Diseño de Experimentos. Guía docente. Unidad 10

Introducción a los Modelos Lineales Mixtos con R

Introducción

Los datos agrupados surgen en casi todas las áreas de aplicación estadística. A veces la estructura de agrupación es simple, donde cada caso pertenece a un solo grupo y solo hay un factor de agrupación. Los conjuntos de datos más complejos tienen una estructura jerárquica o anidada o incluyen elementos longitudinales o espaciales. Todos estos datos comparten la característica común de correlación de observaciones dentro del mismo grupo y, por lo tanto, los análisis que asumen independencia de las observaciones serán inapropiadas. El uso de efectos aleatorios es una forma conveniente de modelar dicha estructura de agrupamiento.

Un modelo de efectos mixtos tiene efectos tanto fijos como aleatorios. Un ejemplo simple de tal modelo sería un análisis de la varianza de dos factores (ANOVA):

$$y_{ijk} = \mu + \alpha_i + B_j + \epsilon_{ijk}$$

donde μ y α_i son efectos fijos y el error, ϵ_{ijk} y el factor aleatorio B_j son independientes e idénticamente distribuidos con distribución $N(0, \sigma^2)$ y $N(0, \sigma_B^2)$ respectivamente.

Querríamos estimar α_i y probar la hipótesis $H_0 : \alpha_i = 0, \forall i$ mientras estimaríamos σ_B^2 y probaríamos $H_0 : \sigma_B^2 = 0$. Notar la diferencia: necesitamos estimar y probar varios parámetros de efectos fijos, mientras que solo necesitamos estimar y probar un solo parámetro aleatorio.

Comencemos con el modelo más simple de efectos aleatorios: un diseño ANOVA de un factor con a niveles:

$$y_{ij} = \mu + A_i + \epsilon_{ij} \qquad i = 1, ..., a, \qquad j = 1, ..., n$$

donde A_i es $N(0, \sigma_A^2)$ y ϵ_{ij} es $N(0, \sigma^2)$.

Ejemplo 1

Evaluación del rendimiento de un instrumental especializado con las componentes de la varianza Un fabricante desarrolla un nuevo espectrómetro de uso en laboratorios clínicos. Una componente crítica en el rendimiento de instrumentos es la uniformidad de las mediciones entre los operarios. En este caso específico, el equipo que desarrolló el instrumento deseaba saber si la variabilidad entre operarios estaba dentro de los estándares aceptables para las aplicaciones clínicas.

Se estableció un diseño factorial de tratamientos con `operarios` como factor. Cada operario realizo dos mediciones con cada instrumento.

Se eligieron cuatro operarios especializados al azar de la plantilla. Se prepararon 8 muestras de suero del mismo reactivo, se asignaron al azar 2 muestras a cada operario. La variable respuesta corresponde a los niveles de triglicéridos (mg/dL) en las muestras de suero.

Para empezar, definamos el data.frame con los datos

```
y <- c(148.6, 152.5,
       148.6, 153.1,
       135.5, 140.9,
       152.0, 157.4)
operario <- rep(1:4, each = 2)
df <- data.frame(y, operario)
df$operario <- as.factor(df$operario)
df
```

```
##       y operario
## 1 148.6        1
## 2 152.5        1
## 3 148.6        2
## 4 153.1        2
## 5 135.5        3
## 6 140.9        3
## 7 152.0        4
## 8 157.4        4
```

Los componentes del modelo son,

$$y_{ij} = \mu + A_i + \epsilon_{ij}$$

donde $i = 1, ..., 4$, y $j = 1, 2$, A_i es el efecto operario. Además, A_i es $N(0, \sigma_A^2)$ y ϵ_{ij} es $N(0, \sigma^2)$, todas ellas variables aleatorias independientes. La variable respuesta es por construcción del modelo un variable aleatoria $N(0, \sigma_Y^2)$ donde $\sigma_Y^2 = \sigma_A^2 + \sigma^2$. Asumiremos las suposiciones de la modelización, y pasaremos a resolver los contrastes de interés. Los contrastes a resolver son: $H_0 : \sigma_A^2 = 0$, donde la hipótesis alternativa afirman que la varianza es mayor que cero.

Ajustemos el modelo y obtengamos la tabla anova.

```
modelo <- aov(y ~ operario, data = df)
tAnova<-anova(modelo)
tAnova
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## operario   3 308.46 102.822  8.7713 0.03117 *
## Residuals  4  46.89  11.723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Concluimos que el factor operario es significativo. Es importante, resaltar que a diferencia de los modelos de efectos fijos ahora **NO** haremos comparaciones múltiples, pasaremos a evaluar las componentes de la varianza.

- Varianza del error: $\hat{\sigma}^2 = MS_E = 11.7$
- Varianza del factor operario: $\hat{\sigma}^2_A = \frac{MS_A - MS_E}{n} = \frac{102.8 - 11.7}{2} = 45.55$

La estimación de la varianza de la variable respuesta es $\hat{\sigma}^2_Y = 45.55 + 11.7 = 57.25$.

Este método de estimar los componentes de la varianza se puede utilizar para diseños más complejos. Se construye la tabla ANOVA, se calculan los cuadrados medios esperados y las componentes de la varianza obtenidos al resolver las ecuaciones resultantes. Estos estimadores se conocen como estimadores ANOVA, que son los que hemos visto en este curso hasta este punto.

Históricamente fueron los primeros estimadores desarrollados para componentes de la varianza. Tienen la ventaja de adoptar formas explícitas adecuadas para el cálculo manual, fórmulas que fueron importantes en los días previos a la informática, pero tienen una serie de desventajas:

1. Las estimaciones pueden tomar valores negativos. Por ejemplo, en nuestra situación anterior, ocurrirá si MSA < MSE, esto es conceptualmente imposible ya que las varianzas no pueden tener valor negativo. Se han propuesto varias correcciones, pero todas se alejan de la simplicidad del método de estimación original.
2. Un diseño balanceado tiene el mismo número de observaciones por celda. En tales circunstancias, la descomposición de ANOVA en sumas de cuadrados es única. Para datos no balanceados, este no es cierto y debemos elegir qué descomposición de ANOVA usar, lo cuál, a su vez, afecta la estimación de los componentes de la varianza. Varias reglas han sido sugeridas de cómo se debe realizar la descomposición, pero ninguno de estos métodos es mejor que otros de manera universal.
3. La necesidad de complicados cálculos algebraicos. Las fórmulas para los modelos más simples son fáciles de encontrar en la bibliografia del campo y sencillas de codificar en software. Los modelos más complejos requerirán construcciones opacas.

Idealmente interesa un método que evite estimaciones negativas de varianzas, que funcione sin ambigüedades para datos no balanceados y que se pueda aplicar de manera transparente y sencilla. La estimación de máxima verosimilitud (MLE) satisface estos requisitos. Esto requiere que fijemos alguna distribución para los errores y los efectos aleatorios. La suposición habitual es la normalidad. La máxima verosimilitud es adecuada también para otras distribuciones de probabilidad de las cuales se ocupan los modelos lineales mixtos generalizados (GLMM).

Para un modelo de efectos fijos con errores normales, podemos escribir:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

o, de manera equivalente

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \text{ donde } \boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{R})$$

donde $\boldsymbol{X}$ es la matriz del modelo de orden $n \times p$ y $\boldsymbol{\beta}$ es un vector de longitud p y $\boldsymbol{R}$ denota la matriz de varianzas y covarianzas de $\boldsymbol{y}$. Hasta el momento, hemos tomado $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ (observaciones incorrelacionadas).

Podemos generalizar a un modelo de efectos mixtos con un vector $\boldsymbol{b}$ de q efectos aleatorios con matriz de modelo asociada $\boldsymbol{Z}$ de orden $n \times q$. Entonces podemos modelar la respuesta $\boldsymbol{y}$, dado el valor de la efectos aleatorios como:

$$E(\boldsymbol{y}|\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$$

Si además asumimos que los efectos aleatorios $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D})$ entonces

$var(\boldsymbol{y}) = var(\boldsymbol{Z}\boldsymbol{b}) + var(\boldsymbol{\epsilon}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\intercal + \boldsymbol{R}$ y podemos escribir la distribución incondicional de $\boldsymbol{y}$ como:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\intercal + \boldsymbol{R})$$

a partir de cual observamos que los efectos fijos se involucran en la estimación de la media y los efectos aleatorios en la estimación de la varianza de $\boldsymbol{y}$.

Si conociéramos $\boldsymbol{D}$, entonces podríamos estimar $\boldsymbol{\beta}$ usando mínimos cuadrados generalizados; ver, por ejemplo, Capítulo 6 en Faraway (2004). Sin embargo, la estimación de los componentes de la varianza, $\boldsymbol{D}$, es a menudo uno de los propósitos del análisis. La máxima verosimilitud estándar es un método de estimación que se puede utilizar aquí para encontrar estimaciones de máxima verosimilitud (MLE) de $\boldsymbol{\beta}$, σ^2 y $\boldsymbol{D}$. Esto es sencillo en principio, pero puede haber dificultades en la práctica. Los modelos más complejos que involucran un mayor número de parámetros de efectos aleatorios pueden ser difíciles de estimar. Los errores estándar se pueden obtener utilizando la teoría asintótica de los MLE.

Los MLE tienen algunos inconvenientes. Un problema particular es que pueden estar sesgados. Por ejemplo, consideremos una muestra i.i.d. de datos normales $x_1, ..., x_n$ entonces el MLE es:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Se necesita un denominador de $n-1$ para tener un estimador insesgado. Problemas similares ocurren con la estimación de los componentes de la varianza. Dado que el número de niveles de un factor puede no ser grande, el sesgo del MLE del componente de varianza asociado con ese factor puede ser bastante grande. Los estimadores de máxima verosimilitud restringida (REML) son un intento de solucionar este problema.

La idea es tomar una combinación lineal de la respuesta, $\boldsymbol{k}$, tal que $\boldsymbol{k}^\intercal\boldsymbol{X} = 0$. Entonces tenemos:

$$\boldsymbol{k}^\intercal\boldsymbol{y} \sim N(0, \boldsymbol{k}^\intercal(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\intercal + \boldsymbol{R})\boldsymbol{k})$$

Ahora, podemos proceder a maximizar la verosimilitud basada en $\boldsymbol{k}^\intercal\boldsymbol{y}$ que no implica ninguno de los parámetros de efectos fijos. Una vez estimados los parámetros del efecto aleatorio, es suficientemente directo obtener las estimaciones de los parámetros de efectos fijos. En general, REML produce estimaciones menos sesgadas. Para datos balanceados, las estimaciones de REML suelen ser iguales que las estimaciones de ANOVA.

Mostraremos los estimadores de máxima verosimilitud para los modelos lineales mixtos (LMM) en R. El paquete R original para el ajuste de modelos de efectos mixtos fue `nlme` como se describe en Pinheiro y Bates (2000). Posteriormente, Bates (2005) introdujo una versión mejorada con el paquete `lme4`.

```
mmod<-lmer(y ~ 1+(1|operario), data = df, )
summary(mmod)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ 1 + (1 | operario)
##    Data: df
##
## REML criterion at convergence: 45.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.13407 -0.58222 -0.03033  0.65971  0.99255
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  operario (Intercept) 45.55    6.749
##  Residual             11.72    3.424
## Number of obs: 8, groups:  operario, 4
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  148.575      3.585   3.000   41.44 3.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El modelo tiene componentes de efectos fijos y aleatorios. El efecto fijo aquí es sólo la intercepción representada por el primer `1` en la fórmula del modelo. El efecto aleatorio es representado por `(1|operario)` que indica que los datos están agrupados por operario y el `1` indicando que el efecto aleatorio es constante dentro de cada operario. Los paréntesis son necesarios para garantizar que la expresión se analiza en el orden correcto. El método de ajuste predeterminado es REML. Vemos que esto da estimaciones idénticas al método ANOVA. Para diseños desbalanceados, los estimadores REML y ANOVA no son necesariamente idénticos.

Para extraer las componentes de la varianza disponemos de la función `VarCorr()`.

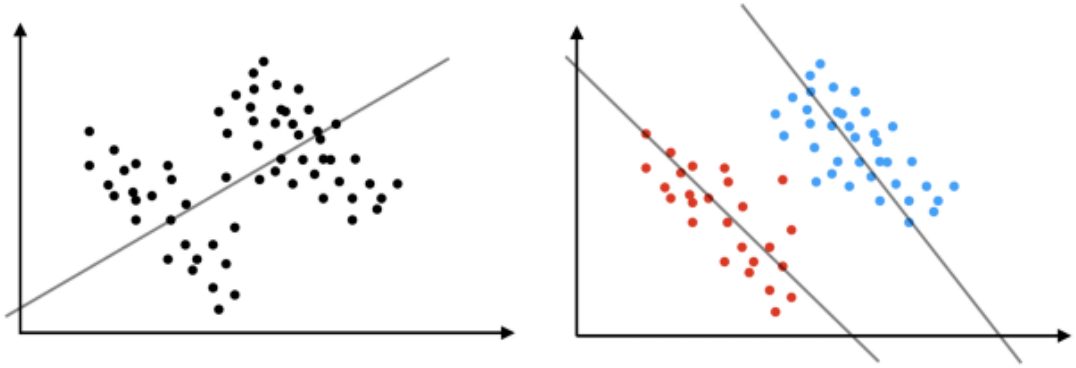
```
VarCorr(mmod)
```

```
##  Groups   Name        Std.Dev.
##  operario (Intercept) 6.7490
##  Residual             3.4238
```

LMM para datos agrupados (clustered data)

La paradoja de Simpson evidencia la necesidad de modelos que tengan en cuenta la agrupación de las observaciones.

Una tendencia o resultado que está presente cuando los datos se colocan en grupos que se invierte o desaparece cuando se combinan los datos.



Veamos la formulación general del modelo lineal mixto para datos agrupados para luego pasar a su aplicación en un caso de ejemplo. El modelo lo formulamos mediante:

$$y_{ij} = \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ij}$$

$i = 1, ..., m$, $j = 1, ..., n_i$, donde

$$\begin{aligned}
y_{ij} &= \text{respuesta del elemento } j\text{-th en el grupo } i, i = 1, ..., m; j = 1, ..., n_i && (1)\\
m &= \text{número de grupos} && (2)\\
n_i &= \text{tamaño del grupo } i && (3)\\
\boldsymbol{x}_{ij} &= \text{vector de covariantes del elemento } j\text{-th del grupo } i \text{ para factores fijos}, \in \mathbb{R}^p && (4)\\
\boldsymbol{\beta} &= \text{vector de parámetros para factores fijos}, \in \mathbb{R}^p && (5)\\
\boldsymbol{z}_{ij} &= \text{vector de covariables del elemento } j\text{-th del grupo } i \text{ para factores aleatorios}, \in \mathbb{R}^q && (6)\\
\boldsymbol{b}_i &= \text{vector de parámetros para efectos aleatorios}, \in \mathbb{R}^q && (7)
\end{aligned}$$

Las suposiciones de la modelización son:

- Los efectos aleatorios $\boldsymbol{b}_i$ son $N(\boldsymbol{0}, \boldsymbol{D})$ siendo $\boldsymbol{D} \in \mathbb{R}^{q\times q}$ la matriz de covarianzas.

- $$\boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}$$

  es $N(\mathbf{0}, \boldsymbol{\Sigma}_i)$, $\boldsymbol{\Sigma}_i \in \mathbb{R}^{n_i \times n_i}$ la matriz de covarianzas del vector error $\boldsymbol{\epsilon}_i$ en el grupo i.

- $\boldsymbol{b}_1, ..., \boldsymbol{b}_m, \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ independientes.

Para las observaciones del grupo i, tenemos la formulación

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} \qquad \boldsymbol{X}_i = \begin{pmatrix} 1 & x_{i11} & ... & x_{i1p} \\ 1 & x_{i21} & ... & x_{i2p} \\ \vdots & \vdots & ... & \vdots \\ 1 & x_{in_i1} & ... & x_{in_ip} \end{pmatrix} \qquad \boldsymbol{Z}_i = \begin{pmatrix} 1 & z_{i11} & ... & z_{i1q} \\ 1 & z_{i21} & ... & z_{i2q} \\ \vdots & \vdots & ... & \vdots \\ 1 & z_{in_i1} & ... & z_{in_iq} \end{pmatrix} \qquad \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}$$

Los vectores de coeficientes de parámetros

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{b}_i = \begin{pmatrix} u_{i0} \\ u_{i1} \\ \vdots \\ u_{iq} \end{pmatrix}$$

Entonces, de manera más compacta, el modelo para las observaciones en el cluster i es:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

donde $\boldsymbol{b}_i$ es $N_q(\mathbf{0}, \boldsymbol{D})$ y $\boldsymbol{\epsilon}_i$ es $N_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. Además, $\boldsymbol{b}_1, ..., \boldsymbol{b}_m, \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ independientes.

Por tanto, el modelo lineal mixto se puede formular por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

Para el total de observaciones el modelo es

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \quad X = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \mathbf{0}_{n_1 \times q} & ... & \mathbf{0}_{n_1 \times q} \\ \mathbf{0}_{n_2 \times q} & \boldsymbol{Z}_2 & ... & \mathbf{0}_{n_2 \times q} \\ \vdots & \vdots & ... & \vdots \\ \mathbf{0}_{n_m \times q} & \mathbf{0}_{n_m \times q} & ... & \boldsymbol{Z}_m \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix}$$

donde $\boldsymbol{b}$ es normal $N_{mq \times mq}(\mathbf{0}, \boldsymbol{G})$, con

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{D} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D} & ... & \mathbf{0} \\ \vdots & \vdots & ... & \vdots \\ \mathbf{0} & \mathbf{0} & ... & \boldsymbol{D} \end{pmatrix}$$

y $\boldsymbol{\epsilon}$ es normal $N_{n\times n}(\mathbf{0}, \boldsymbol{R})$, $n = n_1 + \cdots + n_m$, con

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & ... & \mathbf{0} \\ \vdots & \vdots & ... & \vdots \\ \mathbf{0} & \mathbf{0} & ... & \boldsymbol{\Sigma}_m \end{pmatrix}$$

Ejemplo 2

Se asignaron al azar seis ratones hembra a 3 tratamientos (control, dosis baja y dosis alta). El objetivo del estudio era el de comparar el peso de las crías al nacer según el tratamiento recibido. Se consideró el sexo de la cría como un factor a incluir en el modelo. Se trata de un diseño experimental con dos factores fijos (tratamiento y sexo) y un factor aleatorio (camada) que agrupa las observaciones. No es balanceado dado que el numero de crías por rata no es el mismo, ni tampoco el numero de crías que son hembra o macho.

Tenemos $m = 6$ camadas, para las n_i observaciones de la camada i podemos escribir:

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix} \qquad \boldsymbol{X}_i = \begin{pmatrix} 1 & \boldsymbol{x}_{i1} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{in_i} \end{pmatrix} \qquad \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}$$

donde $\boldsymbol{X}_i$, la matriz de diseño, contiene la primera columna de unos para la intercepción y los vectores $\boldsymbol{x}_{ij}$ $j = 1, ..., n_i$, en filas para indicar los tratamientos y el sexo que determina la condición experimental de la observación correspondiente. Por ejemplo, si consideramos la disposición de niveles de los factores fijos mediante `(intercepcion, control, dosis baja, dosis alta, hembra, macho)`, $(1, 1, 0, 0, 1, 0)$ correspondería a una observación que proviene del control y es hembra, así, el vector $(1, 0, 1, 0, 0, 1)$ correspondería a una observación con dosis baja y es macho.

Los vectores de coeficientes son

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \qquad b_i = \begin{pmatrix} u_i \end{pmatrix}$$

donde $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$, y $b_i \sim N(0, \sigma_b^2)$, siendo $q = 1$.

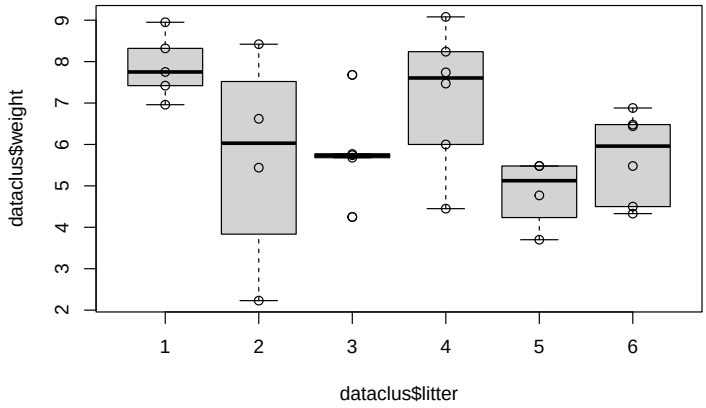
Globalmente,

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 \\ \boldsymbol{Y}_4 \\ \boldsymbol{Y}_5 \\ \boldsymbol{Y}_6 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{Z}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{0} & \boldsymbol{Z}_2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_3 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_3 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_4 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_4 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_5 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_5 & \boldsymbol{0} \\ \boldsymbol{X}_5 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_6 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \boldsymbol{\epsilon}_3 \\ \boldsymbol{\epsilon}_4 \\ \boldsymbol{\epsilon}_5 \\ \boldsymbol{\epsilon}_6 \end{pmatrix}$$

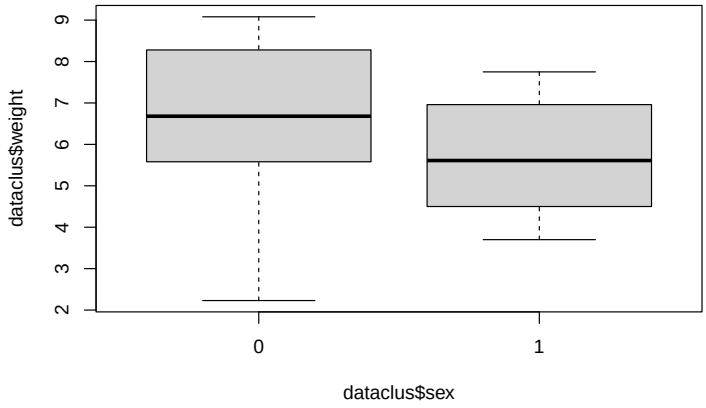
donde $\boldsymbol{Y}_i$ es una matriz $n_i \times 1$ que contiene todas las observaciones de la respuesta para la camada i, $\boldsymbol{X}_i$ son matrices de diseño de los factores fijos $n_i \times 6$ y $\boldsymbol{Z}_i$ son matrices de diseño $n_i \times 1$ de los factores aleatorios y las $\boldsymbol{\epsilon}_i$ es nuevamente una matriz $n_i \times 1$.

Realizamos una breve descriptiva

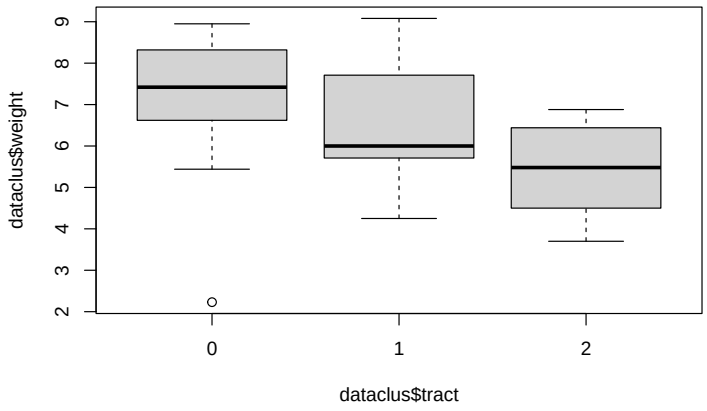
```
plot(dataclus$weight~dataclus$litter)
points(dataclus$weight~dataclus$litter)
```




```
boxplot(dataclus$weight~dataclus$sex)
```



```
boxplot(dataclus$weight~dataclus$tract)
```



A nivel descriptivo, se observa bastante variabilidad entre y dentro de camadas, un ligero efecto del sexo y también un efecto del tratamiento, que sugiere disminución del peso con la dosis. Pasemos a analizar los datos, donde especificamos el efecto aleatorio de la camada atendiendo a que se trata del factor que agrupa las observaciones.

```
mod<-lmer(weight ~ sex+tract+(1|litter), data = dataclus, REML=T)
summary(mod)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: weight ~ sex + tract + (1 | litter)
##    Data: dataclus
##
## REML criterion at convergence: 103.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0809 -0.4536  0.1636  0.6304  1.3314
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  litter   (Intercept) 0.8237   0.9076
##  Residual             1.9681   1.4029
## Number of obs: 30, groups:  litter, 6
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   7.2964     0.8456  3.9507   8.628  0.00105 **
## sex1         -0.8603     0.5212 22.9960  -1.651  0.11241
## tract1       -0.4695     1.1107  2.9454  -0.423  0.70145
## tract2       -1.5675     1.1168  2.9990  -1.404  0.25508
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) sex1   tract1
## sex1   -0.340
## tract1 -0.703  0.087
## tract2 -0.678  0.024  0.512
```

Resultan las estimaciones de la varianza $\sigma_b^2 = 0.8237$ y $\sigma^2 = 1.9681$. La varianza de la camada representa aproximadamente un 30% de la varianza total. En la siguiente tabla Anova se comprueba que **no** resultan significativos los efectos fijos, `sexo` y `tratamiento`.

```
anova(mod)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##       Sum Sq Mean Sq NumDF   DenDF F value Pr(>F)
## sex   5.3620  5.3620     1 22.9960  2.7245 0.1124
## tract 4.1104  2.0552     2  2.9143  1.0443 0.4550
```

Medidas repetidas y datos longitudinales

En un diseño de medidas repetidas se toman más de una medida por cada individuo del estudio. Cuando estas medidas repetidas se toman a lo largo del tiempo, se denomina estudio longitudinal o, en algunas aplicaciones, estudio de panel. En algunos experimentos se registran varias covariables relativas al individuo y el interés se centra en cómo la respuesta depende de las covariables a lo largo del tiempo.

La formulación general del modelo para datos longitudinales o datos agrupados es la misma. El modelo lo formulamos mediante:

$$y_{ij} = \boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i + \epsilon_{ij}$$

$i = 1, ..., m$, $j = 1, ..., n_i$, donde

$$\begin{aligned}
y_{ij} &= \text{respuesta del elemento } i\text{-th en la medida } j, i = 1, ..., m, j = 1, ..., n_i && (8)\\
m &= \text{número de individuos} && (9)\\
n_i &= \text{numero de medidas en el individuo } i && (10)\\
\boldsymbol{x}_{ij} &= \text{vector de covariantes del individuo } i\text{-th en la medida } j, \in \mathbb{R}^p && (11)\\
\boldsymbol{\beta} &= \text{vector de parámetros para factores fijos}, \in \mathbb{R}^p && (12)\\
\boldsymbol{z}_{ij} &= \text{vector de covariables del elemento } i\text{-th en la medida } j \text{ para factores aleatorios}, \in \mathbb{R}^q && (13)\\
\boldsymbol{b}_i &= \text{vector de parametros para efectos aleatorios}, \in \mathbb{R}^q && (14)
\end{aligned}$$

Entonces, de manera más compacta, el modelo para las observaciones en el individuo i es:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

donde $\boldsymbol{b}_i$ es $N_q(\boldsymbol{0}, \boldsymbol{D})$ y $\boldsymbol{\epsilon}_i$ es $N_{n_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$. Además, $\boldsymbol{b}_1, ..., \boldsymbol{b}_m, \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ vectores aleatorios independientes.

Ejemplo 3. Privación de sueño

Como ejemplo de datos longitudinales, consideraremos un estudio de privación del sueño en el que se restringió el tiempo de sueño de 18 personas y se midió una reacción de su organismo en una serie de pruebas durante 10 días. Los datos incluyen tres variables: 1) reacción, 2) días, 3) sujeto, es decir, se siguió a cada individuo durante 10 días.

```
data(sleepstudy)
head(sleepstudy,20)
```

```
##    Reaction Days Subject
## 1  249.5600    0     308
## 2  258.7047    1     308
## 3  250.8006    2     308
## 4  321.4398    3     308
## 5  356.8519    4     308
```

```
## 6  414.6901    5     308
## 7  382.2038    6     308
## 8  290.1486    7     308
## 9  430.5853    8     308
## 10 466.3535    9     308
## 11 222.7339    0     309
## 12 205.2658    1     309
## 13 202.9778    2     309
## 14 204.7070    3     309
## 15 207.7161    4     309
## 16 215.9618    5     309
## 17 213.6303    6     309
## 18 217.7272    7     309
## 19 224.2957    8     309
## 20 237.3142    9     309
```

Para analizar los datos podemos utilizar un modelo mixto con pendientes e intercepciones aleatorias. Dado que sólo tenemos un regresor, este modelo se puede escribir como

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i0} + u_{i1} x_{ij} + \epsilon_{ij}$$

donde y_{ij} denota la j-ésima observación del elemento (individuo) i, y x_{ij} y ϵ_{ij} el predictor y el término de error respectivos. Este modelo se puede expresar en notación matricial de la siguiente manera:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

Supongamos que tenemos m elementos, es decir, $i = 1, ..., m$ y n_i denota el número de observaciones efectuadas en el i-ésimo elemento. Particionado para cada elemento, podemos escribir la fórmula anterior como

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{Z}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{0} & \boldsymbol{Z}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & & \cdots & & \\ \boldsymbol{X}_m & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix}$$

donde $\boldsymbol{Y}_i$ es una matriz $n_i \times 1$ que contiene todas las observaciones del elemento i, $\boldsymbol{X}_i$ y $\boldsymbol{Z}_i$ son matrices de diseño $n_i \times 2$ en este caso y $\boldsymbol{\epsilon}_i$ es nuevamente una matriz $n_i \times 1$.

Escribiéndolas separadamente, tenemos

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} \qquad \boldsymbol{X}_i = \boldsymbol{Z}_i = \begin{pmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix} \qquad \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}$$

Los vectores de coeficientes son

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \boldsymbol{b}_i = \begin{pmatrix} u_{i0} \\ u_{i1} \end{pmatrix}$$

donde $\boldsymbol{b}_i \sim N_2(\boldsymbol{0}, \boldsymbol{D})$, $q = 2$.

Para ver que las dos formulaciones del modelo son de hecho equivalentes, observemos cualquiera de los individuos (escogemos genéricamente el i-ésimo).

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

Aplicando las definiciones anteriores, se puede mostrar que la j-ésima fila del vector resultante es

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i0} + u_{i1}x_{ij} + \epsilon_{ij}$$

donde j varía de 1 a n_i.

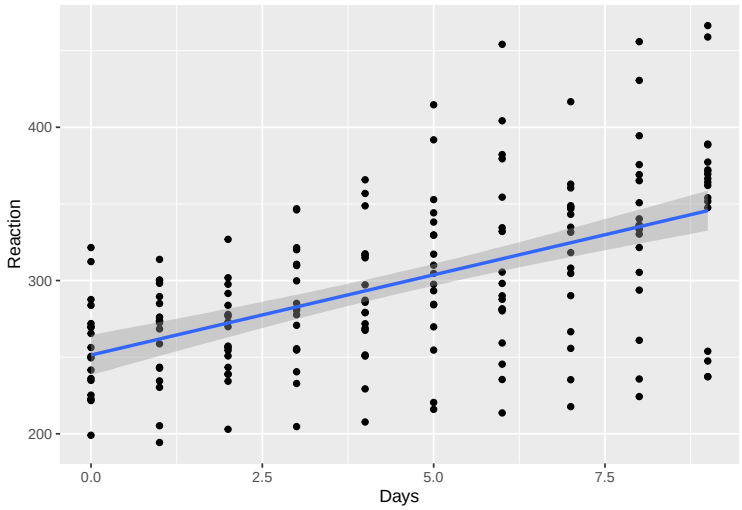
Ejemplo 3 (continuación)

Para observar las limitaciones de los mínimos cuadrados ordinarios (OLS) en datos longitudinales ajustaremos una recta de regresión con `Reaccion` como variable respuesta y `Dias` como variable explicativa con la función `lm` y representaremos el gráfico de dispersión.

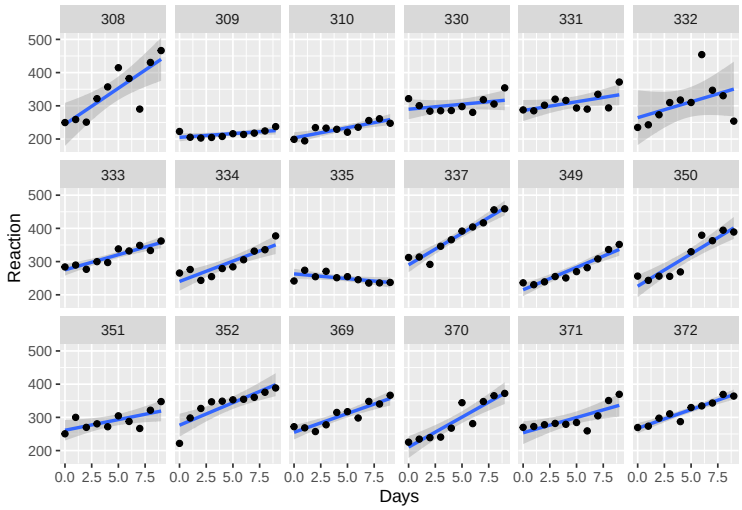
```
summary(lm(Reaction~Days, data = sleepstudy))
```

```
##
## Call:
## lm(formula = Reaction ~ Days, data = sleepstudy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.848  -27.483    1.546   26.142  139.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  251.405      6.610  38.033  < 2e-16 ***
## Days          10.467      1.238   8.454 9.89e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.71 on 178 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2825
## F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

```
ggplot(sleepstudy,aes(x=Days,y=Reaction)) + geom_point() + geom_smooth(method = "lm")
```




```
ggplot(sleepstudy,aes(x=Days,y=Reaction)) + geom_smooth(method = "lm",level = 0.95) +
  geom_point() + facet_wrap(~Subject, nrow = 3, ncol = 6)
```



Podemos ver que la mayoría de los individuos tienen un perfil de reacción creciente, mientras que algunos tienen un perfil neutral o incluso perfil decreciente. ¿No parece extraño que la reacción general aumente mientras que a nivel de individuo las pendientes podrían estar disminuyendo? ¿El ajuste global es realmente lo suficientemente bueno? ¿Capturamos toda la variación en los datos con los mínimos cuadrados ordinarios usuales?

La respuesta es NO porque no hemos tenido en cuenta la falta de independencia entre observaciones. Como veremos más adelante, podemos hacerlo mucho mejor con un modelo lineal mixto (LMM) que tenga en cuenta la no independencia entre las observaciones a través de efectos aleatorios.

Un punto fuerte de LMM es que el ajuste se realiza en todos los individuos simultáneamente en el contexto de cada uno, es decir, todos los ajustes individuales "saben" acerca de cada otro. Por lo tanto, las pendientes, intercepciones e intervalos de confianza de los ajustes individuales se ven afectados por su estadística común, varianza compartida.

A continuación, ajustaremos un LMM con pendientes e intercepciones aleatorias para el efecto de `Dias` para cada `individuo` usando la función `lmer` del paquete `lme4` de R. Esto corresponderá a agregar el término `(Days | Subject)` al al modelo lineal `Reaccion ~ Dias` que se utilizó anteriormente dentro de la función `lm`.

```
summary(lmer(Reaction ~ Days + (Days | Subject), sleepstudy))   # Random intercept and slope
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.10   24.741
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  251.405      6.825  17.000  36.838  < 2e-16 ***
## Days          10.467      1.546  17.000   6.771 3.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

Podemos ver dos tipos de estadísticas: efectos fijos y aleatorios.

La pendiente y los valores de intercepción para efectos fijos se ven bastante similares a los obtenidos anteriormente con OLS.

Por otro lado, las estadísticas de efectos aleatorios es donde se realiza el ajuste por la no independencia entre muestras.

Podemos ver dos tipos de varianza: la que se comparte entre pendientes y intercepciones, `Name = (Intercept` y `Name = Days`, que refleja la agrupación de los puntos de datos por `Subject`, y una varianza residual, aquella que permanece sin modelar.

Además, si comparamos los errores residuales entre modelos de efectos fijos (`lm`) y aleatorios (`lmer`), podemos ver que el error residual disminuyó para el modelo de efectos aleatorios, lo que significa que capturó más variación de la variable respuesta con el modelo de efectos aleatorios.

```
sqrt(sum(residuals(lm(Reaction~Days,data=sleepstudy))^2)/(dim(sleepstudy)[1]-2))
```

```
## [1] 47.71472
```

```
sqrt(sum(resid(lmer(Reaction~Days+(Days|Subject),sleepstudy))^2)/(dim(sleepstudy)[1]-2))
```

```
## [1] 23.56935
```

La misma conclusión se puede extraer de la comparación de los valores AIC y BIC para los dos modelos, nuevamente el LMM con efectos aleatorios simplemente se ajusta mejor a los datos.

```
fit1 <- lm(Reaction ~ Days, data = sleepstudy)
fit2 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy, REML = FALSE)
anova(fit2, fit1)
```

```
## Data: sleepstudy
## Models:
## fit1: Reaction ~ Days
## fit2: Reaction ~ Days + (Days | Subject)
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit1    3 1906.3 1915.9 -950.15   1900.3
## fit2    6 1763.9 1783.1 -875.97   1751.9 148.35  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

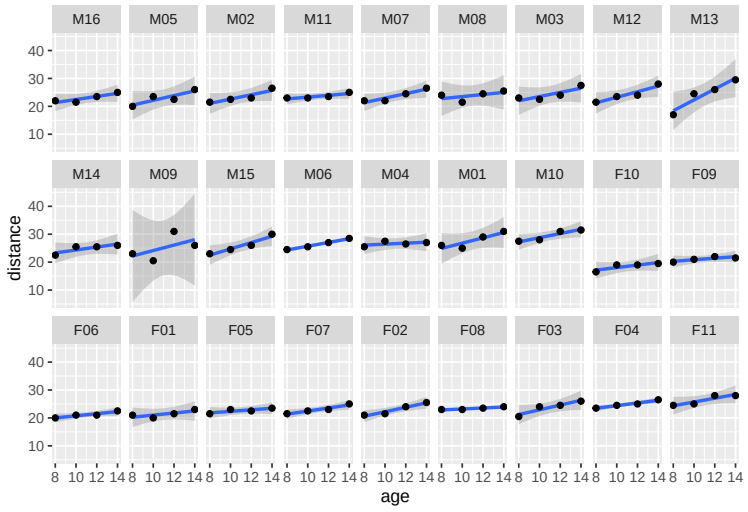
Ejemplo 4

Un ejemplo clásico son los datos, en Potthoff y Roy (1964), es un conjunto de mediciones de la distancia desde la glándula pituitaria hasta la fisura pterigomaxilar tomada cada dos años desde los 8 años hasta los 14 años de edad en una muestra de 27 niños: 16 niños y 11 niñas.

```
data(Orthodont)
head(Orthodont)
```

```
## Grouped Data: distance ~ age | Subject
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

```
ggplot(Orthodont,aes(x=age,y=distance)) + geom_smooth(method = "lm",level = 0.95) +
  geom_point() + facet_wrap(~Subject, nrow = 3, ncol = 9)
```



De la Figura 1.11 parece que existen diferencias cualitativas entre niños y niñas en sus patrones de crecimiento para esta medida. Por ahora es más fácil restringir nuestro modelado a los datos del grupo de niñas.

```
OrthoFem <- Orthodont[ Orthodont$Sex == "Female", ]
```

```
fm1OrthF.lis <- lmList( distance ~ age, data = OrthoFem )
coef( fm1OrthF.lis )
```

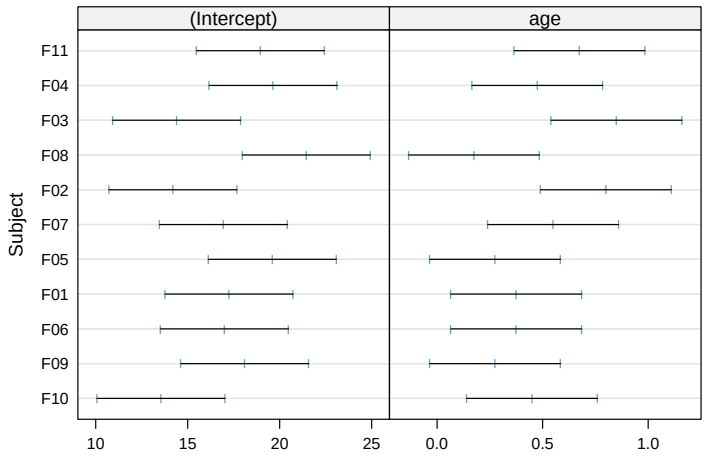
```
##     (Intercept)   age
## F10       13.55 0.450
## F09       18.10 0.275
## F06       17.00 0.375
## F01       17.25 0.375
## F05       19.60 0.275
## F07       16.95 0.550
## F02       14.20 0.800
## F08       21.45 0.175
## F03       14.40 0.850
## F04       19.65 0.475
## F11       18.95 0.675
```

```
intervals( fm1OrthF.lis )
```

```
## , , (Intercept)
##
##        lower  est.    upper
## F10 10.07138 13.55 17.02862
## F09 14.62138 18.10 21.57862
## F06 13.52138 17.00 20.47862
## F01 13.77138 17.25 20.72862
## F05 16.12138 19.60 23.07862
## F07 13.47138 16.95 20.42862
## F02 10.72138 14.20 17.67862
## F08 17.97138 21.45 24.92862
```

```
## F03 10.92138 14.40 17.87862
## F04 16.17138 19.65 23.12862
## F11 15.47138 18.95 22.42862
##
## , , age
##
##           lower  est.     upper
## F10  0.14009962 0.450 0.7599004
## F09 -0.03490038 0.275 0.5849004
## F06  0.06509962 0.375 0.6849004
## F01  0.06509962 0.375 0.6849004
## F05 -0.03490038 0.275 0.5849004
## F07  0.24009962 0.550 0.8599004
## F02  0.49009962 0.800 1.1099004
## F08 -0.13490038 0.175 0.4849004
## F03  0.54009962 0.850 1.1599004
## F04  0.16509962 0.475 0.7849004
## F11  0.36509962 0.675 0.9849004
```

```
plot( intervals ( fm1OrthF.lis ) )
```



Nos damos cuenta que los intervalos para las intercepciones son todos del mismo ancho, al igual que los intervalos para la pendiente con respecto a la edad. Esta es una consecuencia de haber equilibrado datos; es decir, todos los sujetos fueron observados el mismo número de veces y a las mismas edades. También notamos que existe una superposición considerable en el conjunto de intervalos para la pendiente con respecto a la edad. Puede ser factible usar un modelo con una pendiente común.

Lo sorprendente del plot es que no muestra la diferencias sustanciales en las intercepciones que el plot inicia a nivel individual nos llevaría a suponer.

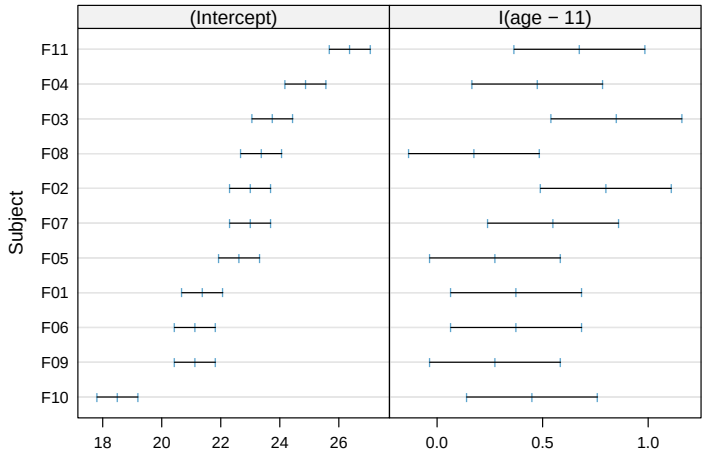
Además, aunque hemos ordenado los grupos, de la distancia media más pequeña (sujeto F10) al que tiene la distancia media más grande (sujeto F11), este orden no se refleja en las intercepciones.

Finalmente, vemos que el patrón entre sujetos en los intervalos para las intercepciones es casi un reflejo del patrón en los intervalos para las pendientes.

Aquellos con experiencia en el análisis de modelos de regresión ya pueden suponer por qué ocurre este reflejo del patrón. Ocurre porque todos los los datos se recopilaron entre los 8 y los 14 años, pero la intercepción representa una distancia a la edad 0. La extrapolación a la edad 0 dará como resultado un alto correlación negativa (alrededor de -0,98) entre las estimaciones de las pendientes y su estimación de la intercept correspondiente.

Eliminaremos esta correlación si centramos los datos. En este caso, nosotros encajaría la distancia como una función lineal de la edad - 11 por lo que los dos coeficientes que se estiman son la distancia a los 11 años y la pendiente o velocidad de crecimiento. Si ajustamos este modelo revisado y representamos gráficamente los intervalos de confianza entonces estos intervalos muestran la tendencia esperada en el intercepción, que ahora representa la distancia ajustada a los 11 años.

```
fm1OrthF.lis <- lmList( distance ~ age, data = OrthoFem )
fm2OrthF.lis <- update(fm1OrthF.lis, distance ~ I( age - 11 ) )
plot( intervals( fm2OrthF.lis ) )
```



Inicialmente, ajustaremos un modelo de efectos mixtos con sólo **intercepciones aleatorias**.

```
fm1OrthF <- lmer( distance ~ age + (1 | Subject) , data = OrthoFem, REML = T)
summary( fm1OrthF )
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: distance ~ age + (1 | Subject)
##    Data: OrthoFem
##
## REML criterion at convergence: 141.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2736 -0.7090  0.1728  0.4122  1.6325
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 4.2786   2.068
##  Residual             0.6085   0.780
## Number of obs: 44, groups:  Subject, 11
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 17.37273    0.85874 27.57177  20.230  < 2e-16 ***
## age          0.47955    0.05259 32.00000   9.119 2.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.674
```

También podríamos ajustar un modelo con la fórmula distancia `~ I (age - 11)` pero, debido al requisito de una pendiente común, las propiedades del modelo centrado son esencialmente equivalentes a las de un modelo no centrado.

Para obtener las estimaciones de máxima verosimilitud, haremos

```
fm1OrthFM <- lmer( distance ~ age + (1 | Subject) , data = OrthoFem, REML = F)
summary( fm1OrthFM )
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: distance ~ age + (1 | Subject)
##    Data: OrthoFem
##
##      AIC      BIC   logLik deviance df.resid
##    146.0    153.2    -69.0    138.0       40
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3056 -0.7192  0.1764  0.4258  1.6689
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 3.88     1.9699
##  Residual             0.59     0.7681
## Number of obs: 44, groups:  Subject, 11
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 17.37273    0.83107 31.01612   20.90  < 2e-16 ***
## age          0.47955    0.05179 33.00000    9.26 1.07e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.685
```

Hemos asumido una pendiente común para todos los sujetos. Para probar esto, podemos ajustar un modelo con efectos aleatorios para ambos - **la intercepción y la pendiente** mediante MLE y comparamos los modelos con las estimaciones MLE.

```
fm2OrthF <-lmer( distance ~ age + (age | Subject) , data = OrthoFem, REML = F)
anova( fm1OrthF, fm2OrthF )
```

```
## Data: OrthoFem
## Models:
## fm1OrthF: distance ~ age + (1 | Subject)
## fm2OrthF: distance ~ age + (age | Subject)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fm1OrthF    4 146.03 153.17 -69.015   138.03
## fm2OrthF    6 146.51 157.21 -67.255   134.51 3.5211  2     0.1719
```

Dado que el p-valor para el segundo modelo frente al primero es de aproximadamente el 15%, concluimos que el modelo más simple es adecuado (el de sólo intercepciones aleatorias)

El modelo con intercepciones aleatorias que se está ajustando se expresaría en notación matricial como

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, ..., 11$$
$$\boldsymbol{b}_i \sim N(0, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{I})$$

con las matrices

$$X_1 = ... = X_{11} = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 11 \\ 1 & 14 \end{pmatrix}, Z_1 = ... = Z_{11} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

El vector bidimensional de efectos fijos $\boldsymbol{\beta}$ consta del intercept, β_1, para la población y la pendiente común, β_2. Los efectos aleatorios, b_i, $i = 1, ..., 11$, son para cada sujeto y describen un cambio en la intercepcion. Debido a que existe una pendiente común, estos cambios en la intercepcion se conservan para todos los valores de edad. La matriz $\boldsymbol{D} = \sigma_b^2$ será una matriz unidimensional 1×1 en este caso. Representa la varianza de las medidas en la población a un valor fijo de edad. Las estimaciones de REML para los parámetros son

$$\beta_1 = 17.37273,\ \beta_2 = 0.47955,\ \sigma_b^2 = 4.2786,\ \sigma^2 = 0.6085$$

La derivación de los valores predichos para la respuesta y para los efectos aleatorios en el modelo lineal de efectos mixtos la podemos extraer mediante la función `random.effects` que nos proporcionan las mejores predicciones lineales insesgadas de los efectos aleatorios (BLUP).

```
random.effects(fm1OrthF)
```

```
## $Subject
##     (Intercept)
## F10 -4.00532866
## F09 -1.47044943
## F06 -1.47044943
## F01 -1.22903236
## F05 -0.02194701
## F07  0.34017860
## F02  0.34017860
## F08  0.70230420
## F03  1.06442981
## F04  2.15080662
## F11  3.59930905
##
## with conditional variances for "Subject"
```

La función `coefficients` (o su forma más corta `coef`) se utiliza para extraer los coeficientes de las rectas ajustadas para cada sujeto. Para el modelo ajustado fm1OrthF el intercept de la recta ajustada para el sujeto i es $\beta_1 + b_i$ y la pendiente es β_2.

```
coef(fm1OrthF)
```

```
## $Subject
##     (Intercept)       age
## F10    13.36740 0.4795455
## F09    15.90228 0.4795455
## F06    15.90228 0.4795455
## F01    16.14369 0.4795455
## F05    17.35078 0.4795455
## F07    17.71291 0.4795455
## F02    17.71291 0.4795455
## F08    18.07503 0.4795455
## F03    18.43716 0.4795455
## F04    19.52353 0.4795455
## F11    20.97204 0.4795455
##
## attr(,"class")
## [1] "coef.mer"
```

Referencias

- Pinheiro y Bates (2000). Mixed-Effects Models in S and S-PLUS.
- McCullogh and Searle (2001) Generalized, Linear, and Mixed Models.
- J.J Faraway (2004)
- J. J Faraway. Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models. Capítulos 8 y 9. Chapman & Hall/CRC. 2006.
- Bates (2005) Fitting Linear Mixed-Effects Models Using lme4.
- West B et al. (2007) Linear Mixed Models: a practical guide using statistical software.