Introducción a los modelos mixtos (SIA 3011003)

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

Ilustración de un uso de estadísticos de resumen para evaluar normalidad de los efectos aleatorios. De la teoría estadística se sabe que si una variable aleatoria multivariada tiene una distribución normal multivariada, sus marginales son normales univariadas.

Teorema[1]. Si $\mathbf{Y} \sim N_n(\boldsymbol{\theta}, \mathbf{D})$, entonces la distribución marginal de cualquier subconjunto de elementos de $\mathbf{Y}$ se distribuye también de acuerdo a una distribución multivariada.

Por ejemplo, si un vector

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N_2 \left(\boldsymbol{\theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_0 b_1} \\ \sigma_{b_1 b_0} & \sigma^2_{b_1} \end{pmatrix} \right)$$

Entonces, de acuerdo a este teorema $b_0 \sim N_1\left(0, \sigma^2_{b_0}\right)$ y $b_1 \sim N_1\left(0, \sigma^2_{b_1}\right)$.

[1] Seber, G.A.F. (1977). *Linear Regression Analysis*. John Wiley & Sons: New York

Gráficamente, por ejemplo, la situación es como sigue para $n = 2$ ($\mu_1 = \mu_2 = 0, \rho = 0.7, \sigma_1 = 1, \sigma_2 = 0.8$):

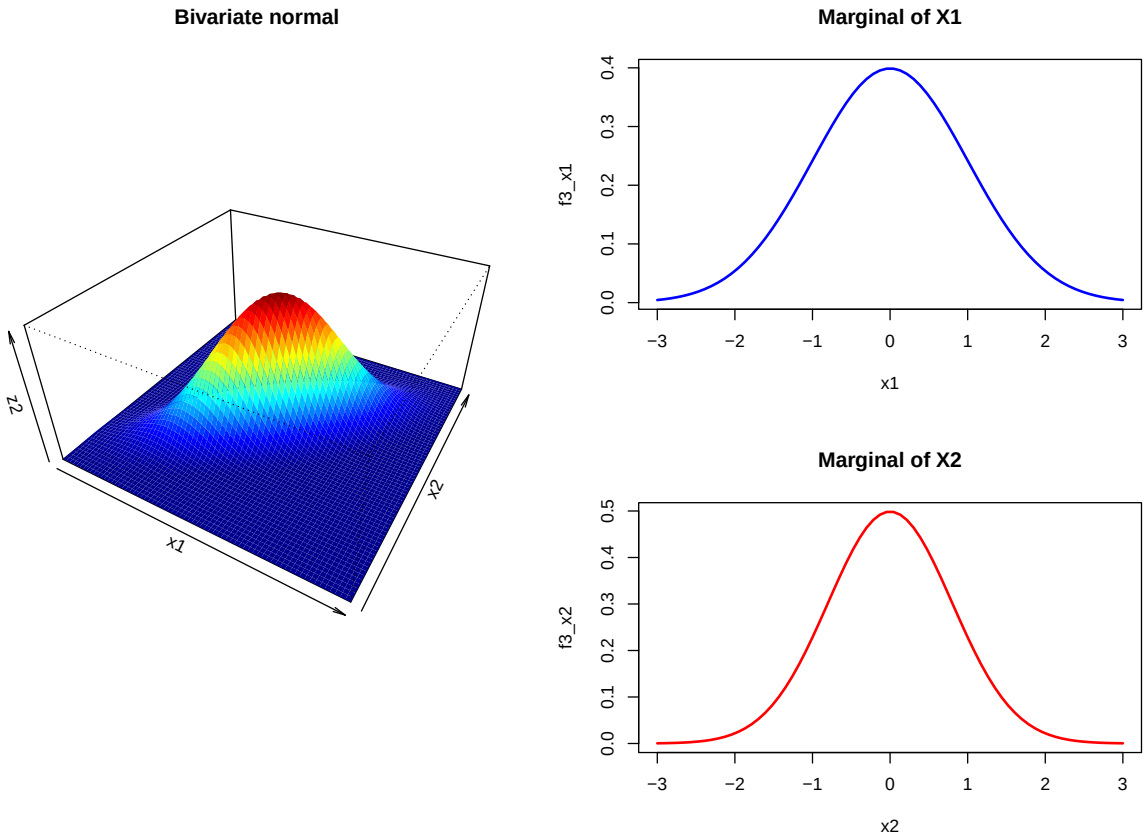

Figura 1: Normal bivariada y sus marginales normales

Supongamos que tenemos los siguientes 5 perfiles de sujetos del Estudio de salud cardíaca de Framingham (Framingham Heart Health Study FHHS)[2]. A partir de estos perfiles, parece que cada sujeto tiene su propio intersecto y pendiente aleatorios.

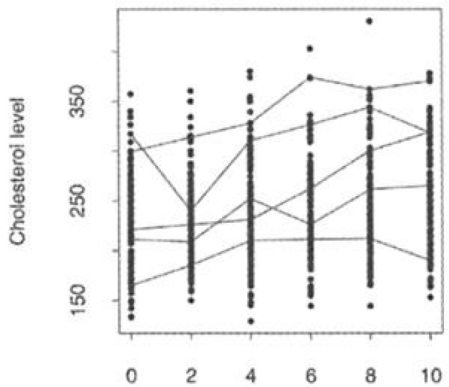

Figura 2: Perfiles de evolución de niveles de colesterol para 5 personas del FHHS

[2]Fuente: Zhang, D y Davidian, M. (2001). Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics*, vol. 57, nº 3, pages. 795-802

Por esta razón, podríamos suponer que el vector de efectos aleatorios:

$$\mathbf{b}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} \sim N_2\left(\mathbf{0}_{2\times 1},\ \mathbf{D}_{2\times 2}\right)$$

Pero Zhang y Davidian ilustraron, con base en los datos disponibles, que β_{0i} no sigue una distribución normal univariada (ver histograma a continuación). Por lo tanto, el supuesto de la distribución para el vector $\mathbf{b}_i$ no es válido. ¿Qué hacer en estos casos? Una posible solución es utilizar distribuciones flexibles como las denominadas Skew-Normal o Skew-T[3].

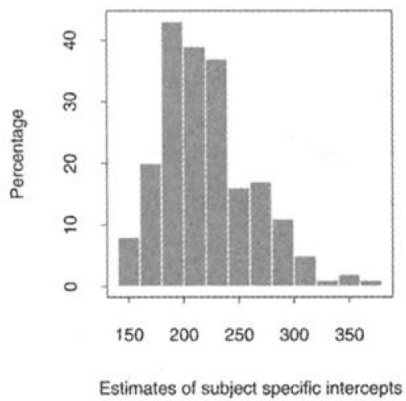

Figura 3: Histograma de los interceptos sujeto-específicos. FHHS

[3] Adelchi Azzalini. The Skew-Normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics*, Vol. 32, No. 2 (Jun., 2005), pp. 159-188

Modelo de intersectos aleatorios revisado. Esta es una de las formulaciones más comunes del LMM. Cada sujeto tiene un perfil que es aproximadamente paralelo a la tendencia de un grupo en particular (recuerde el ejemplo de datos de Sitka Spruce). Este modelo se puede expresar como:

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

donde

- El sujeto se representa por i: $i = 1, 2, \cdots, N$.
- Las observaciones para el sujeto son j: $j = 1, 2, \cdots, m_i$.
- $b_{0i} \sim_{iid} N(0, \sigma_b^2)$. $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$. b_{0i} indep. de ϵ_{ij}

MLM para el modelo de regresión simple

Caso 1: Supongamos que ambos $\mathbf{D}$ y Σ_i se conocen. El MLE de β está dado por:

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

(Solución de mínimos cuadrados generalizados, GLSS). Aquí,

$$\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m)'$$

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m)'$$

$$\mathbf{V} = diag\{\mathbf{V}_i\}$$

y

$$\mathbf{V_i} = \mathbf{Z_i}\mathbf{D}\mathbf{Z_i}' + \Sigma_i$$

Si $\mathbf{V_i}$ no se conoce, se puede usar $\widehat{\mathbf{V_i}}$.

Puesto que $\mathbf{V}$ es una matriz diagonal en bloques, se tiene que

$$\widehat{\beta} = \left(\sum_{i=1}^{m} \mathbf{X_i}' \mathbf{V_i}^{-1} \mathbf{X_i} \right)^{-1} \left(\sum_{i=1}^{m} \mathbf{X_i}' \mathbf{V_i}^{-1} \mathbf{Y_i} \right)$$

También, se tiene que $\widehat{\beta}|\mathbf{b_i}$ es un estimador insesgado para β con matriz de varianzas y covarianzas:

$$Cov\left(\widehat{\beta}|\mathbf{b_i}\right) = \left(\sum_{i=1}^{m} \mathbf{X_i}' \mathbf{V_i}^{-1} \mathbf{X_i} \right)^{-1}$$

Case 2: Asuma que $\mathbf{D}$ y/o Σ_i no se conocen. Para encontrar el MLE de β debemos usar la distribución conjunta de $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m$. En esta situación surgen algunos problemas que serán discutidos más adelante. Ahora discutimos un caso especial que ilustrará algunas de estas complicaciones.

Caso especial: Este caso especial se define para un solo grupo que tiene datos balanceados en el modelo de intersectos aleatorios.

1. Datos balanceados donde $n_i = n$.
2. Errores de medición indep.: $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}_{n \times n}$.
3. El intercepto aleatorio genera solo una componente para la matriz $\mathbf{D} = \sigma_b^2$.

Bajo estas condiciones,

$\mathbf{V_i} = \mathbf{V_o}$ Todos los sujetos tienen la misma matriz de covarianzas.

donde $\mathbf{V_o} = \sigma^2\mathbf{I} + \sigma_b^2\mathbf{J}$ ($\mathbf{J}$ es una $n \times n$ matriz de unos) (es decir, todos los sujetos tienen la misma matriz de covarianza definida por $\mathbf{V_o}$). Cuando σ_b^2 se conoce, el MLE del intercepto y la pendiente está dado por:

$$\widehat{\beta_o} = \bar{Y}_{..} - \widehat{\beta_1}\bar{X}_{..}$$

$$\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}}$$

Se verá que $\mathbf{V_o} = \sigma^2\mathbf{I} + \sigma_b^2\mathbf{J}$ donde se tienen errores independientes: $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$.

Prueba:

$$Var\left(\mathbf{y}_i\right) = \mathbf{Z_i D Z_i}' + \mathbf{\Sigma}_i$$

Si $\mathbf{D} = \sigma_b^2$ se tiene que

$$Var\left(\mathbf{b}_i\right) = \sigma_b^2\mathbf{Z_i Z_i}' + \mathbf{\Sigma}_i = \sigma_b^2\mathbf{Z_i Z_i}' + \sigma^2\mathbf{I}$$

Sea $\mathbf{Z_i} = \mathbf{1}_n$, entonces $\mathbf{Z_i Z_i}' = \mathbf{J}_n$. Así,

$$\mathbf{V_o} = \sigma_b^2\mathbf{J} + \sigma^2\mathbf{I}$$

Para ilustrar, considere los datos de abetos. En este ejemplo, $\mathbf{I}_{4\times 4}$ y $\mathbf{J}_{4\times 4}$, entonces

$$
\begin{aligned}
\mathbf{V_o} &= \sigma^2\mathbf{I} + \sigma_b^2\mathbf{J} \\
&= \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma^2+\sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma^2+\sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma^2+\sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma^2+\sigma_b^2 \end{pmatrix} \\
&= \left(\sigma^2+\sigma_b^2\right) \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix} \quad \text{, donde } \rho = \frac{\sigma_b^2}{\sigma^2+\sigma_b^2} = ICC
\end{aligned}
$$

ICC=Coeficiente de correlación intraclase.[4]

McCulloch y Searle discuten este caso extensamente y muestran que cuando se conoce σ_b^2, el MLE del intersecto y la pendiente comunes son

$$\widehat{\beta_0} = \widehat{Y}_{..} - \widehat{\beta}_1 \widehat{X}_{..}$$

y

$$\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}}$$

Respectivamente.

[4]ICC es la proporción de la variabilidad total atribuible a la varianza dentro de los sujetos. Tiene sentido interpretarlo solo si estamos tratando con LMM con interceptos aleatorios y supuestos de distribuciones normales, pero hay algunas ideas relacionadas para otros modelos.

Propiedades de estos estimadores: Están definidos libres de σ^2 y σ_b^2.

Son iguales al estimador obtenido cuando $\sigma_b^2 = 0$ (uso de los estimadores MCO en ausencia de efecto aleatorio).

La matriz de covarianza de estos estimadores se define en términos de σ^2 y σ_b^2 como sigue:

$$\begin{aligned} Var\left(\widehat{\beta}_0\right) &= \frac{1}{m}\left\{\sigma_b^2 + \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}_{\cdot}^2}{S_{XX}}\right\} \\ Var\left(\widehat{\beta}_1\right) &= \frac{1}{m}\left\{\frac{\sigma^2}{S_{XX}}\right\} \end{aligned}$$

y

$$Cov\left\{\widehat{\beta_0}, \widehat{\beta_1}\right\} = -\bar{X}_{\cdot} Var\left(\widehat{\beta_1}\right)$$

Si σ_b^2 no se conoce, luego, después de considerar la verosimilitud de β_0, β_1 y σ_b^2, encontramos que los MLE de los efectos fijos son iguales a los anteriores (y por lo tanto no dependen de los componentes de varianza). Pero, la matriz de covarianza se define en términos de los componentes de varianza desconocidos. En concreto, los estimadores de los componentes de la varianza son:

$$\widehat{\sigma}^2 = \frac{SSR}{m(n-1)}$$

y

$$\widehat{\sigma}_b^2 = \frac{1}{n}\left(\frac{SSB}{m} - \widehat{\sigma}^2\right)$$

donde,

$$SSR = SSE - m\frac{\widehat{\beta}_1^2}{S_{XX}}$$

$$SSB = n\sum_{i=1}^{m}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$$

$$SSE = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

En esta formulación, asumimos que $\widehat{\sigma}_b^2 > 0$, sin embargo, podría ser que $\widehat{\sigma}_b^2 < 0$ cuando $\widehat{\sigma}^2 > \frac{SSB}{m}$. Para hacer frente a esta dificultad, tomamos

$$\widehat{\sigma}^2 = \frac{SSB + SSR}{mn}$$

y

$$\widehat{\sigma}_b^2 = 0$$

cuando $SSB < \frac{SSR}{(n-1)}$. Si tenemos $\widehat{\sigma}_b^2$ y $\widehat{\sigma}^2$ podemos obtener el Coeficiente de Correlación Intraclase (ICC):

$$\widehat{\rho} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \widehat{\sigma}^2}$$

Este ICC mide la covarianza entre las medidas tomadas sobre el mismo sujeto.

Recuerde el modelo lineal clásico[5]:

$$\begin{aligned} \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times 1} + \mathbf{e}_{n\times 1} \\ E(\mathbf{e}) &= \mathbf{0} \\ Var(\mathbf{e}) &= \sigma^2\mathbf{I} \\ \mathbf{e} &\sim N\left(\mathbf{0},\ \sigma^2\mathbf{I}\right) \end{aligned}$$

Complete lo siguiente:

$$\begin{aligned} E(\mathbf{Y}) &= ? \\ Var(\mathbf{Y}) &= ? \\ \mathbf{Y} &\sim ? \end{aligned}$$

[5]Machiavelli, R. (2014) *Introduccion a los modelos mixtos*. Proceedings X Coloquio de Estadística. Medellín, Colombia.

Solución:

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\beta + \mathbf{e}) = \mathbf{X}\beta \\ Var(\mathbf{Y}) &= Var(\mathbf{e}) = \sigma^2\mathbf{I} \\ \mathbf{Y} &\sim N_n\left(\mathbf{X}\beta,\ \sigma^2\mathbf{I}\right) \end{aligned}$$

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Cuando estamos estudiando la influencia de un factor sobre un factor cuantitativo (respuesta) es común encontrar otras variables o factores que influyen en esa respuesta y deben ser controlados. Estas variables se conocen como bloques. Estos bloques tienen las siguientes propiedades:

(I) Estudiar su efecto no es un objetivo directo, pero aparecen de forma natural en el estudio.

(II) No interactúan con el factor de interés.

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Ejemplo: Diseño en bloques al azar. Un investigador quiere evaluar las necesidades energéticas de una persona cuando camina, come o hace ejercicio. Supongamos que tenemos 10 personas para realizar un experimento y la respuesta de interés son las calorías consumidas por segundo. Los resultados varían de un sujeto a otro. Aquí el factor es la actividad (tres niveles: caminar (W), comer (E) y hacer ejercicio (G)). Si a cada persona se le asigna una actividad diferente, puede ser que la variabilidad observada se deba a diferencias entre sujetos.

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Una posible solución a este problema es asignar cada sujeto a cada actividad. Así, la variable bloque es el sujeto (cada sujeto es un bloque). Cada persona (o bloque) se asigna aleatoriamente a cada actividad:

Persona 1	Persona 2	· · ·	Persona 10
E	G	· · ·	W
W	W	· · ·	G
G	E	· · ·	E

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Un modelo lineal clásico es el llamado diseño de bloques aleatorios[6]. Veamos dos ejemplos:

EXPERIMENTO 1:

1. Elegimos al azar cinco parcelas homogéneas.
2. Dividimos cada parcela en 4 parcelas.
3. En cada parcela asignamos un tratamiento diferente (para un total de cuatro tratamientos). Repetimos el proceso en las demás parcelas. ¿Cuántos tratamientos? ¿Cuántos bloques?

[6]Normalmente, un factor de bloqueo es una fuente de variabilidad que no es de interés principal para el experimentador. Un ejemplo de un factor de bloqueo podría ser el sexo de un paciente; al bloquear por el sexo o género, se controla esta fuente de variabilidad, lo que conduce a una mayor precisión

EXPERIMENTO 2:

1. Elegimos al azar 30 hombres de 25 a 35 años de un grupo de voluntarios.
2. Con estas personas creamos grupos de tamaño 3 cada uno (un total de 10 grupos) según su peso. En otras palabras, creamos grupos con los de peso similar.
3. En el primer grupo de sujetos, asignamos aleatoriamente a cada persona a una de las tres dietas (tratamientos). Repetimos el proceso en los demás grupos. ¿Cuántos tratamientos? ¿Cuántos bloques?

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

El experimento anterior tiene un modelo lineal común (compárelo con el modelo de intersectos aleatorios):

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + b_j + e_{ij} \\ b_j &\sim N(0,\ \sigma_b^2) \\ e_{ij} &\sim N(0,\ \sigma^2) \\ b_j &\text{ indep. of } e_{ij} \end{aligned}$$

Este modelo también podría formularse como

$$\begin{aligned} Y_{ij}|b_j &\text{ i.d } N(\mu + \tau_i + b_j,\ \sigma^2) \\ b_j &\text{ i.d } N(0,\ \sigma_b^2) \end{aligned}$$

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Aquí μ es la gran media o media global (de hecho, es el efecto de la gran media), τ_i se refiere al efecto incremental sobre la media debido al nivel i del tratamiento, b_j es el efecto incremental sobre la media debido al nivel j del bloque.

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Usando estas formulaciones, encuentre lo siguiente:

$$E(Y) =$$
$$E(Y|b_j) =$$
$$Var(Y) =$$
$$Var(Y|b_j) =$$

$$Cov(Y_{1j}, Y_{3j}) =$$
$$Corr(Y_{1j}, Y_{3j}) =$$
$$Cov(Y_{11}, Y_{34}) =$$
$$Corr(Y_{11}, Y_{34}) =$$

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Solución:

$$
\begin{aligned}
E(Y) &= \mu + \tau_i \\
E(Y|b_j) &= \mu + \tau_i + b_j \\
Var(Y) &= \sigma_b^2 + \sigma^2 \\
Var(Y|b_j) &= \sigma^2
\end{aligned}
\qquad
\begin{aligned}
Cov(Y_{1j},\ Y_{3j}) &= \sigma_b^2 \\
Corr(Y_{1j},\ Y_{3j}) &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \\
Cov(Y_{11},\ Y_{34}) &= 0 \\
Corr(Y_{11},\ Y_{34}) &= 0
\end{aligned}
$$

Note que ambos, $Cov(Y_{11},\ Y_{34}) = 0$ y $Corr(Y_{11},\ Y_{34}) = 0$ ya que son la covarianza y correlación entre miembros de diferentes bloques. Por otra parte, $Cov(Y_{1j},\ Y_{3j}) = \sigma_b^2$[7] y $Corr(Y_{1j},\ Y_{3j}) = \frac{\sigma_b^2}{\sigma_b^2+\sigma^2}$ ya que son la covarianza y la correlación dentro de un mismo bloque.

[7] ver prueba detallada en el archivo DEMOSTRACION_CLASE7_INTROMLM.pdf

MODELO LINEAL MIXTO COMO UN MODELO EN BLOQUES AL AZAR

Modelo clásico:

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

Modelo mixto (Diseño en bloques aleatorizados):

$$Y_{ij} = \mu + \tau_i + b_j + e_{ij}$$

b_j y e_{ij} se distribuyen normalmente y en el segundo modelo, son independientes.

¿Cuándo un efecto es fijo o aleatorio?

Si los niveles estudiados son los únicos que nos interesan, el efecto podría considerarse fijo (por ejemplo, el efecto de una dieta específica).

Si los niveles estudiados corresponden a una muestra aleatoria de un conjunto de niveles posibles, el efecto podría considerarse aleatorio (por ejemplo, el efecto de los grupos de peso, random factor).

MODELO LINEAL MIXTO DE INTERCEPTOS Y PENDIENTES ALEATORIAS

Supongamos que tenemos los siguientes 5 perfiles de colesterol de sujetos de Framingham Heart Health Study[8]. A partir de estos perfiles, parece que cada sujeto tiene su propio intersecto y pendiente aleatorios.

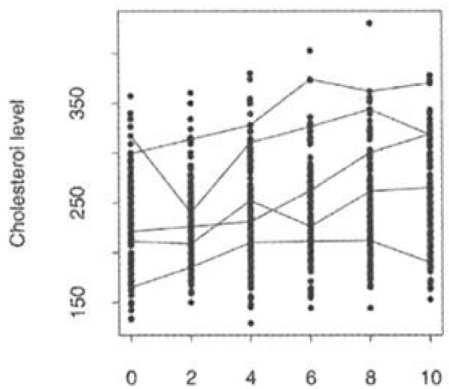

Figura 4: Perfiles de evolución de niveles de colesterol para 5 personas del FHHS

[8]Source: Zhang, D, and Davidian, M. (2001). Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics*, Vol. 57, No. 3 , pp. 795-802

MODELO LINEAL MIXTO DE INTERCEPTOS Y PENDIENTES ALEATORIAS

Por esta razón, podríamos suponer que el vector de efectos aleatorios:

$$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N_2\left(\mathbf{0}_{2\times 1},\ \mathbf{D}_{2\times 2}\right)$$

y formular el siguiente modelo:

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

conocido como Modelo de intersectos y pendientes aleatorias.

MODELO LINEAL MIXTO DE INTERCEPTOS Y PENDIENTES ALEATORIAS

Este es quizás el otro MLM más utilizado. Cada sujeto difiere de los demás en términos tanto del intersecto como de la tendencia (pendiente) del grupo. De nuevo, este 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

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