Introducción a los modelos mixtos (SIA 3011003)

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

MLM para efectos aleatorios del tratamiento. A veces no estamos interesados en la media de los tratamientos sino en la varianza de los efectos del tratamiento. Por ejemplo, estamos comparando la producción promedio de cuatro máquinas que son operadas por diferentes personas (operador de máquina). El efecto de la máquina es interesante, pero el efecto del operador de la máquina no lo es, por lo que podemos elegir al operador de la máquina al azar e incluirlo en el estudio. Lo que nos interesa es el efecto de la variabilidad asociada a cada operador de máquina y la variabilidad debida a la interacción máquina-máquina-operador.

Si los operadores de la máquina operan las 4 máquinas, también podemos estudiar la variabilidad de los efectos de interacción. Podríamos estimar los efectos fijos (las máquinas) que se operan para aquellos operadores de máquinas que representan una muestra aleatoria de una población de operadores de máquinas (efecto aleatorio debido al operador de máquinas).

MLM para datos longitudinales. Hay varias observaciones para cada item (generalmente a través del tiempo). La variable 'tiempo' no se puede aleatorizar, por lo que debemos modelar la correlación debido a medidas repetidas. Además, el tiempo es lineal en el sentido de que t_1 es primero que t_2, t_2 es primero que t_3, y así sucesivamente.

Debido a la estructura longitudinal de los datos, que no es aleatoria (es decir, la segunda observación viene después de la primera, y así sucesivamente), tenemos que modelar la correlación entre las observaciones longitudinales. Podríamos hacer esto explícitamente usando los elementos fuera de la diagonal de la matriz $\mathbf{\Sigma}$ (matriz de covarianza para los errores o, si el LMM se formula de manera condicional, matriz de covarianza condicional de las observaciones dados los efectos aleatorios: $Var\left(\mathbf{Y}|\ \mathbf{b}\right)$.

Otra forma de hacerlo es introduciendo efectos aleatorios de los sujetos (unidades que contienen varias observaciones que se tomaron a lo largo del tiempo). En otras palabras, mediante efectos aleatorios se induce una correlación positiva entre observaciones longitudinales. Si hay pocas observaciones a lo largo del tiempo (y en cada momento medimos la misma característica) podríamos formular directamente un modelo mixto con una estructura de correlación específica para los errores (matriz $\boldsymbol{\Sigma}$).

Si, por el contrario, hay demasiadas observaciones a lo largo del tiempo, y no necesariamente son las mismas, probablemente sea mejor introducir la correlación de forma indirecta a través de efectos aleatorios del sujeto o mediante el uso de perfiles aleatorios específicos del sujeto (o curvas aleatorias específicas del sujeto).

Supongamos por un momento que no hemos especificado efectos aleatorios. Si tenemos N unidades independientes, entonces la matriz Σ tendrá una estructura de bloques diagonales:

$$\Sigma = \begin{pmatrix} \Sigma_1 & , \mathbf{0} & , \dots & , \mathbf{0} \\ \mathbf{0} & , \Sigma_2 & , \dots & , \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & , \mathbf{0} & \dots & , \Sigma_N \end{pmatrix}$$

Cada uno de estos bloques corresponde a una unidad experimental particular y tendrá una estructura que reflejará tanto las varianzas como las covarianzas dentro de esa unidad.

Aquí presentamos algunas de las más comunes en la práctica pero más adelante discutiremos otras estructuras.

Si los errores son independientes,

$$\Sigma_i = \sigma_e^2 \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

(recuerde que si no se especifican efectos aleatorios, $var(\mathbf{y}_i) = \Sigma_i$)

Por otro lado, si especificamos los efectos aleatorios y la estructura de simetría compuesta:

$$Var(\mathbf{y}_i) = \mathbf{V}_i = \left(\sigma_b^2 + \sigma_e^2\right) \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}$$

Si $\rho > 0$, esto es equivalente a pensar en $\mathbf{\Sigma}_i$ como independientes y agregar un intersecto aleatorio al nivel del sujeto. Las varianzas son homogéneas. Existe una correlación entre dos mediciones separadas, pero se supone que la correlación es constante independientemente de cuán separadas estén las mediciones.

Modelo con errores incorrelacionados. Suponga que el modelo para un sujeto en particular (tenemos T medidas por sujeto y n sujetos independientes) viene dado por (no se especifican efectos aleatorios)

$$\begin{aligned}
Y_t &= \mathbf{x}_t'\beta + \boldsymbol{\epsilon}_t, \quad t = 1, \cdots, T \\
\epsilon_t &= \xi_t + \rho\epsilon_{t-1}, \quad t = 2, \cdots, T \\
\rho &\in (-1, 1) \\
\xi_t &\sim N\left(0, \sigma^2\right), \quad t = 2, \cdots, T \\
\epsilon_t &\sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)
\end{aligned}$$

De estos supuestos tenemos que los llamados 'shocks' o 'impactos' $\{\xi_2, \xi_3, \cdots, \xi_T\}$ son variables aleatorias independientes e idénticamente distribuidas (i.i.d).

ϵ_1 es independiente de las $\{\xi_2, \xi_3, \cdots, \xi_T\}$ and $\epsilon_1 \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$.

Escriba los términos de error como una serie geométrica en términos de 'shocks' pasados

$$
\begin{aligned}
\epsilon_2 &= \xi_2 + \rho\epsilon_1 \\
\epsilon_3 &= \xi_3 + \rho\epsilon_2 = \xi_3 + \rho\xi_2 + \rho^2\epsilon_1 \\
\epsilon_4 &= \xi_4 + \rho\epsilon_3 = \xi_4 + \rho\xi_3 + \rho^2\xi_2 + \rho^3\epsilon_1 \\
\vdots &\quad \vdots \\
\epsilon_t &= \xi_t + \rho\epsilon_{t-1} = \xi_t + \rho\xi_{t-1} + \rho^2\xi_{t-2} + \cdots + \rho^{t-1}\epsilon_1
\end{aligned}
$$

Las varianzas de los ϵ_t's son iguales a σ_ϵ^2

$$\begin{aligned} Var(\epsilon_t) &= \sigma^2 + \rho^2 Var(\epsilon_{t-1}) \\ \sigma_\epsilon^2 &= \sigma^2 + \rho^2 \sigma_\epsilon^2 \\ \sigma_\epsilon^2 &= \frac{\sigma^2}{1-\rho^2} \end{aligned}$$

Usando la serie geométrica anterior para ϵ_t, las covarianzas son:

$$E\left(\epsilon_t \epsilon_{t-u}\right) = \sigma^2 \frac{\rho^u}{1-\rho^2}, \qquad u > 0$$

La matriz de covarianzas de los términos de error (que es igual a la matriz de covarianzas de las respuestas y_t)

$$E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right) = \boldsymbol{\Sigma} = \frac{\sigma^2}{1-\rho^2}\boldsymbol{\Gamma}$$

donde

$$\boldsymbol{\Gamma} = \left[\begin{array}{ccccc} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ \rho & 1 & \rho & \cdots & \rho^{t-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \cdots & 1 \end{array}\right]$$

$$\boldsymbol{\Gamma}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & 0 & \cdots & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & -\rho & 1+\rho^2 & -\rho \\ 0 & \cdots & \cdots & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

$$det(\boldsymbol{\Gamma}) = \left(1-\rho^2\right)^{T-1}$$

Defina los residuales

$$r_t = y_t - \mathbf{x}_t' \boldsymbol{\beta}$$

La distribución condicional de $\mathbf{Y}$ dado $\boldsymbol{\beta}$, σ y ρ is

$$\begin{aligned} f\left(\mathbf{Y} \,|\boldsymbol{\beta}, \sigma, \rho\right) &\propto |\boldsymbol{\Sigma}|^{-1/2} \, exp\left\{-\frac{1}{2}\left(\mathbf{Y} - \mathbf{X}\beta\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{Y} - \mathbf{X}\beta\right)\right\} \\ &\propto \frac{\left(1-\rho^2\right)^{1/2}}{\sigma^T} exp\left\{-\frac{1}{2\sigma^2}\left[\left(1-\rho^2\right) r_1^2 + \sum_{t=2}^{T}\left(r_t - \rho r_{t-1}\right)^2\right]\right) \end{aligned}$$

Note que

$$r_t - \rho r_{t-1} = y_t - \rho y_{t-1} - \left(\mathbf{x}_t - \rho \mathbf{x}_{t-1}\right)' \beta$$

Autorregresivo de orden 1 (AR(1)[1], válido si las observaciones están igualmente espaciadas, o al menos de manera aproximada):

$$Var(\mathbf{y}_i) = \mathbf{V}_i = \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}$$

If observations are not equally spaced one can use:

$$Corr\left(e_j, e_{j'}\right) = \rho^{\left|t_j - t_{j'}\right|}, \quad j \neq j'$$

[1] La estructura AR(1) tiene varianzas homogéneas y correlaciones que disminuyen exponencialmente con la distancia. Significa que la variabilidad en una medición, digamos el recuento de glóbulos blancos, es constante independientemente de cuando lo mides. También significa que dos medidas que están una al lado de la otra en el tiempo van a estar bastante correlacionadas (dependiendo del valor de ρ), pero a medida que las medidas se alejan más y más, son menos correlacionados.

Selección de una estructura de covarianza (algunos preliminares). Seleccionar modelos para la media es una tarea necesaria. Un enfoque consiste en ajustar diferentes estructuras de covarianza y efectos aleatorios (matrices $\mathbf{D}$ y $\boldsymbol{\Sigma}$) y luego, mediante un criterio adecuado (como LRT, BIC, AIC) elegimos el sugerido por el criterio. Sin embargo, se debe tener cuidado y evitar mezclar estructuras que puedan causar que el modelo no sea identificable[2] (por ejemplo, un intersecto aleatorio y CS al mismo nivel)

[2] Los parámetros o modelos se denominan no identificables si dos conjuntos de parámetros diferentes conducen a la misma distribución de probabilidad. La identificabilidad del modelo está estrechamente relacionada con la indistinguibilidad del modelo. El objetivo del análisis de indistinguibilidad es determinar si diferentes modelos son capaces de ajustarse a los datos de entrada y salida disponibles

Estrategia de análisis. Una vez seleccionada la estructura para la covarianza, comprobamos de nuevo la parte fija del modelo. Por ejemplo, cuando ajustamos un modelo de regresión, primero seleccionamos las variables y luego verificamos los supuestos. Luego interpretamos los coeficientes estimados, probamos las hipótesis nulas y sacamos conclusiones. En el ANOVA interpretamos interacciones, comparamos medias y probamos contrastes para sacar conclusiones.

Modelo autorregresivo de orden uno (AR(1)). Supongamos que $\{Y(t) : t = \cdots, -1, 0, 1, \cdots\}$ es una serie de tiempo discreta. Una serie de tiempo autorregresiva de orden uno, designada por $AR(1)$, es generada por la ecuación

$$Y(t) = \rho Y(t-1) + \epsilon(t)$$

donde $\epsilon(t)$ son variables aleatorias independientes con media 0 y varianza σ_ϵ^2. Los parámetros del proceso son ρ y σ_ϵ^2.

Aquí presentamos algunos resultados conocidos:

- $Cov\left[\epsilon(t),\, \epsilon(t-s)\right] = 0$ para todos los $s > 0$
- $Cov\left[Y(t-s),\, \epsilon(t)\right] = 0$ para todos los $s > 0$
- Si $\rho = 0$ el proceso se conoce como "ruido blanco"
- Si $\rho > 0$ el proceso varía alrededor de $E\left[Y(t)\right] = \mu$

Si el proceso empieza en $Y(0) = 0$ entonces

$$\begin{aligned} Y(1) &= \epsilon(1) \\ Y(2) &= \epsilon(1) + \epsilon(2) \\ \vdots & \quad \vdots \\ Y(t) &= \sum_{i=1}^{t} \epsilon(i) \end{aligned}$$

de aquí

$$Var\left[Y(t)\right] = t\sigma_\epsilon^2$$

Si se cumple el supuesto de estacionariedad de segundo orden, implica que la varianza $Var\left[Y(t)\right]$ es constante y no depende de t. Suponga que $E\left[Y(t)\right]=0$. La función de covarianza $C(k)$, para $k>0$ se define como:

$$\begin{aligned}
C(k) &= E\left[Y(t)Y(t-k)\right] \\
&= E\left[\left\{\rho Y(t-1)+\epsilon(t)\right\} Y(t-k)\right] \\
&= \rho E\left[Y(t-1)Y(t-k)\right]+E\left[\epsilon(t)Y(t-k)\right] \\
&= \rho E\left[Y(t-1)Y(t-k)\right]+0 \\
&= \rho C(k-1)
\end{aligned}$$

de aquí, se tiene que

$$C(k)=\rho^{k}C(0)=\rho^{k}Var\left[Y(t)\right]$$

Para encontrar la $Var\left[Y(t)\right]$ considere

$$
\begin{aligned}
C(0) &= E\left[Y(t)Y(t)\right] \\
&= E\left[\left\{\rho Y(t-1)+\epsilon(t)\right\}\left\{\rho Y(t-1)+\epsilon(t)\right\}\right] \\
&= \rho^2 E\left[Y(t-1)^2\right] + E\left[\epsilon(t)^2\right] \\
&= \rho^2 C(0) + \sigma_\epsilon^2
\end{aligned}
$$

Así,

$$
C(0) = \frac{\sigma_\epsilon^2}{1-\rho^2}
$$

La función de autocorrelación será:

$$R(k) = Cor\left[Y(t),\ Y(t-k)\right] = \frac{C(k)}{C(0)} = \rho^k \frac{C(0)}{C(0)} = \rho^k$$

La matriz de covarianza de

$$\mathbf{Y}(t) = \begin{bmatrix} \vdots \\ Y(t-2) \\ Y(t-1) \\ Y(t) \end{bmatrix}$$

está dada por

$$VarCov\left(\mathbf{Y}(t)\right) = \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & 1 \end{bmatrix}$$

Inferencia en el MLM. Se discutirán algunos detalles técnicos sobre MLM. Específicamente, se presenta teoría relacionada con el llamado modelo marginal[3] y con el modelo jerárquico (por ejemplo, el MLM en sí mismo es un modelo jerárquico (Verbeke y Molenberghs 2000, capítulos 3 y 5)). Luego discutimos aspectos asociados a la estimación del MLM y presentamos las ventajas y desventajas de los métodos estudiados.

[3] Los modelos marginales (o modelos de promedio poblacional) son a menudo contrastados con modelos condicionales (sujeto-específicos o modelos de efectos aleatorios o multinivel)

Muchos tipos de datos, incluidos los datos de observación recopilados en las ciencias humanas y biológicas, tienen una estructura jerárquica o agrupada. Por ejemplo, los niños con los mismos padres tienden a ser más parecidos en sus características físicas y mentales que los individuos elegidos al azar de la población en general. Los individuos pueden anidarse aún más dentro de áreas geográficas o instituciones tales como escuelas o empleadores. Las estructuras de datos multinivel también surgen en estudios longitudinales donde las respuestas de un individuo a lo largo del tiempo se correlacionan entre sí[4]

[4] Fuente: http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html

Los modelos multinivel reconocen la existencia de tales jerarquías de datos al permitir componentes residuales en cada nivel de la jerarquía. Por ejemplo, un modelo de dos niveles que permita agrupar los resultados de los niños dentro de las escuelas, incluiría residuos a nivel del niño y de la escuela. Así, la varianza residual se divide en un componente entre escuelas (la varianza de los residuales a nivel de escuela) y un componente dentro de la escuela (la varianza de los residuales a nivel de niño). Los residuos escolares, a menudo llamados efectos escolares, representan características escolares no observadas que afectan los resultados de los niños. Son estas variables no observadas las que conducen a la correlación entre los resultados de los niños de la misma escuela[5].

[5] Fuente: http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html

- Los modelos son herramientas de inferencia (Un modelo estadístico es una aproximación a la realidad)
- La elección del modelo está determinada por la pregunta científica

¿Objetivo científico para la inferencia?

- **Media marginal**: respuesta promedio en la población
- **Media condicional**: Dadas otras respuestas del sujeto o dados efectos aleatorios no observados.

Modelos jerárquicos versus Modelos marginales. Recuerde que el MLM está especificado como:

$$\begin{aligned}
\mathbf{Y}_i &= \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\
\mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}) \\
\boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)
\end{aligned}$$

y

$$\mathbf{b}_1, \ldots, \mathbf{b}_m \text{ y } \epsilon_1, \ldots, \epsilon_m$$

son independientes.

También podría formularse un **Modelo condicional**:

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i &\sim N(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \Sigma_i) \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}) \end{aligned}$$

(Primero se formula un modelo para $\mathbf{Y}_i$ dado $\mathbf{b}_i$ y luego un modelo para $\mathbf{b}_i$. No se restringue solo al caso formal)

La función de densidad marginal de $\mathbf{Y}_i$ es:

$$f(\mathbf{y}_i) = \int f\left(\mathbf{y}_i|\mathbf{b}_i\right) f(\mathbf{b}_i) d\mathbf{b}_i$$

la cual está asociada a una distribución normal n-dimensional con vector de medias $\mathbf{X}_i\beta$ y matriz de varianzas y covarianzas

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \Sigma_i$$

Por lo tanto, el **modelo marginal** obtenido mediante un enfoque de dos etapas, establece supuestos específicos sobre la estructura de dependencia para la media y la covarianza $\mathbf{X}_i$ y $\mathbf{Z}_i$, respectivamente. Específicamente:

$$\textbf{Media implicita}: \; \mathbf{X}_i\beta$$

y

$$\textbf{Covarianza implicita}: \; \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \Sigma_i$$

Nótese que el modelo jerárquico implica el modelo marginal pero no viceversa. La razón de esto, es que en la formulación de la media y la covarianza implícitas no hay efectos aleatorios, solo componentes de covarianza, por lo que no se puede recuperar el modelo jerárquico del modelo marginal. El modelo marginal es importante ya que la inferencia en el MLM se basa en este modelo y no en el modelo jerárquico. El modelo marginal se puede ajustar a un conjunto de datos usando SAS PROC MIXED© o la función *lme* de la librería *lme*4 del R©.

Notas:

1. En la práctica, basar la inferencia en el modelo marginal, podría causar problemas numéricos durante la optimización del proceso de verosimilitud ya que no hay garantía de convergencia. Sin embargo, esto es válido para cualquier modelo ya que la convergencia depende fuertemente de los datos a mano.
2. Las inferencias basadas en el modelo marginal no asumen explícitamente la presencia de efectos aleatorios.