, las mediciones repetidas de la PAD del mismo paciente están más cerca unas de otras o más relacionadas entre si que aquellas medidas de un paciente diferente.

La dependencia dentro del grupo (within-cluster dependency) hace que el modelo de regresión clásico basado en MCO sea inapropiado, pero se pueden usar modelos multinivel a fin de acomodar o tener en cuenta tal dependencia. Sin embargo, la correlación intra-grupos es más que una simple molestia. El diseño jerárquico proporciona valiosa información sobre cómo operan los procesos en diferentes niveles.

Los modelos Multinivel permiten desentrañar dicha información al incluir covariables en diferentes niveles y asignar variabilidad no explicada, a diferentes niveles. Por ejemplo, un modelo de tres niveles permite estimar los efectos de covariables a nivel de visita, paciente y centro en el estudio multi-centro. Además, puede incluir efectos aleatorios para abordar la variabilidad que no se explica por esas covariables. Estos efectos aleatorios son especificados en niveles definidos por clústeres anidados. El resultado es que los modelos multinivel para datos jerárquicos son un caso especial de modelos de efectos mixtos.

LOS MODELOS MULTINIVEL SON MODELOS MIXTOS. Uno de los puntos clave de los modelos multinivel es que la estructura jerárquica de los datos hace que sea natural concebir el modelo por etapas. Para ver esto, considere un modelo de tres niveles que tiene efectos fijos en el primer y segundo nivel e intersectos aleatorios y pendientes en el segundo y tercer nivel. En el siguiente desarrollo, un superíndice ℓ denota el nivel ℓ, y i, j y k denotan los índices de nivel-1, nivel-2, y unidades de nivel-3, respectivamente.

Un modelo para estos datos se puede especificar en tres etapas. En cada etapa, usted incorpora covariables y efectos aleatorios para explicar la variación específica del nivel alrededor del intercepto promedio y la pendiente promedio. El modelo de nivel-1 postula una relación lineal entre la respuesta observada y_{ijk} y la covariable de nivel-1 (o etapa 1, stage 1) $x_{ijk}^{(1)}$ $(\ell = 1)$ de la siguiente manera:

$$Y_{ijk} = \alpha_{0jk} + \alpha_{1jk} x_{ijk}^{(1)} + e_{ijk}$$

Al siguiente nivel o etapa, el intercepto y la pendiente de este modelo de nivel-1 varía de acuerdo a las siguientes relaciones con la covariable de nivel-2 $x_{jk}^{(2)}$:

$$\alpha_{0jk} = \beta_{00k} + \beta_{01k} x_{jk}^{(2)} + \gamma_{0jk}^{(2)}$$

$$\alpha_{1jk} = \beta_{10k} + \beta_{11k} x_{jk}^{(2)} + \gamma_{1jk}^{(2)}$$

Finalmente, los interceptos del nivel-2 varían entre las unidades del nivel-3 de acuerdo a los modelos del nivel-3 siguientes:

$$\beta_{00k} = \lambda_{00} + \gamma_{0k}^{(3)}$$

$$\beta_{10k} = \lambda_{10} + \gamma_{1k}^{(3)}$$

Además de las respuestas, covariables y los parámetros que las relacionan, este modelo de tres niveles incorpora términos aleatorios en cada uno de los tres niveles: el residual de nivel-1 es e_{ijk}, y los vectores de efectos aleatorios en el nivel-2 y el nivel-3 son

$$\gamma_{jk}^{(2)} = \left(\gamma_{0jk}^{(2)},\ \gamma_{1jk}^{(2)}\right)$$

y

$$\gamma_{k}^{(3)} = \left(\gamma_{0k}^{(3)},\ \gamma_{1k}^{(3)}\right)$$

Respectivamente.

El supuesto distribucional usual para los efectos aleatorios es normalidad:

$$\gamma_{jk}^{(2)} \sim N(0, G^{(2)})$$

y

$$\gamma_{k}^{(3)} \sim N(0, G^{(3)})$$

Las matrices de covarianza $G^{(2)}$ y $G^{(3)}$ especifican cómo varían el intercepto aleatorio y la pendiente entre las unidades de nivel-2 y unidades de nivel-3, respectivamente.

El vector residual de una unidad de nivel-3 se maneja de manera similar:

$$e_k^{(3)} \sim N(0, R_k^{(3)})$$

Esto completa la formulación del modelo de tres etapas de un modelo multinivel para estos datos multinivel.

Con el fin de ajustar este modelo de tres niveles, es necesario distinguir los parámetros fundamentales. Sustituyendo los modelos de nivel-2 en los modelos de nivel-1 y luego los modelos de nivel-2 en las ecuaciones del nivel-3 se obtiene la siguiente expresión. Veamos en detalle esta afirmación. Considere estas 5 ecuaciones del modelo multinivel:

$$Y_{ijk} = \alpha_{0jk} + \alpha_{1jk} x_{ijk}^{(1)} + e_{ijk} \qquad (1)$$

$$\alpha_{0jk} = \beta_{00k} + \beta_{01k} x_{jk}^{(2)} + \gamma_{0jk}^{(2)} \qquad (2)$$

$$\alpha_{1jk} = \beta_{10k} + \beta_{11k} x_{jk}^{(2)} + \gamma_{1jk}^{(2)} \qquad (3)$$

$$\beta_{00k} = \lambda_{00} + \gamma_{0k}^{(3)} \qquad (4)$$

y

$$\beta_{10k} = \lambda_{10} + \gamma_{1k}^{(3)} \qquad (5)$$

Reemplazando (2) y (3) en (1)

$$\begin{aligned} Y_{ijk} &= \beta_{00k} + \beta_{01k}x_{jk}^{(2)} + \gamma_{0jk}^{(2)} + \left(\beta_{10k} + \beta_{11k}x_{jk}^{(2)} + \gamma_{1jk}^{(2)}\right) x_{ijk}^{(1)} + e_{ijk} \\ &= \beta_{00k} + \beta_{01k}x_{jk}^{(2)} + \gamma_{0jk}^{(2)} + \beta_{10k}x_{ijk}^{(1)} + \beta_{11k}x_{jk}^{(2)}x_{ijk}^{(1)} + \gamma_{1jk}^{(2)}x_{ijk}^{(1)} + e_{ijk} \\ &= \beta_{00k} + \beta_{10k}x_{ijk}^{(1)} + \beta_{01k}x_{jk}^{(2)} + \beta_{11k}x_{jk}^{(2)}x_{ijk}^{(1)} + \gamma_{0jk}^{(2)} + \gamma_{1jk}^{(2)}x_{ijk}^{(1)} + e_{ijk} \end{aligned}$$

Reemplazando (4) y (5) en β_{00k} y β_{10k} en esta última ecuación:

$$\begin{aligned} Y_{ijk} &= \lambda_{00} + \gamma_{0k}^{(3)} + \left(\lambda_{10} + \gamma_{1k}^{(3)}\right) x_{ijk}^{(1)} + \beta_{01k}x_{jk}^{(2)} + \beta_{11k}x_{jk}^{(2)}x_{ijk}^{(1)} + \gamma_{0jk}^{(2)} + \gamma_{1jk}^{(2)}x_{ijk}^{(1)} + e_{ijk} \\ &= \lambda_{00} + \lambda_{10}x_{ijk}^{(1)} + \beta_{01k}x_{jk}^{(2)} + \beta_{11k}x_{ijk}^{(1)}x_{jk}^{(2)} + \gamma_{0jk}^{(2)} + \gamma_{1jk}^{(2)}x_{ijk}^{(1)} + \gamma_{0k}^{(3)} + \gamma_{1k}^{(3)}x_{ijk}^{(1)} + e_{ijk} \end{aligned}$$

Es decir, se obtiene la siguiente expresión:

$$\begin{aligned} Y_{ijk} &= \lambda_{00} + \lambda_{10} x_{ijk}^{(1)} + \beta_{01k} x_{jk}^{(2)} + \beta_{11k} x_{ijk}^{(1)} x_{jk}^{(2)} + \\ &\quad \gamma_{0jk}^{(2)} + \gamma_{1jk}^{(2)} x_{ijk}^{(1)} + \\ &\quad \gamma_{0k}^{(3)} + \gamma_{1k}^{(3)} x_{ijk}^{(1)} + \\ &\quad e_{ijk} \end{aligned}$$

Esta ecuación permite identificar múltiples fuentes de variación. La primera línea del lado derecho de la ecuación (la de color azul) especifica los efectos fijos de este modelo: el intercepto general, las covariables de nivel-1 y nivel-2, y su interacción. La segunda y tercera líneas (las de color rojo) especifican los efectos aleatorios en el segundo y tercer nivel, que son seguidas por el residual en la última línea. Esta forma de especificación del modelo, especialmente la partición de efectos fijos y la separación adicional de los efectos aleatorios de acuerdo con sus niveles, facilita la implementación de estos modelos en los software estándar tales como SAS y R, entre otros.

Este desglose es diferente al de las ecuaciones separadas que los investigadores en ciencias educativas, sociales y del comportamiento a menudo usan para conceptualizar modelos multinivel, pero es más conveniente para el cálculo y vincula los modelos multinivel con el área más grande de los modelos mixtos. Es decir, un modelo multinivel, en forma matrical general, es ahora un modelo mixto de la forma:

$$Y = X\beta + Z\gamma + e$$

donde

$$\gamma \sim N(0, G) \quad \text{y} \quad e \sim N(0, R)$$

γ y e son independientes y $V(Y) = ZGZ^T + R$.

Aquí, Y y e son vectores de respuestas y residuales, respectivamente; X y Z son matrices de diseño para los los efectos fijos y efectos aleatorios, respectivamente; y β y γ son vectores de efectos fijos y aleatorios, respectivamente. Las ecuaciones anteriores definen un modelo lineal mixto y los parámetros de covarianza en R y G, llamelos θ, puede estimarse usando métodos de máxima verosimilitud (ML) o máxima verosimilitud restringida (REML), métodos que se discuten en detalle más adelante en el curso.

Estos métodos ML y REML también se exponen extensa y claramente en Littell et al. (1996)[3]. Después de tener la estimación del parámetro de covarianza, $\widehat{\theta}$, se puede obtener el mejor estimador lineal empírico insesgado (EBLUE) de β y el mejor predictor lineal empírico insesgado (EBLUP) de γ.

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

Modelo de intersectos y pendientes aleatorias. Siendo el MLM de interceptos aleatorios uno de los más usados, este otro, es quizás el más utilizado. Cada sujeto difiere de los demás en términos tanto de su intersecto como de su pendiente. El modelo se expresa como:

$$y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_2 G_i + \beta_3 G_i \times T_{ij} + b_{0i} + b_{1i} T_{ij} + \epsilon_{ij}$$

donde

$$\epsilon_{ij} \sim N\left(0, \sigma^2_\epsilon\right)$$

y

$$\mathbf{b_i} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{b_0} & \sigma_{b_0,b_1} \\ \sigma_{b_0,b_1} & \sigma^2_{b_1} \end{bmatrix}\right)$$

Modelo de intersectos aleatorios y pendientes aleatorias. En forma matricial, se expresa como :

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & T_{i1} & G_i & G_i \times T_{i1} \\ 1 & T_{i2} & G_i & G_i \times T_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_{in_i} & G_i & G_i \times T_{in_i} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 & T_{i1} \\ 1 & T_{i2} \\ \vdots & \vdots \\ 1 & T_{in_i} \end{bmatrix} \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}$$

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b_i} + \boldsymbol{\epsilon}_i$$

con $Var(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma_\epsilon^2\mathbf{I}_{n_i}$. Note que cuando no se especifican efectos aleatorios $Var(\mathbf{y}_i) = \sigma_\epsilon^2\mathbf{I}_{n_i}$.

Observe que en la expresión para la varianza de Y de un MLM arbitrario,

$$Var\left(\mathbf{Y}_i\right) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \Sigma_i, \quad i = 1, \ldots, N$$

se observan dos matrices muy importantes, $\mathbf{D}$ y Σ_i[4]. Estas matrices especifican de manera explicita las componentes de varianza de las fuente de variabilidad presentes en el modelo, que son los efectos aleatorios y la componente de error, respectivamente. Pero, ¿cómo seleccionarlas de manera apropiada? Se verá, que no hay una respuesta única a esta importante pregunta, pero se mostrará que esta expresión constituye una gran fortaleza de estos MLM.

[4]Algunas veces y dependiendo del texto o paper de referencia $\mathbf{D} = \mathbf{G}$ y $\Sigma_i = \mathbf{R_i}$. En estas notas se pueden usar cualquiera de las notacione disponibles siempre y cuando se especifique de antemano cuál es cuál.