Introducción a los modelos mixtos

María Durbán

Departamento de Estadítica, Universidad Carlos III de Madrid

Índice general

Capítulo 1

Conceptos básicos

Los modelos mixtos (MMs) para variables de respuesta continua son modelos estadísticos en los que los residuos están normalmente distribuidos pero puede que no sean independientes o no tengan varianza constante. Este tipo de datos aparecen en muchas situaciones, sobre todo en experimentos donde se realiza algún tipo de muestreo: 1) estudios con datos agrupados, como por ejemplo, alumnos en una clase, individuos en una ciudad, 2) estudios longitudinales o de medidas repetidas, donde un individuo es medido repetidamente a lo largo del tiempo o bajo condiciones distintas. Este tipo de diseños se pueden encontrar en diferentes áreas como la Medicina, Biología, Ciencias Experimentales y Sociales.

1. Tipo y estructura de los datos

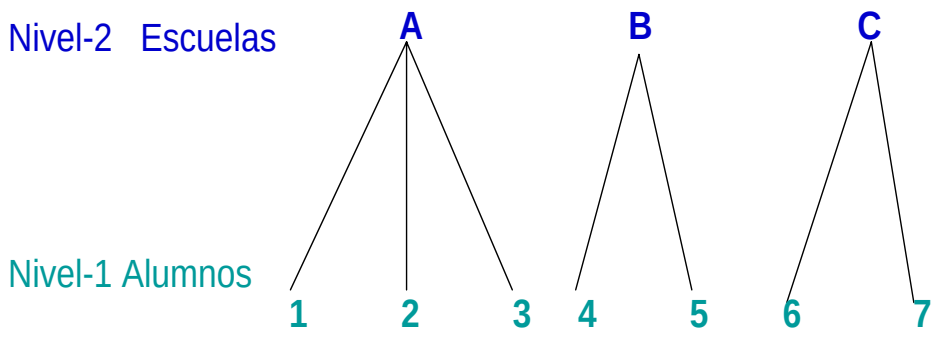
La estructura de los datos con la que estamos trabajando es el factor determinante para saber si hemos de utilizar modelos mixtos, y en su caso, qué tipo de modelo.

1.1. Datos jerárquicos (o agrupados)

En este tipo de datos la variable dependiente se mide una sóla vez en cada individuo (la unidad de análisis), y los individuos está agrupados en (o anidados) en unidades mayores. Muchos tipos de datos tienen una estructura jerárquica:

- Alumnos en escuelas
- Personas en distritos
- Pacientes en hospitales
- Plantas en una parcela

Las jerarquías son una forma de representar la relación de dependencia que hay entre los individuos y los grupos a los que pertenecen (Goldstein, 2002). Por ejemplo, supongamos que hacemos un estudio sobre el rendimiento escolar de alumnos en distintas escuelas, tendríamos una estructura a dos niveles: muchos individuos al nivel 1 (alumnos) que está agrupados (o anidados) en unas pocas unidades de nivel 2 (escuelas).



Las estructuras multinivel pueden aparecer también como consecuencia del diseño del estudio que estamos llevando a cabo. Por ejemplo, una encuesta sobre el estado de salud puede dar lugar a un diseño a tres niveles: primero muestreamos regiones, luego distritos y después individuos.

En cada nivel de la jerarquía podemos medir variables. Algunas estarán medidas en su nivel *"natural"*, por ejemplo en el nivel de la escuela podríamos medir el tamaño, y al nivel de los alumnos podríamos medir su situación socio-económica.
Además, podemos mover las variables de un nivel a otro mediante agregación o desagregación:

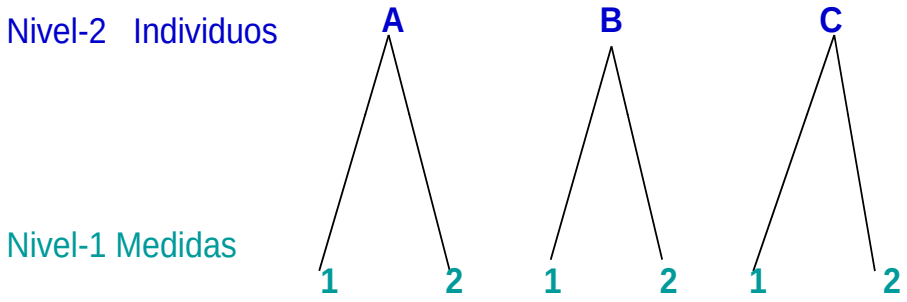
1. **Agregación:** La variable al nivel más bajo se mueve a un nivel más alto, por ejemplo, podemos asociar a cada escuela la media del nivel socioeconómico de sus alumnos.

2. **Desagregación:** Mover las variables a un nivel más bajo, por ejemplo, asignarle a cada alumno una variable que indique el tamaño de la escuela a la que pertenece.

1.2. Medidas repetidas y datos longitudinales

En este tipo de datos la variable dependiente se mide más de una vez a un mismo individuo (Singer et al., 2003). Por ejemplo, medimos los niveles de glucosa de un enfermo antes y después de haberle inyectado insulina. Este tipo de datos también puede ser considerados como datos multinivel (o jerárquicos) donde el Nivel 2 representa a los individuos y el Nivel 1 representa a las diferentes medidas tomadas. Dado que las medidas se toman a un mismo individuo, es probable que dichas medidas no sean independientes, por lo que utilizar un modelo lineal ordinario no sería apropiado.

Por **datos longitudinales**, entendemos datos en los que la variable dependiente se ha medido en distintos instantes de tiempo en cada una de las unidades de análisis. En algunos casos, cuando la variable dependiente se mide a lo largo del tiempo, puede ser difícil identificar si los datos son medidas repetidas o datos longitudinales. Desde el punto de vista del análisis



de los datos mediante MMs esta distinción no es un elemento crítico. Lo importante es que en ambos tipos de datos la variable dependiente se ha medido repetidas veces en la misma unidad de análisis, y que por tanto las observaciones estarán correlacionadas.

2. ¿Efectos fijos o aleatorios?

En un modelo mixto la clave se encuentra en la distinción entre efectos fijos y aleatorios (Snijers, 2003). Esto es importante porque la inferencia y el análisis de efectos fijos y aleatorios es distinta.

Los **efectos fijos** son variables en las cuales el investigador ha incluído sólo los niveles (o tratamientos) que son de su interés. Por ejemplo, en un experimento podemos estar interesados en comparar dos grupos, uno al que se le aplica un tratamiento y otro de control. En este caso, el objetivo del estudio compara los grupos y no estamos interesados en generalizar los resultados a otros tratamientos que podrían haber sido incluídos. Otro ejemplo sería el caso en el que hacemos un encuesta y elegimos 10 ciudades. Si sólo estamos interesados en los resultados para esas 10 ciudades y no queremos generalizar los resultados al resto de ciudades que podrían haber sido seleccionadas, la variable *ciudad* será un efecto fijo. Si elegimos las ciudades de forma aleatoria de una población grande de ciudades consideraríamos la variables *ciudad* como un **efecto aleatorio**.

Una cantidad se considera alatoria cuando cambia sobre las unidades de una población. Cuando un efecto en un modelo estadístico es considerado aleatorio, estamos asumiendo que queremos extraer conclusiones sobre la población de la cual se han elegido las unidades observadas, y no tenemos interés en esas unidades en particular. En este contexto se habla de *"intercambiabilidad"*, en el sentido de que podríamos cambiar una unidad de la muestra por otra de la población y nos sería indiferente. Este es el caso de los factores de agrupamiento o diseño, como son los bloques en un experimento agrícola, o los días cuando un experimento se lleva a cabo en días distintos, o un técnico de laboratorio cuando hay varios haciendo el

experimento; también lo serían los sujetos en un diseño de medidas repetidas o las localizaciones donde se recogen muestras en un río, si el objetivo es generalizar a todo el río.

Los métodos estándar utilizados para construir tests e intervalos de confianza para los efectos fijos, no son válidos para los efectos aleatorios, ya que los efectos observados son sólo una muestra de todos los posibles efectos.

La clave para distinguir, estadísticamente hablando, entre efectos fijos y aleatorios es si los niveles de la variable se pueden interpretar como extraídos de una población con una cierta distribución de probabilidad. En el caso de un efecto fijo estaremos, normalmente, interesados en comparar los resultados de la variable dependiente para los distintos niveles del la variable explicativa, es decir, estaremos interesados en la diferencia entre las medias. En el caso de efectos aleatorios, no estamos interesados específicamente en comparar si las medias son distintas, sino en cómo el efecto aleatorio explica la variabilidad en la variable dependiente. Por lo tanto, para que un efecto pueda considerarse aleatorio, es necesario que la variable dependiente presente cierta variabilidad no explicada asociada con las unidades del efecto aleatorio. Por ejemplo, en un estudio sobre satisfacción en el trabajo (variable dependiente) de los empleados (unidades observadas) de un cierto número de empresas (efecto aleatorio), si el nivel de satisfacción de los empleados de unas empresas es mayor que el de otras y el investigador no lo tiene en cuenta, habrá una cierta variabilidad residual asociada con en efecto *empresa*. Si esta variabilidad fuera próxima a cero, no sería necesario incluir el efecto aleatorio asociado con la empresa.

2.1. ¿Por qué hay que utilizar modelos mixtos?

Cuando las observaciones están agrupadas en niveles o siguen una cierta jerarquía, las unidades se ven afectados por el grupo al que pertenecen. Las jerarquías (o niveles) nos permiten representar la relación de dependencia entre los individuos y los grupos a los que pertenecen. Los alumnos que están en una misma escuela se parecen más entre sí que si los hubiéramos selecionado aleatoriamente de entre toda la población de alumnos. Los modelos mixtos nos permiten tener en cuenta que las observaciones no son independientes.

El hecho de tener variables medidas en distintos niveles hizo que hasta la aparición de los modelos mixtos, se analizaran los datos a un solo nivel, mediante agregación o desagregación de las variables, y utilizando modelos de regresión múltiple. Sin embargo, esto es inadecuado, y hacerlo de este modo, ignorando los distintos niveles, da lugar a problemas desde dos puntos de vista:

1. Estadístico: Si agregamos los datos, combinamos muchas observaciones para dar lugar a unas pocas, y como resultado perdemos información. Por el contrario, si desagragamos los datos, estos son tratados como si fueran observaciones independientes, lo que hace que los errores estándar sean menores de lo que en realidad son, y por tanto, consideraríamos significativas algunas variables que no lo son. En el caso de modelos con medidas repetidas, ignorar la correlación entre las medidas obtenidas en un mismo individuo afectaría al cálculo de los errores estándar. En general, el deseo de generalizar los resultados de la muestra a los de la población (es decir si consideramos efectos

aleatorios) hace que los intervalos de confianza sean más anchos ya que hay más fuentes de variabilidad.

2. Conceptual: Si no tenemos cuidado a la hora de interpretar los resultados cometeríamos un error conocido como *"falacia del nivel equivocado"* (Dansereau et al., 2006), y que consiste en analizar los datos a un nivel e interpretarlos al otro nivel:

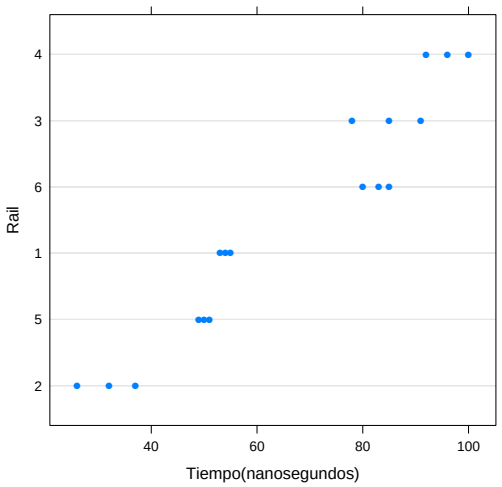
   - *Falacia ecológica* (Jargowsky, 2005): Establecer la relación entre la variable respuesta y una característica del nivel superior y atribuirle esta relación a los individuos del nivel más bajo cuando la relación no se ha establecido a ese nivel. Por ejemplo, supongamos que la renta per cápita en España sea superior a la renta per cápita en Albania. Dar por supuesto que cualquier español elegido al azar tendrá una renta mayor que cualquier albanés elegido al azar es un ejemplo de falacia ecológica, ya que la renta per cápita es un promedio y con ese solo dato no sabemos cual es la distribución de la renta entre los individuos en cada país. Para entenderlo mejor: un caso extremo sería que un solo individuo español tuviera de renta 1.000.000 euros y el resto de los españoles 1 euro cada uno, mientras que todos los albaneses tienen una renta de 2 euros. Cualquiera de los albaneses tiene una renta superior a todos los españoles excepto a uno, en cambio la renta per cápita española sería más alta. Consecuentemente, conociendo la media, que es una característica del grupo, no podemos inferir características de los individuos.
   - *Falacia atomista*(Subramanian et al., 2009): Consiste en atribuir la relación al nivel superior basándose en una análisis llevado a cabo a un nivel inferior. Por ejemplo, en un estudio con individuos puede observarse que el mayor renta individual se asocia a una menor mortalidad por cardiopatía coronaria. Si se infiere de estos datos que a escala de país el mayor ingreso per cápita se asocia con la reducción de la mortalidad por cardiopatía coronaria, el investigador quizá esté incurriendo en una falacia atomística (porque entre países, los mayores ingresos per cápita pueden, en realidad, asociarse con una mayor mortalidad por cardiopatía coronaria).

2.2. Ejemplo

Vamos a trabajar con datos en los que se pretende testar el estrés que sufren los railes, para ello se mide el tiempo que tarda en recorrerlo un cierto tipo de onda (Pinheiro and Bates, 2000). Se seleccionaron 6 railes y se testaron 3 veces cada uno (los datos se encuentran en la librería `nlme`). Los ingenieros estaban interesados en testar en este experimento el tiempo medio de recorrido *"típico"* de los railes (tiempo esperado), la variación del tiempo medio entre railes (variabilidad entre-railes), la variabilidad entre los tiempos observados de un mismo raíl (variabilidad dentro del raíl)

Claramente los datos están agrupados (o en clúster) por raíl. Esta agrupación tiene dos implicaciones:

- Es de esperar que las observaciones hechas sobre un mismo raíl se parezcan más entre sí que a las observaciones de otros railes.



- Es de esperar que el tiempo medio varíe de un raíl a otro, además de variar de una medida a la otra.

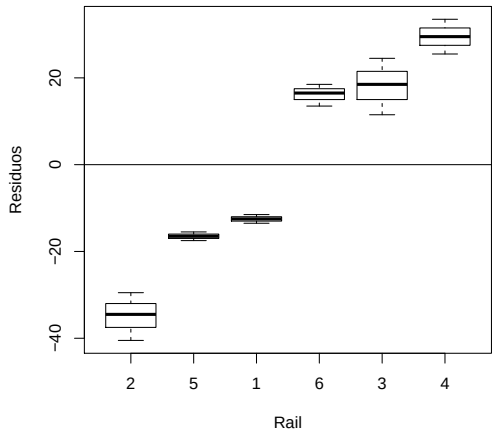
En la figura anterior podemos apreciar que hay bastante variabilidad en los tiempos medios entre los diferentes railes, y que esta variabidad es mayor que la variabilidad dentro de un mismo raíl. Estos datos se pueden analizar con un modelo de efectos fijos o de efectos aleatorios, la elección dependerá de si queremos hacer inferencia sobre los railes específicos que se usaron en el experimento, o si queremos hacer inferecia sobre toda la población de railes de la cual éstos fueros elegidos. Es evidente que que el factor de agrupamiento por raíl debería ser incorporado como un efecto aleatorio, pero para mostrar la importancia de esto tiene empezaremos por ignorar la estructura de agrupación y nos centraremos en el primer objetivo del experimento que se centraba en el tiempo medio:

$$y_{ij} = \mu + \epsilon_{ij}, \quad i = 1, \ldots, 6, \quad j = 1, \ldots, 3 \quad \epsilon ij \sim N(0, \sigma^2) \tag{1.1}$$

donde y_{ij} es el tiempo de recorrido la j-ésima vez en el raíl i-ésimo, y μ es el tiempo medio de recorrido que queremos estimar. Sabemos que su estimador máximo verosímil (ML) es la media muestral, $\overline{y}_{\cdot\cdot} = 66{,}5$, y que el error cuadrático medio (MSE) es un estimador de la varianza, $\hat{\sigma}^2 = s^2 = 23{,}645^2$.

```
rail1=lm(travel~1, data=Rail)

summary(rail1)

lm(formula = travel ~ 1, data = Rail)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.500      5.573   11.93  1.1e-09 ***

Residual standard error: 23.65 on 17 degrees of freedom
```

Si hacemos un gráfico de los residuos de este modelo para cada uno de los railes queda evidente que no es el modelo adecuado:



Al ignorar en el modelo el efecto *Rail*, este aparece en los residuos, de modo que el siguiente paso sería incluirlo en el modelo mediante un Análisis de la Varianza (ANOVA), es decir, vamos a permitir la media de cada raíl quede representada por un parámetro (este sería un modelo de efectos fijos):

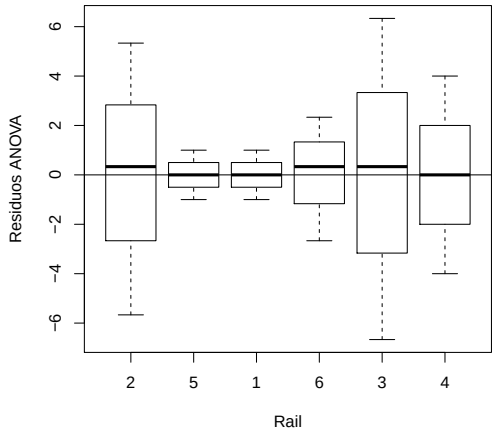
$$y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \epsilon_{ij}, \quad , i = 1, \ldots, 6, \quad j = 1, \ldots, 3 \quad \epsilon ij \sim N(0, \sigma^2), \tag{1.2}$$

donde μ es la media común a todos los railes y α_i es lo que diferencia (al alza o a la baja) a la media de cada raíl de la media global (y la media de cada rail es $\mu_i = \mu + \alpha_i$):

```
rail2=lm(travel~Rail-1, data=Rail)

summary(rail2)

Coefficients:
Rail2  Rail5  Rail1  Rail6  Rail3  Rail4
31.67  50.00  54.00  82.67  84.67  96.00

Residual standard error: 4.021 on 12 degrees of freedom
```

Si hacemos nuevamente un gráfico de los residuos, vemos que los residuos están centrados en cero pero aún hay varios problemas:

- El modelo sólo es útil para los railes específicos usados en el experimento, mientras que el interés del mismo está en la población de railes de la cual se eligieron los que se han usado.
- El modelo no da una estimación de la variabilidad entre railes (que era de interés en el experimento).



- El número de parámetros incrementa al incrementar el número de railes.

Estos problemas se pueden solucionar con un modelo de efectos aleatorios, este modelo trataría el efecto del raíl como variaciones aleatorias alrededor de la media poblacional. El modelo (1.2) puede reescribirse somo:

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

el modelo de efectos aleatorios reemplazaría los parámetros fijos $\mu_i - \mu = \alpha_i$ por un efecto aleatorio u_i que es una variable aleatoria específica para el i-ésimo raíl, con media cero y varianza desconocida σ_u^2, y que representa las desviación de la media del raíl i respecto de la media poblacional (se llaman efectos ya que representan desviaciones). El modelo sería:

$$y_{ij} = \mu + u_i + \epsilon_{ij}, \quad b_i \sim N(0, \sigma_u^2) \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{1.3}$$

además, es normal asumir que los u_i's son independientes entre sí, y de los ϵ_{ij}'s.

Es importante caer en la cuenta del cambio de interpretación de μ, antes era el tiempo de recorrido medio de los 6 railes incluidos en el experimento, y ahora es el tiempo medio de la población de railes de la cual se han elegido estos 6. Además, ahora no estimamos la media de cada raíl, μ_i (ya que no es de interés), sino que estimamos la media poblacional, μ y la varianza entre los railes en la población σ_u^2. Ésta mide la heretogeneidad entre los railes, la cual es consecuencia de tener las observaciones agrupadas por raíl. Dado que no estimamos μ_i, el número de parámetros no aumenta con el número de railes, es decir, independientemente del número de railes, los parámetros a estimar son: μ, σ^2 and σ_u^2.

En el modelo (1.2), las observaciones y_{ij} eran independientes, ya que los ϵ_{ij} lo eran. Ahora, sin embargo, algunas de las observaciones tienen un efecto aleatorio común (todas las que pertenecen al mismo raíl), y por lo tanto están correladas:

$$
\begin{aligned}
Var(y_{ij}) &= Var(u_i) + Var(\epsilon_{ij}) + 2\underbrace{Cov(u_i, \epsilon_{ij})}_{0} = \sigma_u^2 + \sigma^2 \\
Cov(y_{ij}, y_{ik}) &= \sigma_u^2 \\
Cov(y_{ij}, y_{lk}) &= 0 \\
Corr(y_{ij}, y_{ik}) &= \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}
\end{aligned}
$$

Entonces la matrix de varianzas-covarianzas de las observaciones que pertenecen al mismo raíl sería:

$$
Var(y_i) = \begin{pmatrix} \sigma_u^2 + \sigma^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma^2 + \sigma_u^2 \end{pmatrix}, \tag{1.4}
$$

mientras que las observaciones de diferentes railes son independientes. Esta matrix de varianzas se llama de **simetría compuesta**, a ρ se le llama **coeficiente de correlación intra-clase** y a σ^2 y σ_u^2 **componentes de la varianza**.

Capítulo 2

Formulación del modelo mixto lineal

El nombre *modelos mixtos lineales* viene del hecho de que estos modelos son lineales en los parámetros, y en las covariables, y pueden implicar efectos fijos o aleatorios. Son, por lo tanto, una extensión de los modelos lineales de regresión.

La formulación general de un modelo mixto tiene la siguiente forma:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\epsilon} \sim (\boldsymbol{0}, \boldsymbol{R}) \tag{2.1}$$

donde $\boldsymbol{X}$ es una matrix $n \times k$ (k es el número de efectos fijos), $\boldsymbol{Z}$ es una matrix $n \times p$ (p es el número de efectos aleatorios), y $\boldsymbol{G}$ es la matriz de varianzas-covarianzas de los efectos aleatorios, con dimensión $p \times p$.

Recordemos que en el ejemplo anterior el modelo era:

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

donde

y_{ij}	es el tiempo de recorrido de la j-ésima medida en el raíl i-ésimo
μ	es la media global de toda la población de railes
u_i	es el efecto aleatorio del i-ésimo raíl
ϵij	es el término de error

El modelo para los datos del i-ésimo raíl sería:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} u_i + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix}$$

o equivalentemente:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i \quad \boldsymbol{u}_i \sim N(0, \sigma_u^2)$$

y el modelo para todas las observaciones sería el modelo (2.1), donde:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_6 \end{pmatrix} \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_6 \end{pmatrix} \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}) \quad \boldsymbol{G} = \sigma_u^2 \boldsymbol{I}_6 \quad \boldsymbol{R} = \sigma^2 \boldsymbol{I}_{18}$$

1. Estimación en modelos mixtos

1.1. Estimación de los efectos fijos y predicción de efectos aleatorios

Estimación de $\boldsymbol{\beta}$

Una forma de obtener un estimador de $\boldsymbol{\beta}$ utilizar el **modelo marginal**, es decir, reescribir (2.1) como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \text{ donde } \boldsymbol{\epsilon}^* = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}.$$

Este es un modelo con errores correlados, ya que:

$$Cov(\boldsymbol{\epsilon}^*) = \boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}'\boldsymbol{G}\boldsymbol{Z}.$$

Dada la matrix $\boldsymbol{V}$, el estimador de $\boldsymbol{\beta}$ se obtiene mediante *Mínimos Cuadrados Generalizados*, minimizando la siguiente función:

$$\boldsymbol{Q} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

obtieniendo:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{2.2}$$

que además se corresponde con el estimador máximo verosímil (recordad que el método de máxima verosimilitid busca los valores de los parámetros bajo los cuales los datos observados son los más probables).

Predicción de $\boldsymbol{u}$

En el caso de los efectos aleatorios hablamos de predicción y no de estimación, ya que los efectos aleatorios no son parámetros, sino variables aleatorias.

Hay varias maneras de obtener predictores de $\boldsymbol{u}$ que tengan la propiedad de ser el mejor predictor lineal insesgado (mejor en el sentido de tener menor error cuadrático medio de predicción), una de ellas es mediante lo que se llaman *ecuaciones de modelos mixtos de Henderson*. Es un método que nos permite obtender el mejor estimardor lineal insesgado de $\boldsymbol{X}\boldsymbol{\beta}$ y el mejor predictor lineal insesgado de $\boldsymbol{u}$. Se obtiene maximizando la densidad conjunta de $\boldsymbol{y}$ y $\boldsymbol{u}$:

$$f(\boldsymbol{y}, \boldsymbol{u}) = f(\boldsymbol{y}|\boldsymbol{u})f(\boldsymbol{u}), \quad \boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{R}) \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$$

en térmimos de la verosimilitud tendríamos:

$$l \propto -\frac{1}{2}\left[log|\boldsymbol{R}| + log|\boldsymbol{G}| + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u}\right],$$

derivando con respecto a $\boldsymbol{\beta}$ y $\boldsymbol{u}$ obtenemos las siguientes ecuaciones:

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}.$$

Las soluciones a estas equaciones son:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{2.3}$$

$$\hat{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{2.4}$$

donde $\boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$.

Pero, $\boldsymbol{V}$ depende de los parámetros de la varianza en el modelo que forman parte de $\boldsymbol{G}$ y $\boldsymbol{R}$; a continuación mostramos comos estimar dichos parámetros.

1.2. Estimación de los componentes de la varianza

Los métodos más comunes para la estimación de los parámetros de las matrices de covarianza son: Máxima verosimilitud (MV) o Maxima verosimilitid restringida (REML) (Searle, 1992).

Máxima verosimilitud

Como hemos visto anteriormente, $Cov(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$. Entonces el estimador MV de los parámetros de $\boldsymbol{V}$ está basado en el modelo:

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V}),$$

de modo que:

$$l(\boldsymbol{\beta}, \boldsymbol{V}) \propto \frac{1}{2}\left[(log(|\boldsymbol{V}|) + (\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}).\right] \tag{2.5}$$

El estimador MV de β coincide con (2.3). Sustituyendo esta expresión en (2.5), obtenemos el perfil de verosimilitud para $\boldsymbol{V}$:

$$l_P(\boldsymbol{V}) = \frac{1}{2}\left[(log|\boldsymbol{V}| + \boldsymbol{y}'\boldsymbol{V}^{-1}\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\boldsymbol{y}\right]. \tag{2.6}$$

Los estimadores MV de los parámetrios de $\boldsymbol{V}$ se obtienen maximizando esa función. Por ejemplo, en el caso de los datos de railes

$$\boldsymbol{V} = \sigma_u^2\boldsymbol{Z}'\boldsymbol{Z} + \sigma^2\boldsymbol{I},$$

de modo que (2.6) es una función de los parámetros (σ_u^2, σ^2). No existe una solución cerrada para los estimadores que resultan de maximizar (2.6), y se hace de forma numérica.

Máxima verosimilitud restringida (REML)

Hay que recordar que en un modelo lineal clásico, el estimador máximo verosímil de σ^2, $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})}{n}$ es sesgado, y usábamos $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})}{n-k}$ (donde $k = dim(\boldsymbol{\beta})$). Ahora, la idea es similar, y lo que hacemos es estimar los componentes de la varianza basándonos, no en $\boldsymbol{y}$, sino en los residuos obtenidos después de estimar los efectos fijos, es decir $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Esto significa que el método utilizado maximiza la **verosimilitud marginal o restringida** (Patterson and Thompson, 1971):

$$l_R(\boldsymbol{V}) = l_P(\boldsymbol{V}) - \frac{1}{2} log|(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|.$$

La principal ventaja de del REML sobre MV, es que REML tiene en cuenta los grados de libertad utilizados para estimar los efectos fijos en el modelo. Si el tamaño de la muestra con la que estamos trabajando es pequeño, REML dará mejores estimaciones que MV, en el caso de tamaño nuestral grande, no habra prácticamente ninguna diferencia.

En el ejemplo de los railes:

- Modelo efectos fijos: $\hat{\mu} = 66{,}5$, $\hat{\sigma} = 4{,}02$
- Modelo efectos aleatorios: $\hat{\mu} = 66{,}5$, $\hat{\sigma} = 4{,}02$ $\hat{\sigma}_u = 24{,}8 (22{,}62 MV)$

Esto nos dice que la variabilidad entre railes es mucho mayor que dentro de los railes. El hecho de que en ambos modelos En general $\hat{\sigma}$ no va a ser igual en un modelo sin effectos aleatorios que con efectos aleatorios, aquí lo es porque tenemos un diseño balanceado, es decir, que hay el mismo número de observaciones en todos los railes.

2. Contrastes de hipótesis

2.1. Contrastes de hipótesis para $\boldsymbol{\beta}$

Los métodos estándar para realizar contrastes de hipótesis sobre $\boldsymbol{\beta}$ son el **Wald test** y el **test de la razón de verosimilitud**.

Wald test

Es posible probar que la distribución approximada (para muestras grandes) del estimador (restringido) máximo verosímil es:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \underbrace{(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}}_{Var(\hat{\boldsymbol{\beta}})}\right) \tag{2.7}$$

de modo que si

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = 0 \quad \boldsymbol{L}\boldsymbol{\beta} \neq 0$$

el estadístico de contraste sería:

$$T_W = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \boldsymbol{L}' \left(\boldsymbol{L} (\boldsymbol{X}' \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{L}' \right)^{-1} \boldsymbol{L} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \tag{2.8}$$

que asintóticamente sique una distribución ji-cuadrado con grados de libertad igual al rango de $\boldsymbol{L}$. Importante: la varianza de $\hat{\boldsymbol{\beta}}$ se estima incerporando el estimadot (RE)ML de los componentes de la varianza en (2.7).

El estadístico del Wald test está basado en errores estándar que subestiman la verdadera variabilidad de $\hat{\boldsymbol{\beta}}$, ya que ignoran la variabilidad introducida al estimar los componentes de la varianza. Este problema se puede aliviar utilizando un **test t** (para hipótesis sobre parámetros individuales), o um **test F** para hipótesis globales como la anterior:

- **Test t**:

  $$T = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t(v)$$

  los grados de libertad de t, v, se estima a partir de los datos. Un método frecuente es la aproximación tipo Satterthwaite (Satterthwaite, 1941).

- **Test F**:

  $$F = \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \boldsymbol{L}' \left(\boldsymbol{L} (\boldsymbol{X}' \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{L}' \right)^{-1} \boldsymbol{L} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{k} \sim F(k, v)$$

  donde $k = Rango(\boldsymbol{L})$.

Tests de razón de verosimilitud

Como en el caso de modelos lineares de regresión, se basan en la comparación de los valores de la función de verosimilitud para los dos modelos a los que dan lugar las hipótesis que se están comparando.

Dada una hipótesis nula H_0 y una alternativa H_1 anidadas, el test de razón de verosimilitud se define,

$$LRT = -2 \left[log(l_{H_0}) - log(l_{H_1}) \right] \approx \chi^2_{df}$$

Donde *df* corresponde a la diferencia en el número de parámetros bajo H_1 y H_0. Si LRT es suficientemente grande, habrá suficiente evidencia para rechazar la hipótesis nula, y por lo tanto concluir que el modelo más complejo es más adecuado. Este test sólo se puede utilizar si el método de estimación es ML, ya que REML no depende de los parámetros correspondientes a los efectos fijos.

2.2. Contrastes de hipótesis para los parámetros de varianza

En este caso, utilizamos el test de la razón de verosimilitid, pero la distribución asintótica del estadístico del test depende de si el valor del paramámetro bajo la hipótesis nula está en la frontera del espacio paramétrico o no:

- *Caso 1*: El valor de los parámetros de varianza bajo la hipótesis nula no están en la frontera del espacio paramétrico (por ejemplo, si queremos contrastar que los parámetros de varianza de dos efectos aleatorios son iguales o no). En ese caso utilizamos el test normalmente.

- *Caso 2*: El valor de los parámetros de varianza bajo la hipótesis nula están en la frontera del espacio paramétrico (por ejemplo, si queremos contrastar si un efecto aleatorio es necesario, estaríamos contrastando si la varianza del efecto alearorio es cero o no). En este caso el la distribución asintótica del estadístico del test es una mixtura entre una χ^2_p y χ^2_{p-1}, donde p es el número de parámetros de la varianza que se hacen cero bajo la hipótesis nula (Self and Liang, 1987)

2.3. Otras consideraciones

Criterios de información

Busca el modelo que mejor ajusta los datos. Se basan en la verosimilitud, pero penalizando por el número de parámetros en el modelo. La ventaja de este tipo de criterios es que permiten comparar modelos que no están anidados. El más utilizado es el **criterio de información de Akaike**:

$$AIC = -2 \times l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}) + 2p,$$

donde p es el número total de parámetros estimados. Cuanto más pequeño mejor es el modelo (de ahí que el término $2p$ hace la función de penalización).

Diagnosis del modelo

En el caso de modelos mixtos hemos de verificar la hipótesis de normalidad tanto para los residuos al nivel más bajo (medidas repetidas en los railes) como para los efectos aleatorios (railes), y también las de independencia (en su caso).

En el caso de los modelos mixtos, utilizamos los residuos condicionales y su versión *studentizada*. Los **residous condicionales** son la diferencia entre los valores observados y el valor predicho condicional:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{u}},$$

pero estos residuos no son adecuados en muchos casos, ya que tiende a estar correlados y sus varianzas pueden cambiar de un grupo a otro, aunque en el verdadero modelo los resiuod seanincorrelados y con varianza constante. Para solucionar este problema se pueden escalar los residuos por sus desviaciones estándar (o las estimaciones de éstas), dando lugar a los **residuos estandarizados** (si las desviaciones estandar son conocidas), o a los **residuos studentizados** (si son desconocidas y utilizamos estimaciones de las mismas). Con estos residuos haría un análisis similar al caso de modelos de regresión lineal.

3. Funciones de R para ajustar modelos mixtos

Hay varios paquetes de R para el ajuste de modelos mixtos. Vamos a centrarnos en dos nlme y lme4. El segundo es una versión del primero que incluye modelos más generales, y mejora los gráficos. Sin embargo, el primero es el más usado. a continuación describimos las funciones para el ajuste de modelos mixtos con ambos paquetes.

3.1. La función lme

Esta función permite el uso de efectos aleatorios anidados y de errores correlados o heterocedásticos dentro de los grupos. En general para definir un modelomixto necesitamos especificar la estructura de la media y de la parte aleatoria del modelo, incluídos los factores de agrupamiento, asi como la estructura de correlación (si la hay).

La forma general de esta función es :

```
lme(respuesta~predictores,random= , correlation= , method=)
```

La forma de introducir la estructura de los efectos aleatoria será particular para cada caso. Si hay un sólo efecto aleatorio :

```
lme(y~x,random=~1|g)
```

donde g es la variable que define el efecto aleatorio (por ejemplo, Rail). Para especificar un efecto aleatorio para alguno de los efectos fijos:

```
lme(y~x,random=~x|g)
lme(y~x,random=~1+x|g)
```

Si hay más de un efecto aleatorio, se introducen dentro de una lista:

```
lme(y~x,random=list(~1|g1, ~1|g2))
```

Cuando los efectos aleatorios están anidados:

```
lme(y~x,random=~1|g1/g2)
```

Esto significa que g2 está dentro de g1.

También podemis especificar el mét odo de estimación: “REML” o “ML”. En el ejemplo de los railes, el modelo (1.3) se ajustaría del siguiente modo:

```
library(nlme)

rail3=lme(travel~1, random=~1 |Rail ,data=Rail)
```

En la parte aleatoria, | separa las variables de agrupamiento de las predictoras, en este caso no había variables predictoras, por lo tanto aparece un 1.

```
rail3
  Log-restricted-likelihood: -61.0885
  Fixed: travel ~ 1
(Intercept)
       66.5
Random effects:
 Formula: ~1 | Rail
        (Intercept) Residual
StdDev:    24.80547 4.020779
```

La siguiente tabla muestra cómo extraer la información del modelo ajustado. Supongamos que el modelo ajustado se encunetra en un objeto llamado `modelo`, y

```
summ=summary(modelo)
```

Método de estimación	`modelo$method`
$\hat{\boldsymbol{\beta}}$	fixef(modelo)
$\hat{\boldsymbol{\beta}}$ se($\hat{\boldsymbol{\beta}}$) t-test	`summ$tTable`
$\widehat{Var}(\hat{\boldsymbol{\beta}})$	`vcov(model)`
I.C. 95 % para $\boldsymbol{\beta}$	`intervals(modelo, which=''fixed'')`
$\hat{\sigma}$	`summ$sigma`
I.C. 95 % para σ, σ_u	`intervals(modelo, which=''var-cov''`
$\hat{\boldsymbol{u}}$	`ranef(modelo)`
$\beta_0 + \boldsymbol{u}$	`coef(modelo)`
$\hat{\boldsymbol{G}}$	`getVarCov(modelo)`
$\hat{\boldsymbol{G}}$ y $\hat{\sigma}$	`VarCorr(modelo)`
$\hat{\boldsymbol{R}}$	`getVarCoc(modelo,type=''conditional'')`
$\hat{\boldsymbol{V}}$	`getVarCoc(modelo,type=''marginal'')`
Valor de ML	`logLik(modelo,REML=FALSE)`
Valor de REML	`logLik(modelo,REML=TRUE)`
AIC	`AIC(modelo)`
BIC	`BIC(modelo)`
Valores ajustados:	`fitted(modelo)`
Residuos:	
– Condicionales	`resid(modelo,type=''response'')`
– Marginales	`resid(modelo, type=''response'',level=0)`
Residuos Normalizados	`resid(modelo,type=''normalized'')`
Residuos de Pearson	`resid(modelo, type=''pearson'')`
Valores predichos:	
– Condicionales	`predict(modelo, newdata)`
– Marginales	`predict(modelo, newdata,level=0)`

Hemos visto que la función `VarCorr()` da información sobre la estructura de componentes de varianza:

```
VarCorr(rail3)
Rail = pdLogChol(1)
            Variance  StdDev
(Intercept) 615.31111 24.805465
Residual     16.16667  4.020779
```

A partir de esta información podríamos calcular el coeficiente de correlación intra-clase:

$$ICC = \frac{615{,}31}{615{,}31 + 16{,}16} = 0{,}974$$

$$\Downarrow$$

el 97,4 % de la variabilidad total proviene de la heterogeneidad entre los railes.

Si ajustamos el modelo mediante maxima verosimilitud:

```
rail3.1=lme(travel~1, random=~1 |Rail ,data=Rail,method="ML")
VarCorr(rail3.1)
Rail = pdLogChol(1)
            Variance  StdDev
(Intercept) 511.86111 22.624348
Residual     16.16667  4.020779
```

¿Por qué la σ_u^2 es menor y σ^2 se mantiene igual?.

3.2. La función `lmer`

Es la función del paquete `lme4` para ajustar los modelos mixtos. En este caso hay una única fórmula en el modelo, y los efectos aleatorios se incluyen entre paréntesis:

```
lmer(y~x+(1|g))
lme(y~x,random=~1|g)
```

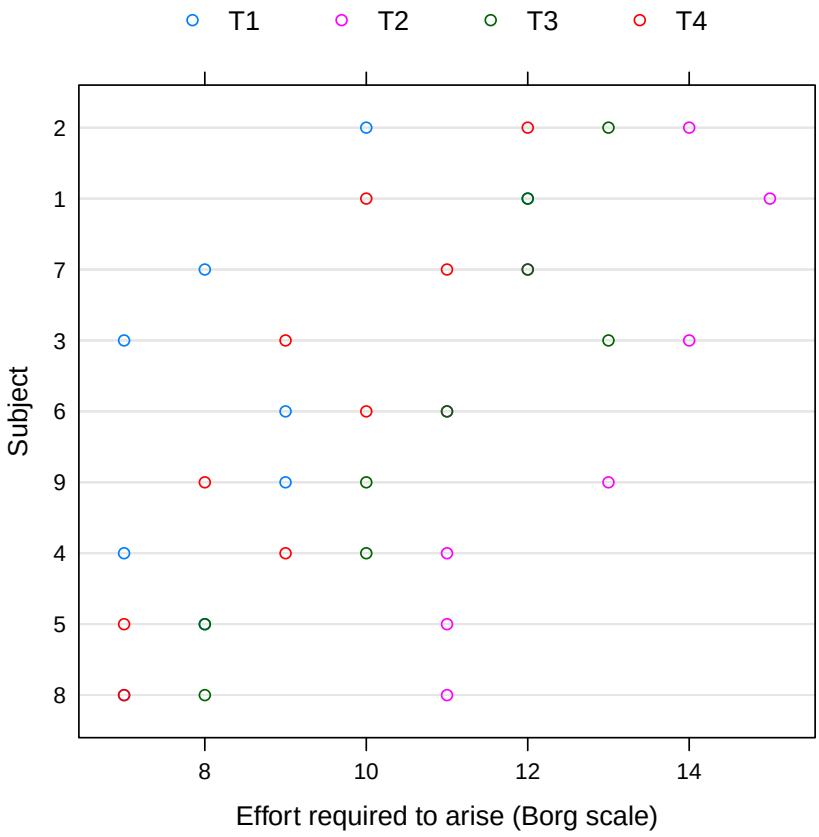
Esta función es más sencilla de usar, pero es limitada en cuanto a que no permite cualquier tipo de estructura para la matriz de varianzas de efectos aleatorios $\boldsymbol{G}$.

En el ejemplo de los railes:

```
rail.lmer=lmer(travel~1+(1|Rail), data=Rail)
   AIC   BIC logLik deviance REMLdev
 128.2 130.8 -61.09    128.6   122.2
Random effects:
 Groups   Name        Variance Std.Dev.
 Rail     (Intercept) 615.311  24.8055
 Residual              16.167   4.0208

Fixed effects:
            Estimate Std. Error t value
(Intercept)    66.50      10.17   6.539
```

3.3. Ejemplo: Diseño completamente aleatorizado por bloques (RCBD)

En el ejemplo de los railes los datos estaban agrupados por raíl, y todas las observaciones se tomaban bajo el mismo tratamiento. En muchas ocasiones, hay más de un tratamiento de interés. En un diseño aleatorizado por bloques cada tratamiento se observa en cada uno de los bloques. Los datos con los que vamos a trabajar provienen de un experimento cuyo objetivo era comparar el diseño ergonómico de diferentes banquetas. Se le pidió a 9 individuos que se sentaran y levantaran de 4 banquetas con distinto diseño, y se midió el esfuerzo (scala de Borg) realizado para levantarse.



En este caso, la variable agrupadora es `Subject`, y todos los tratamientos (tipo de banqueta) se observan en cada bloque. El modelo sería:

$$y_{ij} = \mu_i + u_i + \epsilon_{ij} \quad i = 1, \ldots, 4 \quad j = 1, \ldots, 9.$$

donde μ_j es el esfuerzo medio al levantarse de la banqueta j, y u_i es el efecto aleatorio correspondiente al i-ésimo individuo (es un efecto aleatorio ya que queremos generalizar los resultados a la población a la cual pertenecen estos 9 individuos).

```
stool.lme=lme(effort~Type-1, random=~1|Subject,data=ergoStool)
  Fixed: effort ~ Type - 1
   TypeT1    TypeT2    TypeT3    TypeT4
 8.555556 12.444444 10.777778  9.222222

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.332465 1.100295

VarCorr(stool.lme)
Subject = pdLogChol(1)
            Variance StdDev
(Intercept) 1.775463 1.332465
```

```
Residual    1.210648 1.100295
```

Para elegir la estructura correcta de los efectos aleatorios, hemos de incluir todos los efectos fijos y sus interacciones, y luego comparar los distintos modelos en los que varían los efectos aleatorios, pero los efectos fijos se mantienen. En este caso nos plateamos sólo dos modelos, uno con el efecto aleatorio y otro sin él, y utilizamos el test de la razón de verosimilitud:

```
stool.NULL=lm(effort~Type,data=ergoStool)
test=-2*logLik(stool.NULL, REML=T) +2*logLik(stool.lme, REML=T)
mean(pchisq(test,df=c(0,1),lower.tail=F))
0.000120735
```

Otra opción es usar la AIC:

```
AIC(stool.NULL)
147.3064
AIC(stool.lme)
133.1308
```

Vemos que el modelo de efectos aleatorios el el mejor por ambos criterios.

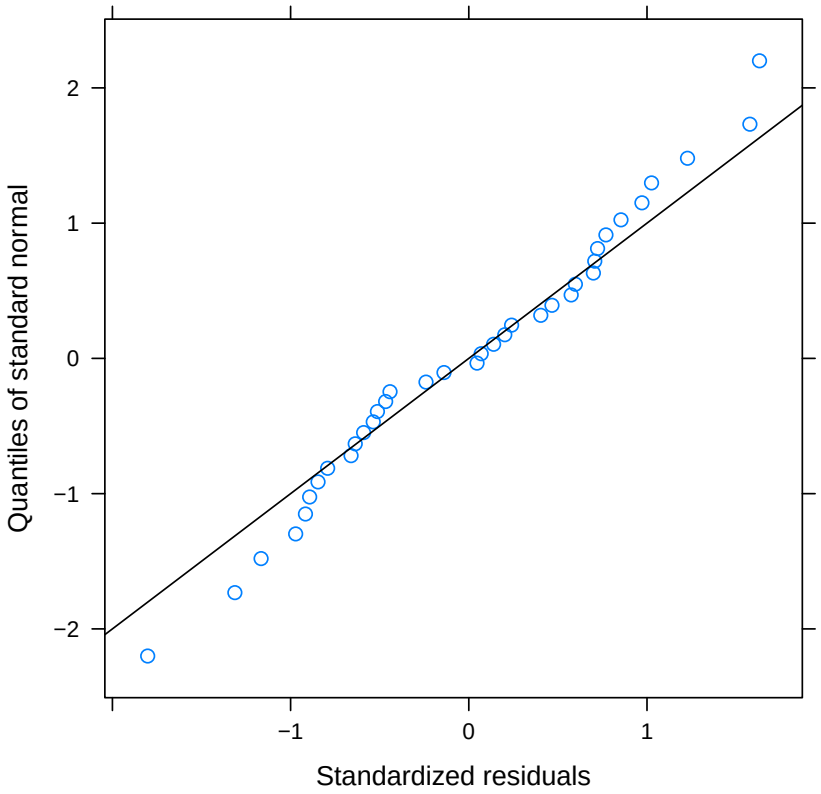
Una ver elegida la estructura de efectos aleatorios, hemos de comprobar la significación de las variables fijas. Para ello comparamos modelos anidados mediante el test de la razón de verosimilitud, basado en el ajuste mediante el método "ML":

```
stool.lme1=lme(effort~Type, random=~1|Subject,data=ergoStool,method="ML")
stool.lme2=lme(effort~1, random=~1|Subject,data=ergoStool,method="ML")
anova(stool.lme2,stool.lme1)
           Model df      AIC      BIC    logLik   Test L.Ratio p-value
stool.lme2     1  3 164.1500 168.9006 -79.07502
stool.lme1     2  6 134.1444 143.6455 -61.07222 1 vs 2 36.0056  <.0001
```

Por lo tanto, el tipo de banqueta es significativo.

Una vez elegida la parte fija y aleatoria del modelo hemos de comprobar que se cumplen las hipótesis del modelo: homocedasticidad, normalidad, etc:

```
plot(stool.lme)
qqnorm(stool.lme, abline=c(0,1))
```



Podemos usar el test de Shapiro-Wilk para contrastar la hipótesis de normalidad más formalmente:

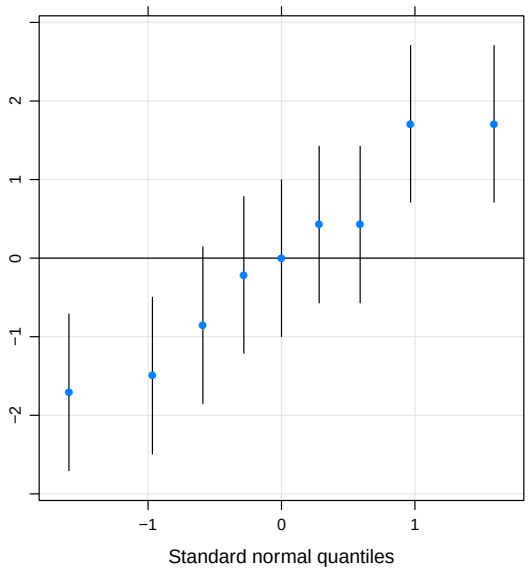
```
shapiro.test(resid(stool.lme))
        Shapiro-Wilk normality test
data:  resid(stool.lme)
W = 0.9778, p-value = 0.6716
```

Podemos ajustar el modelo con la función `lmer`:

```
stool.lmer=lmer(effort~Type-1+(1|Subject), data=ergoStool)
```

Esta función nos permite hacer gráficos para contrastar la normalidad de los efectos aleatorios:

```
library(lattice)
qqmath(ranef(stool.lmer, postVar = TRUE), strip = FALSE)$Subject
```

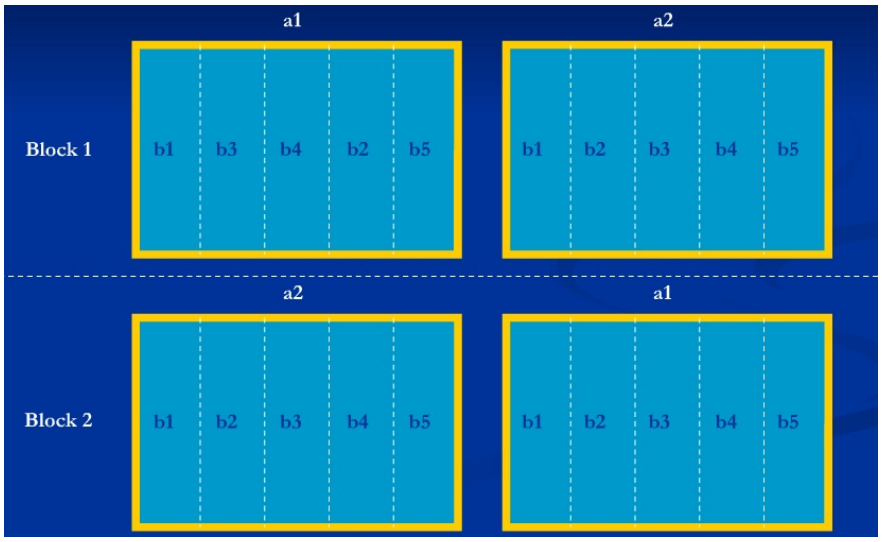


Ejercicio: Diseño incompleto por bloques pracialmete balanceado (PBIBD)

Los datos corresponden al peso de semillas de algodón recogidas en 15 bloques donde se aplican 15 tratamientos (4 tratamientos por bloque). Los datos se encuentra en el data.frame `PBIB` de la librería `SASmixed`. Ajusta un modelo de efectos aleatorios, comprueba si los tratamientos son significativos y si se cumplen las hipótesis del modelo. ¿Cuál es el porcentaje de variabilidad total explicada por el efecto de bloque?.

3.4. Ejemplo: Diseño split-splot

En general, diremos que un factor **B** está anidado en oro **A**, si los niveles de `B` son diferentes para los diferentes niveles de `A`. Por ejemplo, se lleva a cabo un experimento para estudiar la terminación de unas ciertas piezas, para ello se eligen cuatro máquinas, y cada una de ellas está operada por tres trabajadores, los cuales fabrican dos piezas cada una. El factor *trabajador* está anidado en el factor *máquina* ya que los trabajadores son diferentes para cada una de las máquinas.
Un ejemplo bastante frecuente es el de los diseños tipo split-plot. En un diseño split-plot hay dos factores, uno de ellos por ejemplo el factor B es más sencillo de aplicar y otro, el factor A es más difícil. Por ejemplo, el factor A puede representar variedades de una planta e y el factor B el tipo de fertiizante. Los niveles de A se asigna aleatoriamente a los *whole plots* mientras que los niveles de B se asignan aleatoriamente a los *split-plots*. En la siguiente figura mostramos un ejemplo de un posible experimento tipo split-plot.



Los datos que vamos a utilizar se encuentran en el archivo `Avena.txt` y corresponden a seis bloques, cada uno de los cuales consiste en tres *whole-plot*. Los niveles del factor `Variety` se asigna aleatoriamente a los *whole-plots*. Cada *whole-plot* se divide en cuatro *spit-plots* y los niveles de fertilizante (`nitro`) se asignan también de forma aleatoria, y la variable respuesta (`yield`) es la producción de avena. Es importante tener en cuenta la estructura del experimento, ya que las comparacione entre variedades se harán un nivel y la de los fertilizantes a otro nivel. Además, en este caso:

$$\begin{aligned} Var(y_{ijk}) &= \sigma_B^2 + \sigma_V^2 + \sigma^2 \\ Cov(y_{ijk}, y_{ijk^*}) &= \sigma_B^2 + \sigma_V^2 \\ Cov(y_{ijk}, y_{ij^*k^*}) &= \sigma_B^2 \\ Cov(y_{ijk}, y_{i^*jk^*}) &= 0 \\ Cov(y_{ijk}, y_{i^*j^*k^*}) &= 0 \end{aligned}$$

donde $i = 1 \ldots, 6$, $j = 1, 2, 3$, y $k = 1, \ldots, 4$. La matriz de efectos aleatorios $\boldsymbol{Z}$ sería:

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{Z}2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{Z}_6 \end{bmatrix} \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{G}2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{G}_6 \end{bmatrix}$$

donde:

$$\boldsymbol{Z}_i = \begin{bmatrix} 1 & 1 & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & 1 & \\ 1 & & 1 & \\ 1 & & 1 & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \end{bmatrix} \quad \boldsymbol{G}_i = \begin{bmatrix} \sigma^2_B & & & \\ & \sigma^2_V & & \\ & & \sigma^2_V & \\ & & & \sigma^2_V \end{bmatrix}$$

Lo primero que tenemos que ver es cuál es la estructura aletoria correcta, para ello ajustamos tres modelos:

```
Avena1=lm(yield~Variety*factor(nitro),Avena)
Avena2=lme(yield~Variety*factor(nitro),random=~1|Block,Avena)
Avena3=lme(yield~Variety*factor(nitro),random=~1|Block/Variedad,Avena)
```

En los que incluímos todos los efectos fijos. Ahora utilizamos a aproximación del test de la razón de verosimilitud. Primero hacemos un test para saber si es necesario incluir el efecto de Bloque:

```
test12=-2*logLik(Avena1, REML=T) +2*logLik(Avena2, REML=T)
mean(pchisq(test12,df=c(0,1),lower.tail=F))
 7.178078e-08
```

Dado que es significativo, ahora contrastamos si es necesario el efecto variedad dentro de bloque:

```
test23=-2*logLik(Avena2, REML=T) +2*logLik(Avena3, REML=T)
mean(pchisq(test23,df=c(0,1),lower.tail=F))
0.002820642
```

Una vez elegida la estructura de efectos aleatorios, vemos qué efectos fijos son significativos, para ellos ajustamos los siguientes modelos mediante máxima verosimilitud:

```
Avena3.1=lme(yield~1,random=~1|Block/Variety,Avena,method="ML")
Avena3.2=lme(yield~Variety,random=~1|Block/Variety,Avena,method="ML")
Avena3.3=lme(yield~Variety+factor(nitro),random=~1|Block/Variety,Avena,method="ML")
Avena3.4=lme(yield~Variety*factor(nitro),random=~1|Block/Variety,Avena,method="ML")

anova(Avena3.1,Avena3.2,Avena3.3,Avena3.4)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Avena3.1     1  4 675.4840 684.5907 -333.7420
Avena3.2     2  6 676.3727 690.0327 -332.1863 1 vs 2  3.11134  0.2110
Avena3.3     3  9 616.0432 636.5332 -299.0216 2 vs 3 66.32950  <.0001
Avena3.4     4 15 625.9057 660.0557 -297.9529 3 vs 4  2.13746  0.9066
```

Por lo que el modelo final sería el que incluye sólo el nivel de nitrógeno como factor fijo:

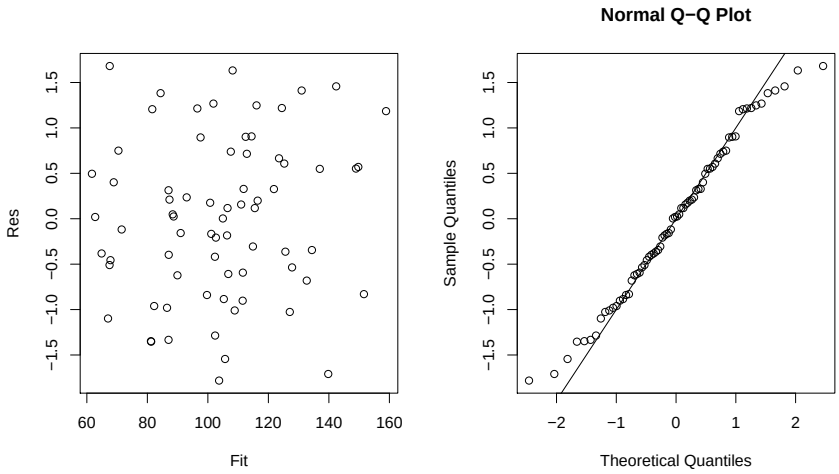
```
Avena.final=lme(yield~factor(nitro),random=~1|Block/Variety,Avena,method="ML")

Random effects:
 Formula: ~1 | Block
        (Intercept)
StdDev:    12.89669

 Formula: ~1 | Variety %in% Block
        (Intercept) Residual
StdDev:    11.14048 12.39064

Number of Observations: 72
Number of Groups:
             Block Variety %in% Block
                 6                 18
```

El modelo nos dice que no hay diferencia significativa entre la producción media de avena de las distintas variedades, pero que la variabilidad entre las variedades dentro de los bloques sí es significativa. Los efectos aleatorios serían:

```
ranef(Avena.final,level=1)
ranef(Avena.final,level=2)
```

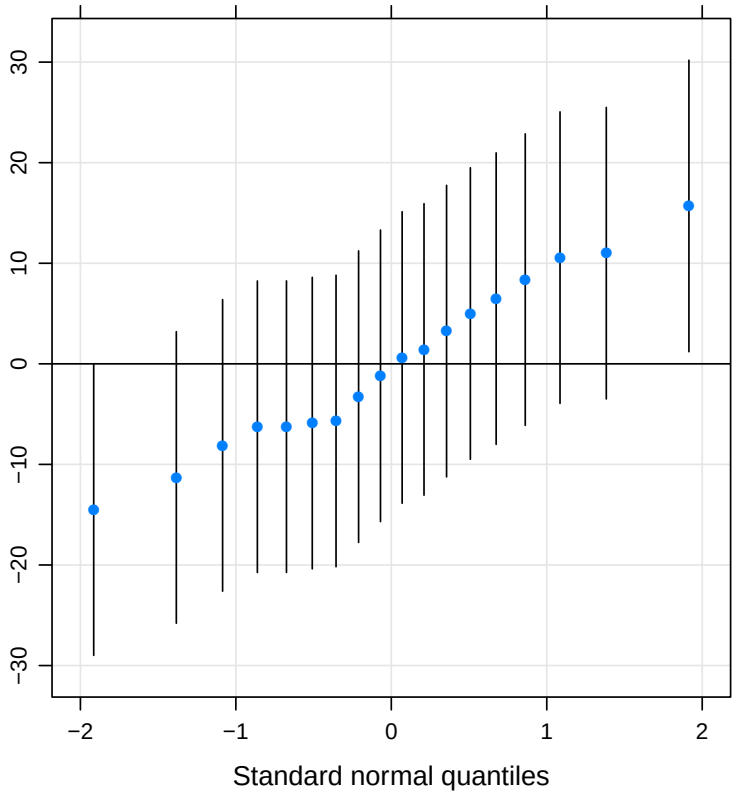
Hacemos gráficos de residuos en los distintos niveles, empezando por los residuos a nivel más bajo:

```
Res=residuals(Avena.final,type="normalized")
Fit=fitted(Avena.final)
par(mfrow=c(1,2))
plot(Res~Fit)
qqnorm(Res)
abline(0,1)
```



Para comprobar la normalidad de los residuos a nivel de bloque y variedad dentro de bloque, ajustamos el modelo con la función `lmer`:

```
Avena.lmer=lmer(yield~factor(nitro)+(1|Block/Variety),Avena)
qqmath(ranef(Avena.lmer, postVar = TRUE), strip = FALSE)$Block
qqmath(ranef(Avena.lmer, postVar = TRUE), strip = FALSE)$`Variety:Block`
```



Ejercicio: Variedades de hierba

Los datos se encuentran en el data.frame `Cultivation` de la librería `SASmixed`. Corresponden a un experimeto en el que se utilizaron 4 bloques, cada bloque se dividió en dos y dos variedades de hierba se asignaron aleatoriamente a cada mitad. A su vez, cada parcela ocupada por una variedad se dividió en tres y se inoculó con tres bacterias distintas. Ajusta el modelo adecuado y comprueba las hipótesis del modelo.

Capítulo 3

Modelos multinivel

Las estructura jerárquicas o multinivel, aparecen con frecuencia en las Ciencias Sociales, Naturales y en Medicina. Ejemplos de ello son los individuos anidados en áreas geográficas o instituciones. Cuando los individuos forman cluster o grupos, es de esperar que dos individuos selecionados aleatoriamente de un mismo grupo se parezcan más entre si que dos individuos de grupos distintos. Por ejemplo, las características de una escuela (pública o privada) es probable que influyan en el rendimiento de los alumnos. Debido a este efecto de la escuela, los resultados de los alumnos de una misma escuela se parecerán entre si más que a los resultados de otras escuelas.

Los niveles se nombran de forma ascendente, empezando por el nivel más elemental. A los niveles se les asocia un subíndice:

- Nivel 1 $\rightarrow i$ (alumnos)
- Nivel 2 $\rightarrow j$ (escuelas)

¿Qué ocurre cuando ignoramos la estructura jerárquica de los datos?. Si los datos están agrupados y no tenemos en cuenta el efecto de agrupamiento en nuestro modelo de regresión, la hipótesis de independencia no se satisface. Supongamos que estamos interesados en predecir los resultados escolares y si hay diferencias debidas al sexo y la raza de los alumnos. Si no estamos interesados en las diferencias entre las escuelas, probablemente ignoremos este efecto de agrupamiento. La consecuencia es que los errores estándar de los coeficientes de regresión en el modelo serán generalmente subestimados y, por lo tanto, los intervalos de confianza serán demasiado estrechos y los p-valores pequeños, lo que nos puede llevar a concluir que una variable explicatiba tiene un efecto real en la variable respuesta cuando en realidad no es cierto. Los errores estándar se estimarán correctamente sólo cuando se tenga en cuenta la variabilidad entre los grupos, y los modelos multinivel son una forma eficiente de conseguirlo. Estos modelos permiten investigar la naturaleza de la variabilidad entre los grupos, y los efectos de las características grupales sobre los resultados individuales.

Los datos con los que vamos a trabajar en este capítulo (`mates.txt`) provienen de un estudio titulado *High School and Beyond* . Los datos corresponden a 7185 estudiantes repartidos en 160 escuelas, el número de alumnos por escuela varía entre 14 y 67. La variable de interés

`mat` es el nivel estandarizado alcanzado en matemáticas (media= 12.75, sd=6.88). Hay dos variables predictoras a nivel del alumnos: `cennivel`, es una medida del nivel socioeconómico del alumno (centrado en la media), y `sexo`. Además hay una variable medida a nivel de la escuela, `sector`, que la identifica como pública o privada. Una cuestión inicial que nos podemos plantear es si el nivel socioeconómico del alumno predice las diferencias en el nivel de matemáticas. Para ello ajustaríamos el modelo:

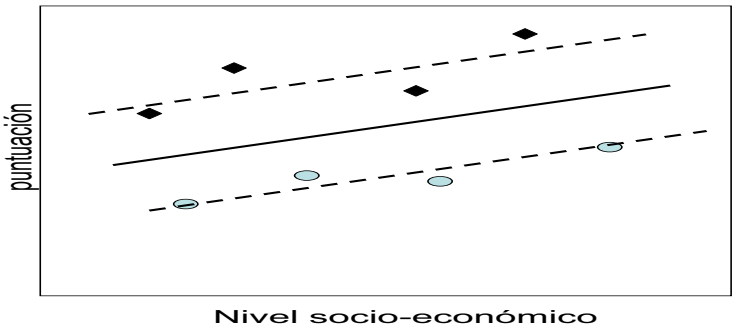
$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j$$

este modelo ignora que los alumnos provienen de distintos centros (por eso solo aparece el subíndice j que es el que representa a las unidades de nivel más bajo, en este caso a los alumnos). Abrimos el archivo `mates.R`:

```
mates=read.table("mates.txt",header=TRUE)
mates$centro=factor(mates$centro)
mates$sexo=factor(mates$sexo)
mates$sector=factor(mates$sector)
attach(mates)
multi0=lm(mat~cennivel)
summary(multi0)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.76099    0.07934  160.84   <2e-16 ***
cennivel     2.19109    0.12010   18.24   <2e-16 ***
```

La ordenada en el origen es 12.76 y la pendiente 2.19, lo que indica que por cada unidad que aumenta el nivel socio-económico, la puntuación del test aumenta en 2.19 unidades, además podemos ver que el coeficiente es significativo.

Pero supongamos que ocurre una situación como la que aparece en la siguiente Figura



Los alumnos de la escuela A sacan, en promedio, mejores notas que las que le asignaría el modelo ajustado, con la escuela B ocurre lo contrario. El gráfico indica que la ordenada en el origen no debería ser la misma para todos los centros, sino que debería ser distinta para distintos centros. Es decir, el valor predicho debe ajustarse hacia arriba o abajo, eso lo podemos conseguir permitiendo que cada escuela tenga su propia ordenada en el origen

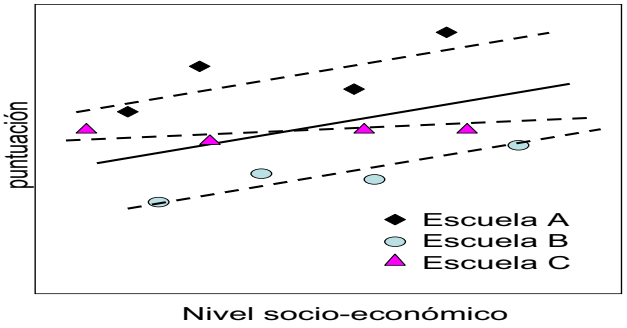
$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \epsilon_{ij}$$

este modelo es similar al anterior pero hemos añadido el subíndice i para identificar el centro al que pertence cada alumno. En realidad, lo que hacemos es utilizar una variable categórica con tantas categorías como escuelas

```
multi1=lm(mat~cennivel+centro)
summary(multi1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.72943    0.88731  10.965  < 2e-16 ***
cennivel     2.19117    0.10865  20.168  < 2e-16 ***
centro2      3.79539    1.50582   2.520 0.011742 *
centro3     -2.08142    1.24830  -1.667 0.095478 .
centro4      6.53921    1.62405   4.026 5.72e-05 ***
.
.
.
```

Esto hace que consideremos a las escuelas como un efecto fijo y no aleatorio, es decir, implícitamente estamos suponiendo que solo nos interesan estas esculas en particular.

Las cosas se pueden complicar más, es posible que el efecto del nivel socio-económico sea distinto para cada centro, es decir, que un aumento de una unidad en ese nivel puede dar lugar a un aumento distinto en la nota del test en cada centro. En el siguiente gráfico vemos como la pendiente de la recta para la escuela C es distinta a las dos anteriores. El modelo



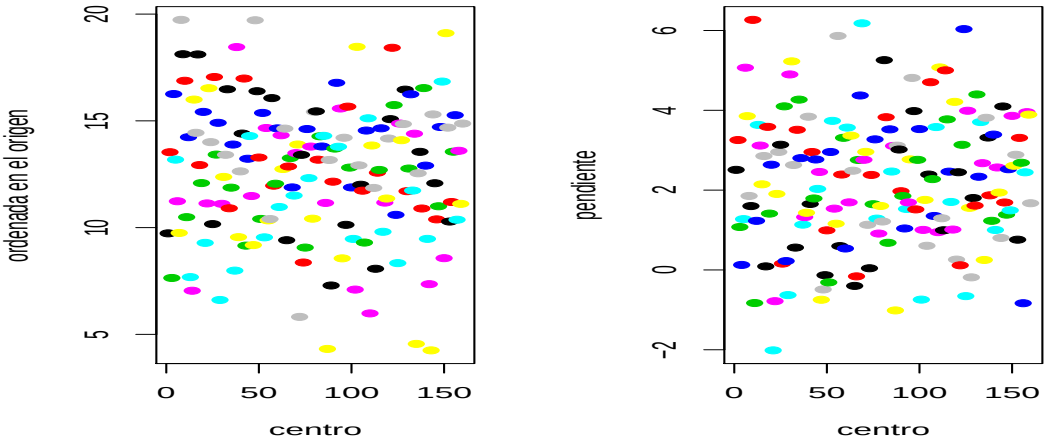
que permite tener en cuenta esta situaciones:

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}$$

```
multi2=lm(mat~cennivel*centro)
summary(multi2)
```

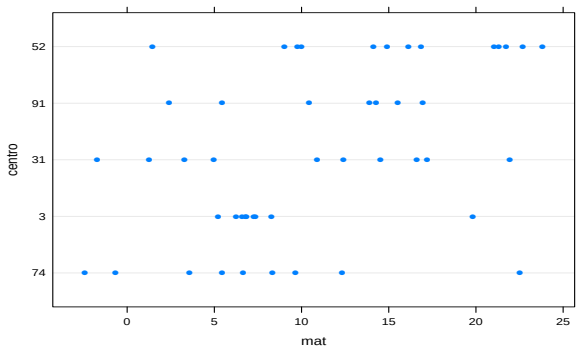
Pero como hemos dicho antes, no estamos interesados en estas escuelas en concreto, sino en la población de la que estas escuelas son una muestra. Esto nos permite comparar escuelas con distintas características.
El gráfico siguiente muestra el valor de la ordenada en el origen y la pendiente de cada una

de las rectas correspondientes a cada escuela, podemos ver que hay mucha variabilidad, sobre todo en la ordenada en el origen. Con un modelo multinivel podemos contestar a preguntas como: ¿Cúales son las causas de esta variabilidad?, ¿qué variables pueden explicarla?.



1. Modelo multinivel para las medias de grupo

Es el modelo multinivel más sencillo. Consideramos que los datos tienen una estructura con dos niveles, los alumnos están en el nivel 1 y están agrupados en escuelas, nivel 2. Vamos a empezar suponiendo que no disponemos de ninguna variable explicativa, y que por lo tanto nuestro único interés es la diferencia entre las notas medias del test de matemáticas entre los distintos centros. La siguiente figura muestra los datos de 5 escuelas; se puede apreciar la variabilidad existente tanto entre escuelas como dentro de las escuelas.



Especificamos los dos niveles del modelo:

$$\text{Nivel 1: } y_{ij} = \mu_i + \epsilon_{ij}$$

El subíndice j corresponde a individuos y el i a escuelas, si consideramos a las escuelas como un efecto aleatorio, entonces μ_i (la media de cada escuela) vendría dada por:

$$\text{Nivel 2: } \mu_i = \beta_0 + u_i$$

donde β_0 es la media de todos los alumnos y ui es la desviación de la media de la escuela i de la media de todas las escuelas.

Poniendo las dos ecuaciones juntas:

$$y_{ij} = \beta_0 + u_i + \epsilon_{ij}, \quad i = 1, \ldots, m \quad j = 1, \ldots, n_m.$$

Lo que indica que la nota de un alumno j en la escuela i es la suma de la media total (β_0), más la desviación de la media de la escuela i a la media total (u_i), más la desviación del alumno j de la media de su escuela (ϵ_{ij}).

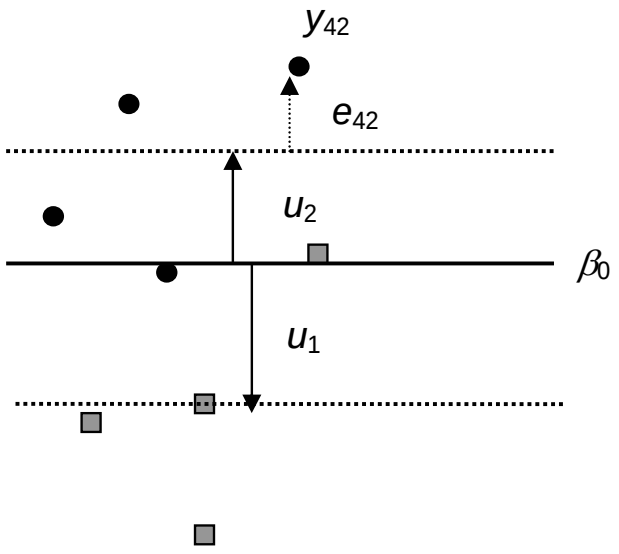
En notación matricial el modelo anterior viene dado por:

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

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{mn_m} \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \mathbf{1_1} \\ \vdots \\ \mathbf{1_m} \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1_1} & 0 & \ldots & 0 \\ 0 & \mathbf{1_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & \mathbf{1_m} \end{bmatrix} \quad \mathbf{1_i} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{\mathbf{n_i \times 1}}$$

La media de $\boldsymbol{y}$ para el grupo i viene dada por $\beta_0 + u_i$, y por lo tanto, u_i es la desviación del la media del grupo i respecto de la media total. Los residuos a nivel individual ϵ_{ij} son la diferencia entre el valor de la variable respuesta del individuo j y la media del grupo al que pertenece. La sigiente figura muestra esta descomposición de los residuos:



escuela	u_i
91	0.918
3	-4.568
31	-0.268
25	2.537
74	-3.907

Los residuos a ambos niveles se suponen que provienen de una población con distribución Normal, $u_i \sim N(0, \sigma_i^2)$ y $\epsilon_{ij} \sim N(0, \sigma^2)$, y ambos son independientes, es decir, las observaciones que vienen de distintas escuelas son independientes.

En nuestro ejemplo:

```
Modelo0=lme(mat~1,random=~1|centro)
Modelo0
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -23558.4
  Fixed: mat ~ 1
(Intercept)
   12.63697

Random effects:
 Formula: ~1 | centro
        (Intercept) Residual
StdDev:    2.934966 6.256862
```

La media total estimada es 12,64, y la media para la escuela i es $12{,}64 + \hat{u}_i$ donde $\hat{u}_i$ es el efecto aleatorio predicho para la escuela. Estos efectos son variables aleatorias con una distribución Normal, por lo tanto, su distribución depende de dos parámetros: la media y la varianza. La estimación de los efectos fijos y la predicción de los efectos aletorios y de los parámetros de la varianza del modelo se hacen mediante máxima verosimilitud o máxima verosimilitud restringida (como vimos el el capítulo anterior). La siguiente tabla muestra los valores predichos de los efectos aleatorios correspondientes a 5 escuelas: Por ejemplo, para la escuela identificada como 3, el efecto aletorio (o residuo) tiene un valor de $-4{,}568$, y la puntuación media para esta escuela vendría dada por $\beta_0 + u_3 = 12{,}64 - 4{,}568 = 8{,}072$. Si hubiéramos considerado la escuela como un efecto fijo (una variable categórica) en vez de como un efecto aleatorio, sólo hubiéramos tenido un componente de la varianza σ^2, en cambio, en un modelo con efecto aleatorio, tenemos dos componentes: σ^2 y σ_u^2. Por lo tanto la variabilidad total en los datos viene dada por:

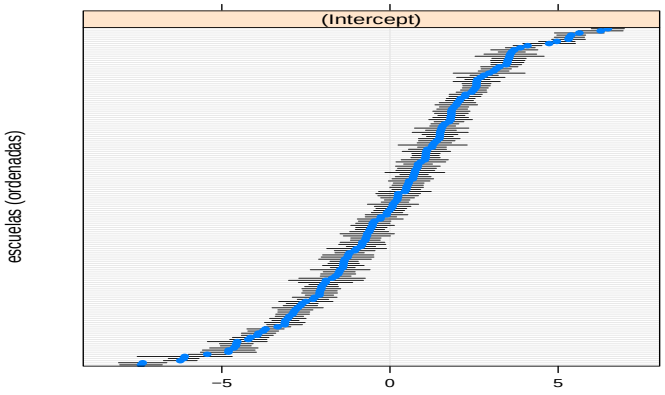
$$Var(y_{ij}) = \sigma^2 + \sigma_u^2$$

(variabilidad dentro de los centros más varibilidad entre los centros).

El ICC sería:

$$\frac{2{,}93^2}{2{,}93^2 + 6{,}27^2} = 0{,}18,$$

Por lo tanto el 18 % de la variabilidad total es debida a las diferencias entre las medias de los centros. El 82 % restante es atribuible a las diferencias entre estudiantes dentro de una misma escuela. El siguiente gráfico podemos ver los efectos aleatorios junto con sus intervalos de confianza (las escuelas han sido ordenadas de atendiendo a su media para apreciar mejor la variabilidad entre las mismas.



1.1. Constrastes para el efecto de grupo

Una primera aproximación para contrastar si hay o no diferencias entre los grupos sería calcular el intervalo de confianza para σ_u:

```
intervals(Modelo0)
 Random Effects:
  Level: centro
                    lower     est.    upper
sd((Intercept)) 2.595887 2.934966 3.318335
```

el intervalo no contiene al cero, pero la forma más correcta de hacerlo sería:

$$\begin{aligned} H_0: &\quad y_{ij} = \beta_0 + \epsilon_{ij} \\ H_1: &\quad y_{ij} = \beta_0 + u_i + \epsilon_{ij} \end{aligned}$$

Esto es equivalente a contrastar $H_0 : \sigma_u^2 = 0$. Lo usual es hacer este tipo de contrastes mediante un test de la razón de verosimilitud:

$$LR = -2(\log L_0 - \log l_1)$$

Este tipo de test se puede usar simpre que los modelos a contrastar estén anidados. Como vimos en el capítulo anterior el resultado del test en este caso se compara con el valor de una mixtura de distribuciones Chi-quadrado $0{,}5\chi_0^2 + 0{,}5\chi_1^2$:

```
Modelo_NULL=lm(mat~1)
test=-2*logLik(Modelo_NULL, REML=T) +2*logLik(Modelo0, REML=T)
mean(pchisq(test,df=c(0,1),lower.tail=F))
```

El siguiente paso sería intentar explicar la variabilidad entre los centros mediante variables explicativas medidas tanto al nivel 1 como al nivel 2.

Modelo 1: Variable explicativa a Nivel 1

Como la variable explicativa está medida al Nivel 1, la introducimos en la equación del Nivel 1:

$$\begin{aligned} &\text{Nivel 1:} \quad y_{ij} = \mu_i + \beta_1 x_{ij} + \epsilon_{ij} \\ &\text{Nivel 2:} \quad \mu_i = \beta_0 + u_i \end{aligned}$$

Si $\boldsymbol{x}$ es una variable continua, este modelo asume que la pendiente de la recta es la misma para todas las escuelas (por eso β_1 no lleva el subíndice i). Poniendo las dos ecuaciones juntas:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{efectos fijos} + \underbrace{u_i + \epsilon_{ij}}_{efectos aleatorios}$$

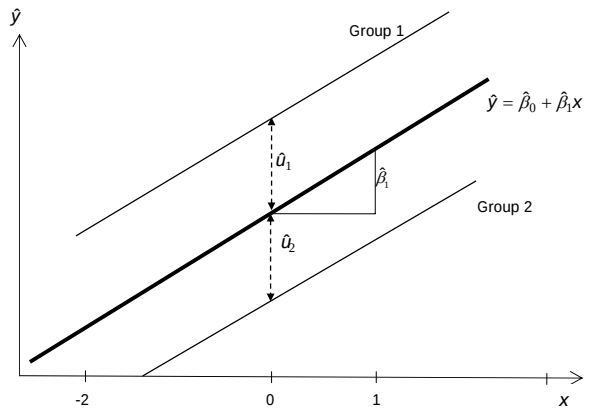
En forma matricial:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{mn_m} \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{mn_m} \end{bmatrix}, \qquad \boldsymbol{\beta} = [\beta_0, \beta_1]^T$$

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1}_1 & 0 & \dots & 0 \\ 0 & \mathbf{1}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{1}_m \end{bmatrix}, \qquad \mathbf{1}_j = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n_j \times 1}$$

En este modelo, la relación global entre $\boldsymbol{y}$ y $\boldsymbol{x}$ viene representada por la línea recta con ordenada en el origen β_0 y pendiente β_1. Sin embargo, la ordenada en el origen para un determinado grupo i viene dada por $\beta_0 + u_i$, es decir, será mayor o menor que que la ordenada en el origen global β_0 por una cantidad u_i. Aunque la ordenada en el origen varía de grupo a grupo, la pendiente es la misma para todos los grupos. Por lo tanto las líneas rectas ajustadas para cada grupo son paralelas. La siguiente figura muestra la línea global ajustada y las líneas para dos grupos.



En nuestro ejemplo, introducimos como variable explicativa `cennivel` (el nivel socioeconómico centrado), de modo que el modelo viene dado por:

$$\texttt{mat} = \beta_0 + \beta_1 \texttt{cennivel} + u_i + \epsilon_{ij}$$

```
Modelo1=lme(mat~cennivel,random=~1|centro)
Modelo1
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -23362
  Fixed: mat ~ cennivel
(Intercept)    cennivel
  12.649286    2.191168
Random effects:
 Formula: ~1 | centro
        (Intercept) Residual
StdDev:    2.944893 6.083618
```

Ahora tenemos dos efectos fijos:

$$\begin{aligned} \hat{\beta}_0 &= 12{,}62 \\ \hat{\beta}_1 &= 2{,}19 \end{aligned}$$

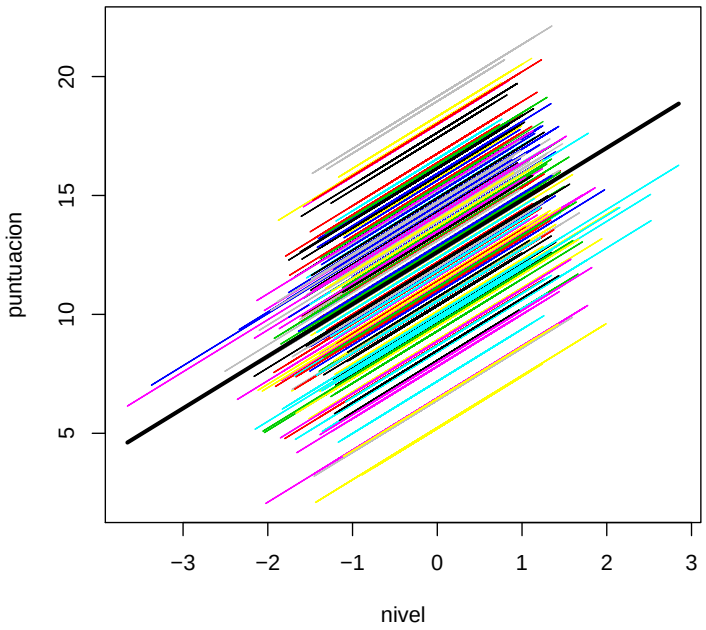
$\hat{\beta}_0$ es la nota media para alumnos con nivel socioeconómico medio (la variable está centrada) y la recta media vendría dada por:

$$12{,}62 + 2{,}19\ \texttt{cennivel}$$

Para contrastar si la pendiente es significativamente distinta de cero tendríamos que hacer un test de la razón de verosimilitus para efectos fijos, usando máxima verosimilitud (no máxima verosimilitud restringida):

```
Modelo0.ML=update(Modelo0,method="ML")
Modelo1.ML=update(Modelo1,method="ML")
anova(Modelo0.ML,Modelo1.ML)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Modelo0.ML     1  3 47121.81 47142.45 -23557.90
Modelo1.ML     2  4 46728.41 46755.93 -23360.21 1 vs 2 395.3969  <.0001
```

Comparado con el modelo sin la variable explicativa, la inclusión del nivel socioeconómic del alumno en el modelo ha reducido la varianza a nivel del alumno en un 5 % ($(6{,}25^2 - 6{,}08^2)/6{,}25^2 = 0{,}05$). El siguiente gráfico muestra las rectas ajustadas para cada escuela.



<u>Modelo 2</u>: Variable explicativa a Nivel 2

Si las variable explicativas son medidas al Nivel 2:

$$\begin{aligned} \text{Nivel 1:} \quad & y_{ij} = \mu_i + \epsilon_{ij} \\ \text{Nivel 2:} \quad & \mu_i = \beta_0 \beta_2 s_i + u_i \\ y_{ij} \;=\; & \underbrace{\beta_0 + \beta_2 s_i}_{\text{efectos fijos}} + \underbrace{u_i + \epsilon_{ij}}_{\text{efectos aleatorios}} \end{aligned}$$

En nuestr caso, la variable utilizada es `sector`:

$$\texttt{mat} = \beta_0 + \beta_2 \texttt{sector} + u_i + \epsilon_{ij}$$

```
Modelo2= lme(mat~sector,random=~1|centro)
Modelo2
Linear mixed-effects model fit by REML
  Log-restricted-likelihood: -23540.07
  Fixed: mat ~ sector
(Intercept)     sector1
  11.393044    2.804887
Random effects:
```

```
 Formula: ~1 | centro
        (Intercept) Residual
StdDev:    2.583981 6.257108
```

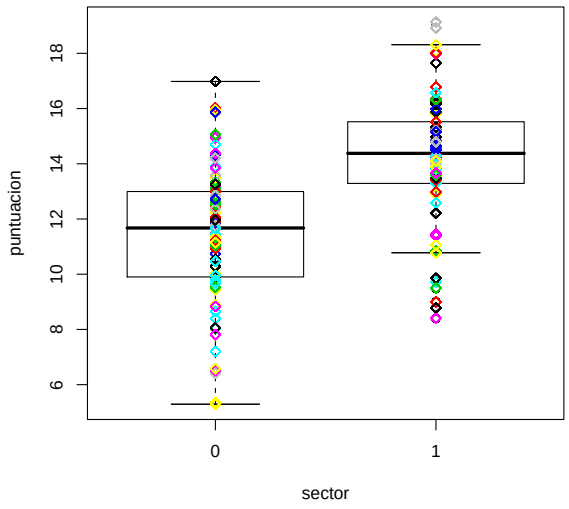
En cuanto a los efectos aleatorios, la varianza del efecto aleatorio de nivel 2 σ_u^2 ha descendidio:$(2{,}93^2 - 2{,}58^2)/2{,}93^2 = 0{,}22$, es decir que se ha reducido en un 22 % la variabilidad no explicada entre los centros al introducir la variable sector. En el caso de los efectos fijos, tenemos que:

$$\begin{aligned} \hat{y}|sector = 0 &= 11{,}39 \\ \hat{y}|sector = 1 &= 11{,}39 + 2{,}8 = 14{,}19 \end{aligned}$$

esto lo interpretaríamos así: la nota un alumno en una escuela privada se espera que sea 2.8 unidades mayor que la de un alumno en una escuela pública (podemos generalizar ya que hemos asumido que las escuelas son un efecto aleatorio). Para contrastar si la variable `sector` es significativa utilizaremos de nuevo la función anova y el LRT.

```
Modelo2.ML=update(Modelo2,method="ML")
anova(Modelo0.ML,Modelo2.ML)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Modelo0.ML     1  3 47121.81 47142.45 -23557.90
Modelo2.ML     2  4 47087.11 47114.62 -23539.55 1 vs 2 36.70476  <.0001
```

El siguiente gráfico muestra las medias para cada centro atendiendo a si son públicos o privados:



2. Modelos con pendiente aleatoria

En este tipo de modelos suponemos que la relación entre la variable respuesta y las variables explicativas va a ser distinta para las distintas unidades de nivel 2, es decir, la relación puede cambiar de un centro a otro. Por ejemplo, el efecto del nivel socioeconómico

en las notas puede ser distinto en distintos centros, de modo que podemos relajar el modelo anterior, en el que la pendiente era la misma para todos los grupos, permitiendo que la pendiente varíe aleatoriamente entre los grupos.

<u>Modelo 3</u>: Variables a Nivel 1

$$\begin{aligned} \text{Nivel 1:} \quad & y_{ij} = \mu_i + \beta_{1i} x_{ij} + \epsilon_{ij} \\ \text{Nivel 2:} \quad & \mu_i = \beta_0 + u_i \\ & \beta_{1i} = \beta_1 + v_i \end{aligned}$$

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{efectos fijos}} + \underbrace{u_i + v_i x_{ij} + \epsilon_{ij}}_{\text{efectos aleatorios}}, \quad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{G}_i) \quad \boldsymbol{G}_i = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}$$

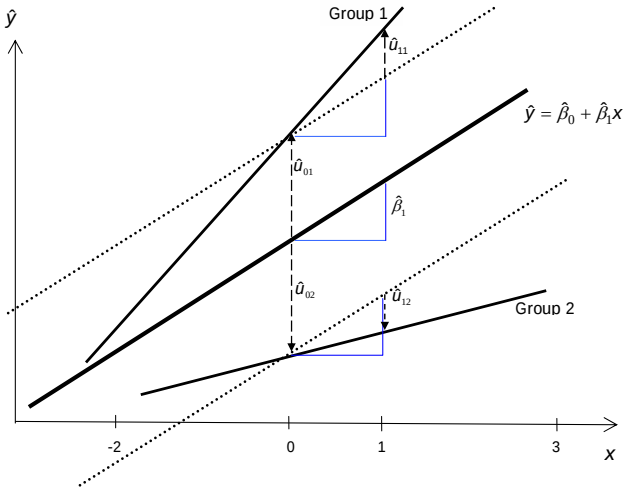
Donde σ_{uv} es la covarianza entre entre las ordenadas en el origen de los grupos y las pendientes. Un valor positivo de la covarianza implica que grupos con una un valor del efecto de grupo u_i elevado, tienden a tener valores elevados de v_i, o equivalentemente, centros con ordenada en el origen alta, tienen pendiente alta. En forma matricial el modelo es:

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

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{mn_m} \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \qquad \boldsymbol{X}_i = \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_m \end{bmatrix}, \qquad \mathbf{1}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n_i \times 1} \qquad \boldsymbol{\beta} = [\beta_0, \beta_1]^T.$$

La siguiente figura muestra una representación de los modelos con y sin pendiente aleatoria, es decir, con y sin interacción entre el nivel socioeconómico del alumno y la escuela



En nuestro ejemplo, este gráfico significaría que el aumento del nivel socioeconómico tiende a mejorar resultados en ambas escuelas, esa mejora es más rápida en la escuela 1 que en la 2, y que esa diferencia se va agrandando con el aumento del nivel socioeconómico.

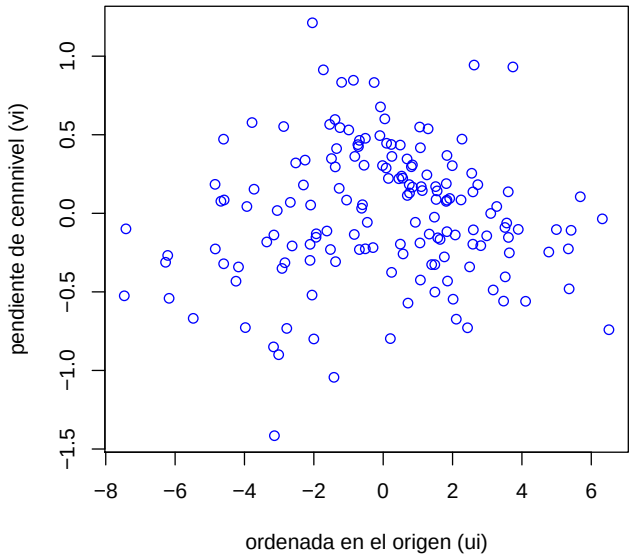
```
Modelo3= lme(mat~cennivel,random=~cennivel|centro)
Modelo3
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -23357.12
  Fixed: mat ~ cennivel
(Intercept)    cennivel
  12.649339    2.193192

Random effects:
 Formula: ~cennivel | centro
 Structure: General positive-definite,
            StdDev    Corr
(Intercept) 2.9464629 (Intr)
cennivel    0.8330628 0.021
Residual    6.0580687
```

El efecto del nivel socieconómico en la escuela i se estima como $2{,}19 + \hat{u}_i$, y la varianza de las pendientes entre escuelas es $0{,}833^2 = 0{,}694$. Para la *escuela promedio* predecimos un aumento de 2,19 en la puntuación cuando el nivel socioeconómico aumenta en una unidad.
Ahora tenemos los siguientes parámetros de la varianza:

$$\hat{\sigma}_u^2 = 8{,}67 \quad \hat{\sigma}_v^2 = 0{,}694 \quad \hat{\sigma}_{uv} = \rho\sigma_u\sigma_v = 0{,}051 \quad \hat{\sigma}^2 = 36{,}7$$

La varianza de ordenada en el origen estimada, 8,67 se interpreta como la variabilidad entre las escuelas para un nivel socioeconómico medio. La varianza de ordenada en el origen estimada, 8,67 se interpreta como la variabilidad entre las escuelas para un nivel socioeconómico medio. En el siguiente gráfico podemos ver que efectivamente no existe correlación entre las odenadas en el origen y las pendientes de las diferentes escuelas, de modo que nos podemos plantear si es necesario el parámetros σ_{uv}.

En este caso H_0: $\sigma_{uv} = 0$ y H_1: $\sigma_{uv} \neq 0$ ya que la covarianza puede tomar cualquier valor, y or lo tanto el test de la razón de verosimilitud es exacto:




```
Modelo3.1= lme(mat~cennivel,random = list(centro=pdDiag(~cennivel)))
Modelo3.1
Random effects:
 Formula: ~cennivel | centro
 Structure: Diagonal
        (Intercept)  cennivel Residual
StdDev:    2.946427 0.8329191 6.058073
anova(Modelo3.1,Modelo3)
          Model df      AIC      BIC    logLik   Test    L.Ratio p-value
Modelo3.1     1  5 46724.25 46758.65 -23357.12
Modelo3       2  6 46726.24 46767.51 -23357.12 1 vs 2 0.01559759  0.9006
```

El siguiente paso sería contrastar si es necesario que las rectas tengan pendientes diferentes, es decir, H_0: $\sigma_v^2 = 0$, H_1: $\sigma_v^2 > 0$, en este caso sí necesitamos la aproximación:

```
test=-2*logLik(Modelo1, REML=T) +2*logLik(Modelo3.1, REML=T)
mean(pchisq(test,df=c(0,1),lower.tail=F))
 0.0008984922
```

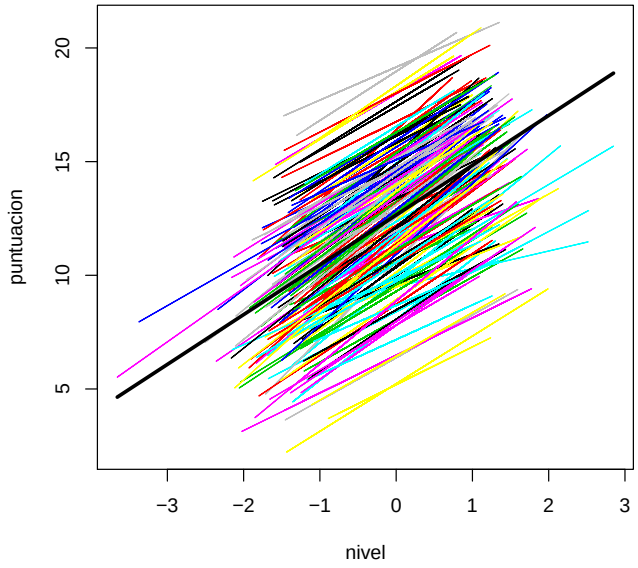
Además, podemos usar algún criterio de información para comparar los modelos:

```
 AIC(logLik(Modelo3))    AIC(logLik(Modelo3.1))    AIC(logLik(Modelo1))
 46726.24                 46724.25                 46732
```

El siguiente gráfico muestra las rectas correspondientes a cada escuela, junto con la recta promedio

Este modelo asume que la variabilidad entre las escuelas depende del nivel socioeconómico. Podemos comprobarlo explícitamente, calculando la varianza entre-escuelas:

$$\begin{aligned} Var(\mu_i) &= Var(u_i + v_i x_{ij}) = Var(u_i) + Var(v_i)x_{ij}^2 + 2x_{ij}\underbrace{Cov(u_i, v_i)}_{0} \\ &= 8{,}68 + 0{,}694\texttt{cennivel}^2 \end{aligned}$$



Modelo 4: Variables a Nivel 1 y 2

En este tipo de modelos suponemos que tanto las pendientes como las ordenadas en el origen para cada centro pueden ser explicadas por variables de nivel 2.

$$\begin{aligned} \text{Nivel 1:} \quad & y_{ij} = \mu_i + \beta_{1i} x_{ij} + \epsilon_{ij} \\ \text{Nivel 2:} \quad & \mu_i = \beta_0 + \beta_2 s_i + u_i \\ & \beta_{1j} = \beta_1 + \beta_3 s_i + v_i \end{aligned}$$

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij} + \beta_2 s_i + \beta_3 x_{ij} : s_i}_{\text{efectos fijos}} + \underbrace{u_i + v_i x_{ij} + \epsilon_{ij}}_{\text{efectos aleatorios}}$$

Al introducir la variable medida al nivel 2, la parte fija se modifica (con respecto al Modelo 3) pero no la parte aleatoria, sin embargo es de esperar que la varianza de los efectos aleatorios se reduzca ya que parte de la variabilidad ha sido explicada por la variable a nivel 2. Estimamos β_2 para saber si los centros privados son diferentes de los públicos en cuanto a su nota media. Estimamos β_3 para saber si los centros privados difieren de los públicos en cuanto a la relación entre el nivel socio-económico y la puntuación.

```
Modelo4= lme(mat~cennivel*sector,random = list(centro=pdDiag(~cennivel)))
summary(Modelo4)
 Data: NULL
       AIC      BIC    logLik
  46662.88 46711.03 -23324.44

Random effects:
 Formula: ~cennivel | centro
 Structure: Diagonal
        (Intercept)  cennivel Residual
StdDev:    2.597427 0.5157795 6.058111
```

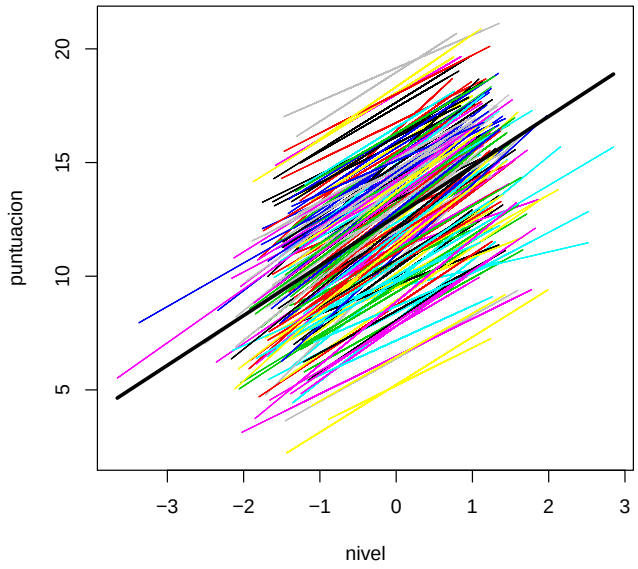
```
Fixed effects: mat ~ cennivel * sector
                     Value Std.Error   DF  t-value p-value
(Intercept)      11.409644 0.2929341 7023 38.94952       0
cennivel          2.784446 0.1556796 7023 17.88575       0
sector1           2.797268 0.4394370  158  6.36557       0
cennivel:sector1 -1.345710 0.2345135 7023 -5.73830       0
```

Todos los efectos fijos sparecen significativos. Para comprobarlo reajustamos el modelo:

```
Modelo4.1=lme(mat~cennivel*sector,random = list(centro=pdDiag(~cennivel)),method="ML")
anova(Modelo4.1)
                numDF denDF  F-value p-value
(Intercept)         1  7023 3393.291  <.0001
cennivel            1  7023  358.752  <.0001
sector              1   158   41.234  <.0001
cennivel:sector     1  7023   33.357  <.0001
```

Los centros privados tienen una nota media significativamente más alta que los públicos (2.79) y tienen una pendiente más suave que la de los centros públicos (-1.34), es decir, que en un colegio privado la mejora de la nota con respecto al nivel socio-económico es más suave que un colegio público.

La figura anterior muestras las rectas promedio para escuelas públicas y privadas. Se puede apreciar



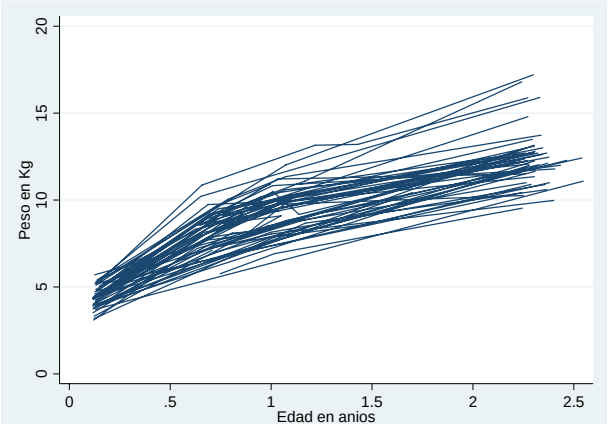
como la pendiente es menor para las escuelas privadas.

Capítulo 4

Medidas repetidas y datos longitudinales

Las repetidas medias repetidas y los datos longitudinales (medidas repetidas a lo largo del tiempo), pueden verse como modelos multinivel donde las medidas repetidas están anidadas en los individuos. Esto nos llevaría a un modelo con dos niveles, donde el nivel más bajo (nivel 1) son las medidas repetidas que están agrupadas por individuos (nivel 2). Las medidas longitudinales pueden tomarse en instantes fijos de tiempo o en ocasiones distintas. La ventaja de los modelos multinivel es que se pueden utilizar en ambos casos, ya que no es necesario que el número de medidas sea la misma para todos los individuos ni se haya tomado en el mismo instante de tiempo, de modo que se pueden analizar datos aunque haya datos faltantes para algún individuo, o que decidan abandonar un ensayo clínico. Cuando tenemos pocas medias repetidas, por ejemplo, cuando medimos algo antes y después de un tratamiento, y estamos simplemente interesados en comparar las medias antes y después del tratamiento, podemos utilizar MANCOVA (análisis de la covarianza multivariante).

Para ilustrar este tipo de modelos vamos a utilizar datos de pesos de 68 niños de una comunidad asiatica del Reino Unido (Singer et al., 2003), los niños fueron pesados entre 1 y 5 veces. Además del peso, conocemos la edad de los niños y el sexo (los datos se encuentran en el fichero `child.txt`).



Cuando estamos trabajando con medidas repetidas en el tiempo, es siempre útil hacer un gráfico de los datos para ver qué tipo de tendencia temporal hay en los datos. A la vista del gráfico nos

podemos plantear:

1. Tendencia global lineal/cuadrática
2. Ordenadas en el origen específicas para cada niño
3. Línea de regresión específicas para cada niño

1. Modelo con ordenada en el origen aleatoria

Comenzamos por ajustar un modelo con una tendencia global cuadrática y un efector aleatorio para cada niño. Es decir, estamos permitiendo que cada niño tenga una trayectoria individual que será paralela a la trayectoria media.

$$\begin{aligned} \texttt{peso}_{ij} &= \mu_i + \beta_1 \texttt{edad}_{ij} + \beta_2 \texttt{edad}_{ij}^2 + \epsilon_{ij} \\ \mu_i &= \beta_0 + u_i \end{aligned}$$

$$\texttt{peso}_{ij} = \beta_0 + \beta_1 \texttt{edad}_{ij} + \beta_2 \texttt{edad}_{ij}^2 + u_i + \epsilon_{ij} \tag{4.1}$$

```
child1=lme(peso~edad+edad2,random=~1|id)
child1
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -281.0327
  Fixed: peso ~ edad + edad2
(Intercept)        edad       edad2
   3.432819    7.818011   -1.705631

Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:   0.9258152 0.7401676
```

Vemos que la variabilidad entre individuos es mayor que dentro de cada individuo. Por lo que debemos buscar alguna manera de explicar esa variabilidad. Para ello permitimos que la diferencia entre la trayectoria global y la de cada niño sea representada mediante una línea, y que la pendiente de esa recta varíe de niño a niño.

2. Modelo con pendiente aleatoria

El modelo que ajustamos para el peso del niño i en el instante j es:

$$\texttt{peso}_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)\texttt{edad}_{ij} + \beta_2 \texttt{edad}_i^2 + e_i,$$

$$\texttt{peso}_{ti} = \underbrace{\beta_0 + \beta_1 \texttt{edad}_{ti} + \beta_2 \texttt{edad}_{ij}^2}_{fijo} + \underbrace{u_i + v_i \texttt{edad}_{ij} + \epsilon_{ij}}_{aleatorio} \tag{4.2}$$

```
child2=lme(peso~edad+edad2,random=~edad|id)
child2
  Log-restricted-likelihood: -262.4327
  Fixed: peso ~ edad + edad2
(Intercept)        edad       edad2
   3.494664    7.703452   -1.660091

Random effects:
 Formula: ~edad | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.6459801 (Intr)
edad        0.5116162 0.258
Residual    0.5780657
```

Vemos que la varianza entre individuos ha descendido, y la correlación es postiva. Vamos a contrastar si ambos parámetros son distintos de cero o no:

```
child2.1=lme(peso~edad+edad2,random = list(id=pdDiag(~edad)))
anova(child2.1, child2)
         Model df      AIC      BIC    logLik   Test   L.Ratio p-value
child2.1     1  6 537.6541 557.2921 -262.8270
child2       2  7 538.8654 561.7764 -262.4327 1 vs 2 0.7886937  0.3745
```

Por lo que podemos considerar que la covarianza es cero. Ahora contrastamos si las pendientes han de ser distintas:

```
test=-2*logLik(child1, REML=T) +2*logLik(child2.1, REML=T)
mean(pchisq(test,df=c(0,1),lower.tail=F))
7.988806e-10
```

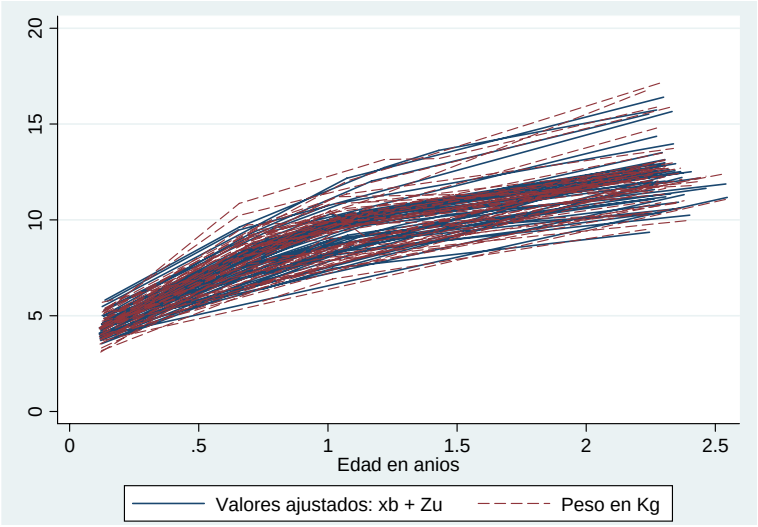
El test de la razón de verosimilitud nos indica que es necesario un modelo con pendiente aleatoria. El siguiente gráfico muestra las trayectorias observadas y las ajustadas. Podemos observar que en la mayoría de los casos el modelo se ajusta bastante bien a los datos.

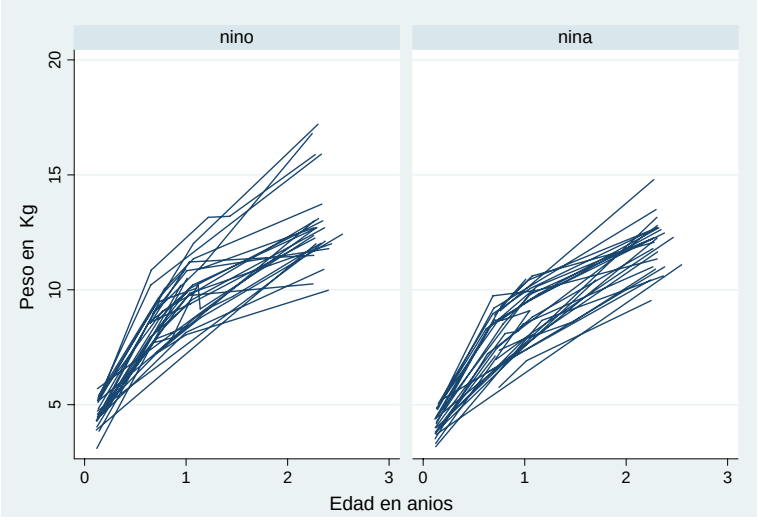
Una cuestión que nos podemos plantear es si hay una diferencia sistemática entre la trayectoria global (de la población) de niños y niñas. Una forma sencilla de havernos una idea es hacer un gráfico de los valores observados por edad:

En el gráfico se aprecia un peso medio mayor en los niños, y como la variabilidad entre los niños aumenta con la edad de forma más pronunciada que entre las niñas. Un primer paso es introducir la variable sexo y su interacción con la edad como un efecto fijo en el modelo:

$$\texttt{peso}_{ij} = (\beta_0 + \beta_3 \texttt{nina}_i + u_i) + (\beta_1 + \beta_4 \texttt{nina}_i + v_i)\texttt{edad}_{ij} + \beta_2 \texttt{edad}_{ij}^2 + \epsilon_{ij},$$

$$\texttt{peso}_{ij} = \beta_0 + \beta_1 \texttt{edad}_{ij} + \beta_2 \texttt{edad}^2{}_{ij} + \beta_3 \texttt{nina}_i + \beta_4 (\texttt{edad}_{ij} : \texttt{nina}_i) + u_i + v_j \texttt{edad}_{ij} + \epsilon_{ij} \quad (4.3)$$






```
child3=lme(peso~edad*nina+edad2,random=list(id=pdDiag(~edad)))
Linear mixed-effects model fit by REML
  Data: NULL
  Log-restricted-likelihood: -259.0672
  Fixed: peso ~ edad * nina + edad2
  (Intercept)          edad      ninagirl         edad2 edad:ninagirl
    3.7543813     7.8068689    -0.5107904    -1.6541318    -0.2296062

Random effects:
 Formula: ~edad | id
 Structure: Diagonal
        (Intercept)      edad  Residual
StdDev:   0.6515836 0.5377058 0.5636585

Fixed effects: peso ~ edad * nina + edad2
                 Value  Std.Error  DF    t-value p-value
(Intercept)    3.749754 0.17127062 127  21.893740  0.0000
edad           7.813443 0.25513536 127  30.624699  0.0000
```

```
ninagirl      -0.505694 0.21158508  66  -2.390028  0.0197
edad2         -1.657646 0.08853549 127 -18.722959  0.0000
edad:ninagirl -0.230204 0.17677471 127  -1.302247  0.1952
```

El efecto de sexo parece significativo pero no la interacción ⇒ en media, los niños pesan más que las niñas, pero su tasa de crecimiento lineal medio no es diferente. ¿Cómo comprobarías si la interacción es significativa?

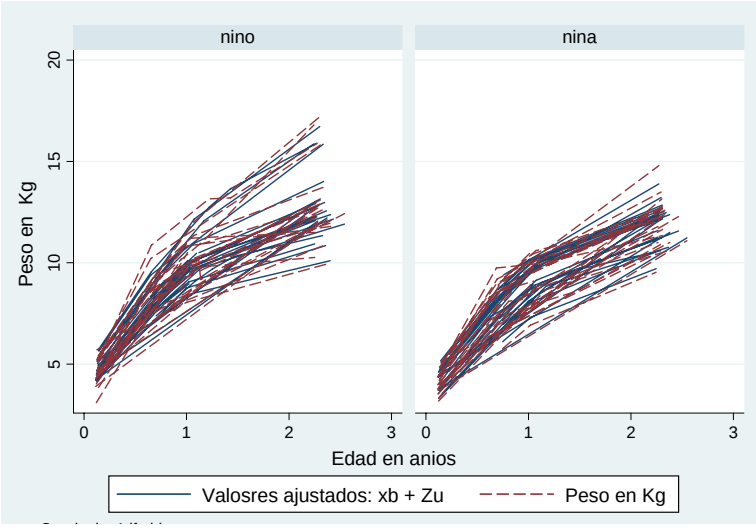
En el modelo anterior, incluimos el efecto del sexo en el crecimiento medio, pero asumimos que la variabilidad en las desviaciones específias de cada niño eran similares para niños y niñaas. Para comprobar esta hipótesis, introducimos la variable sexo dentro de la parte aleatoria del modelo para permitir que las rectas individuales difieran de las rectas medias de cada sexo, es decir,

$$\texttt{peso}_{ij} = (\beta_0 + \beta_3 \texttt{nina}_i + u_i : \texttt{nina}_i) + (\beta_1 + \beta_4 \texttt{nina}_i + v_i : \texttt{nina}_i)\texttt{edad}_{ij} + \beta_2 \texttt{edad}_{ij}^2 + \epsilon_{ij},$$

```
child4=lme(peso~edad*nina+edad2,random =list(id=pdDiag(~nina-1),id=pdDiag(~nina:edad-1)))
child4
  Fixed: peso ~ edad * nina + edad2
  (Intercept)          edad      ninagirl         edad2 edad:ninagirl
    3.7672164     7.7820065    -0.5150251    -1.6399541    -0.2428170

Random effects:
 Formula: ~nina - 1 | id
 Structure: Diagonal
          ninaboy  ninagirl
StdDev: 0.5763516 0.7727271

 Formula: ~nina:edad - 1 | id %in% id
 Structure: Diagonal
        ninaboy:edad ninagirl:edad  Residual
StdDev:    0.7012962     0.2678845 0.5588595
```

En los modelos anteriores tratábamos a todos los ni nos como muestras de una misma población, mientras que en este último modelo permitimos que las ordenadas en el origen y las pendientes aleatorias sean distintas entre niños y niñas. En otras palabras, permitimos que haya heterocedasticidad en los efectos aleatorios debido al sexo (permitimos que las varianza de los efectos aleatorios sean distintas para niños y niñas) (anteriormente vimos como parecía haber más variabilidad en la tasa de creciemiento entre los niños, que entre las niñas).

El siguiente gráfico muestra los valores ajustados y observado para este modelo final.



En todos los modelos que hemos visto, hemos supuesto que los residuos al nivel más bajo ϵ_{ij} son independientes. En el siguiente capítulo veremos cómo es posible relajar esta hipótesis.

Capítulo 5

Extensión del modelo mixto

1. Heterocedasticidad

En los modelos mixtos que hemos visto hasta ahora, hemos asumido que la varianza era constante:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \quad Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$$

Esta suposición es violada en muchas ocasiones:

- Cuando la varianza aumenta al aumentar la magnitud de la variable respuesta.
- Cuando las varianzas son distintas para distintos grupos.
- Cuando la variabilidad depende de una variable explicativa

En estos casos, lo correcto es modelizar la varianza como una función de las covariables, de un factor de agrupación o de la media de la variable repuesta, es decir:

$$Var(\epsilon_{ij}) = \sigma^2 g^2(\boldsymbol{v}_i, \boldsymbol{\theta})$$

donde

$$\begin{array}{rcl} \boldsymbol{v}_i & = & \text{vector de una o más covariables (incluida } E[y_{ij}]) \\ \boldsymbol{\theta} & = & \text{vector de parámetros desconocidos que han de ser estimados} \\ g^2(.) & = & \text{una función conocida} \end{array}$$

La librería `lme4` no incluye esta posibilidad, mientras que `nlme` sí.

Funciones disponibles en el paquete `nlme`

- `varFixed`. La función varianza es $g^2(\boldsymbol{v}_i) = \boldsymbol{v}_i$:

  $$Var(\epsilon_{ij}) = \sigma^2 \boldsymbol{v}_i$$

  la varianza es proporcional a los valores de la covariable (es lo que normalmente entendemos como *weighted least squares*).

- `varIdent`. Corresponde a especificar differentes varianzas en cada nivel de alguna variable de agrupamiento, $\boldsymbol{s}$. La función varianza es $g^2(s_{ij}, \boldsymbol{\theta}) = \theta_{s_{ij}}$:

$$Var(\epsilon_{ij}) = \sigma^2 \theta^2_{s_{ij}}$$

- `varPower`. Generaliza `varFixed`, de modo que la varianza del error es proporcional a una potencia de la covariable: $g^2(\boldsymbol{v}_i) = |\boldsymbol{v}_i|^{2\theta}$, y

$$Var(\epsilon_{ij}) = \sigma^2 |\boldsymbol{v}_i|^{2\theta}$$

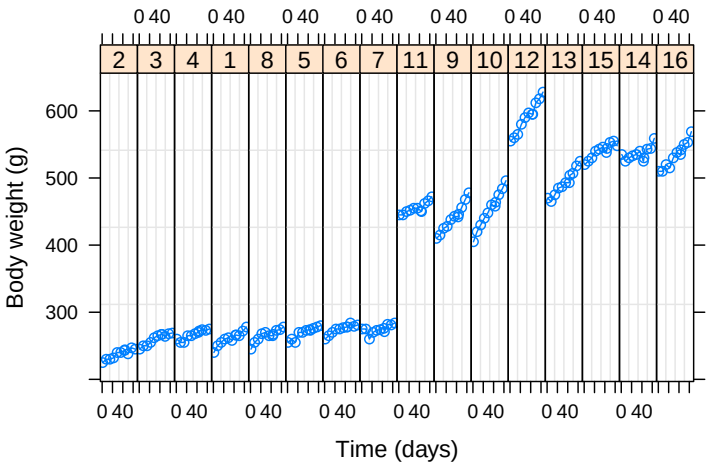
  Un caso particular que se utiliza mucho es cuando la covariable es la media, es decir :

$$Var(\epsilon_{ij}) = \sigma^2 = |\boldsymbol{\mu}_i|^{2\theta}$$

- `varComb`. Permite combinar (como producto) varios tipos de funciones de varianza

Ejemplo:Rats Body Weight

Se mide el peso (en gramos) de 16 ratas cada siete días (nueves veces en total) con una medida extra el día 44. Hay tres grupos de ratas sometidas a tres dietas diferentes.



El gráfico muestra que hay grandes diferencias entre las dietas, y hay una rata en el grupo 2 que tiene un alto peso inicial; además parece que el peso crece de forma lineal con el tiempo, posiblemente con diferentes ordenadas en el origen y pendientes para cada dieta, y con efectos aleatorios para tener en cuenta la variabilidad entre ratas. El modelo vendría dado por:

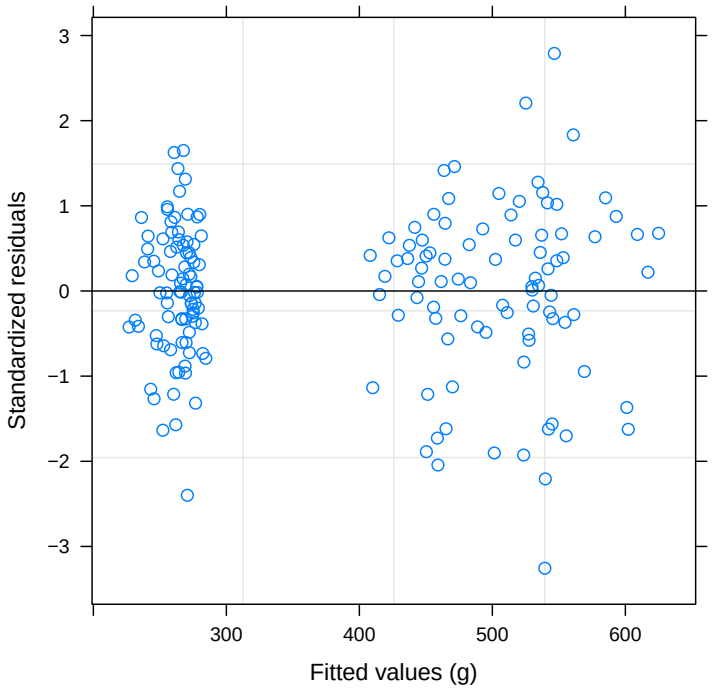
```
rat1=lme(weight~I(Time-33.54545)*Diet, data=BodyWeight,
        random=~I(Time-33.54545)|Rat)
Linear mixed-effects model fit by REML
  Data: BodyWeight
  Log-restricted-likelihood: -575.8599
  Fixed: weight ~ I(Time - 33.54545) * Diet
       (Intercept)       I(Time - 33.54545)                    Diet2
       263.7159075                0.3596391              220.9886336
       Diet3       I(Time - 33.54545):Diet2 I(Time - 33.54545):Diet3
       262.0795441                0.6058392               0.2983375
Random effects:
 Formula: ~I(Time - 33.54545) | Rat
 Structure: General positive-definite, Log-Cholesky parametrization
                   StdDev     Corr
(Intercept)        36.6355111 (Intr)
I(Time - 33.54545)  0.2484113 0.077
Residual            4.4436052
```

Los resultados muestran que hay mucha más variabilidad en las ordenadas en el origen que en las pendientes. Sería de interés contrastar si hay interacción entre dieta y tiempo y si es necesario que las pendientes sean distintas para cada individuo:

```
A=lme(weight~I(Time-33.54545)*Diet, data=BodyWeight,
      random=~I(Time-33.54545)|Rat,method="ML")
B=lme(weight~I(Time-33.54545)+Diet, data=BodyWeight,
      random=~I(Time-33.54545)|Rat,method="ML")
anova(B,A)
  Model df      AIC      BIC    logLik   Test  L.Ratio p-value
B     1  8 1194.219 1219.583 -589.1095
A     2 10 1185.858 1217.563 -582.9291 1 vs 2 12.36078  0.0021

rat1.1=lme(weight~I(Time-33.54545)*Diet, data=BodyWeight,
         random = list(Rat=pdDiag(~I(Time-33.54545))))

anova(rat1.1,rat1)
       Model df      AIC      BIC    logLik   Test    L.Ratio p-value
rat1.1     1  9 1169.792 1198.014 -575.8959
rat1       2 10 1171.720 1203.078 -575.8599 1 vs 2 0.07195702  0.7885


Modelo_NULL=lme(weight~I(Time-33.54545)*Diet, data=BodyWeight,random=~1|Rat)
test=-2*logLik(Modelo_NULL, REML=T) +2*logLik(rat1.1, REML=T)
mean(pchisq(test,df=c(0,1),lower.tail=F))
0
```

Por lo que es necesario pendientes distintsa para cada rata. Aquí hemos asumido que todas las ordenadas en el origen tienen la misma variabilidad (y de manear similar para las pendientes). ¿Cómo ajustarías un modelo que permitiera distinta variabilidad para cada dieta?.

El gráfico de los residuos estandarizados frente a los valores ajustados muestras la heterocedasticidad de los datos.



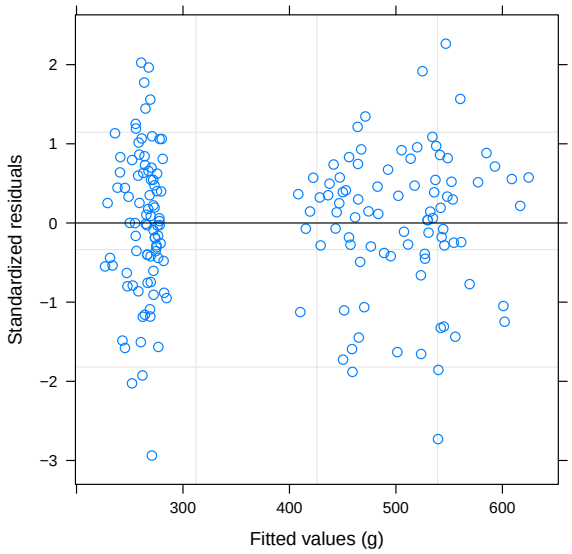
Dado que los valores ajustados son positivos y lejos del cero ⇒ podemos usar la función `VarPower` para modelizar la heterocedasticidad:

```
rat2=update(rat1.1,weights=varPower(form=~fitted(.)))
 Formula: ~I(Time - 33.54545) | Rat
 Structure: Diagonal
        (Intercept) I(Time - 33.54545) Residual
StdDev:    36.61393          0.2435684 0.174975
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
    power
0.5430372
```

Lo primero que observamos es que se ha reducido la varianza residual. Podemos utilzar el test de la razón de verosimilitud para contrastar si $\theta = 0$:

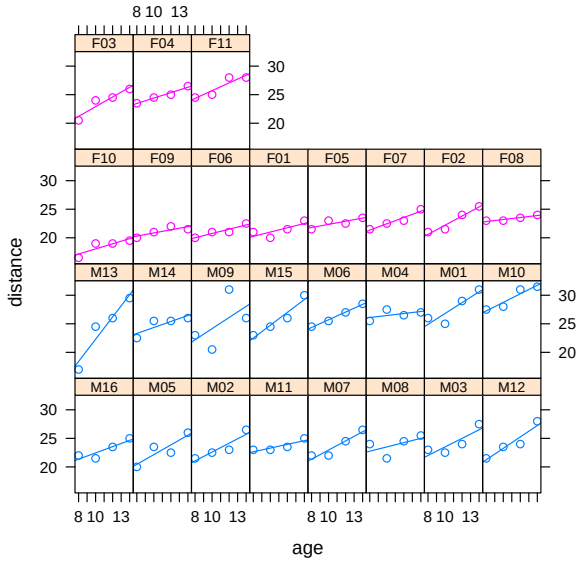
```
anova(rat1.1,rat2)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
rat1.1     1  9 1169.792 1198.014 -575.8959
rat2       2 10 1161.990 1193.348 -570.9949 1 vs 2 9.801898  0.0017
```

Ahora el gráfio de residuos muestra que esta estructura de covarianza representa correctamente la variabilidad.

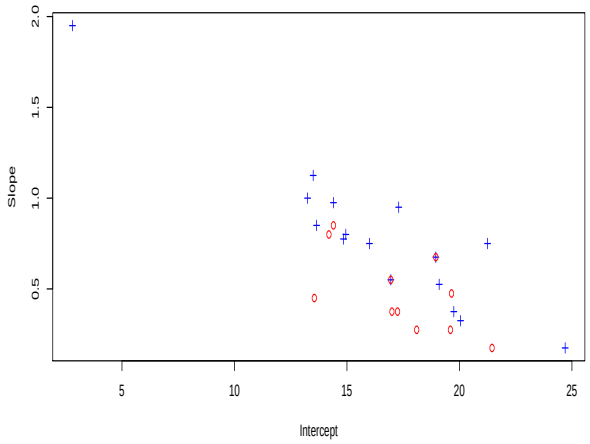


Ejemplo: Orthodont data

Se realizó un estudio a 27 individuos (16 niños y 11 niñas) con edades comprendidas entre los 8 y los 14 años, en el que se midió cada dos años la distancia entre dos puntos del lateral de la cabeza. La siguiente figura muestra los datos:



Un buen resumen de estos datos es hacer un gráfico de la ordenada en el origen y la pendiente para las rectas individuales (ver archivo `ortho.R`). El siguente gráfico muestra que hay diferencias claras entre niños y niñas, y que a ordenadas en el origen más grandes le corresponden pendientes más pequeñas.



Una primera (y burda) forma de saber si las pendiente son diferentes entre ambos sexos será utilizar un test para comparar las dos muestras:

```
 t.test(b[Sex=="Male"], b[Sex=="Female"], var.equal=TRUE)

        Two Sample t-test

t = 2.2624, df = 25, p-value = 0.03261
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.07422099 0.45597382
sample estimates:
mean of x mean of y
0.7843750 0.4795455
```

Ahora consideramos un modelo con efectos aleatorios. Vamos a ajustar dos modelos, uno asumiendo que las pendientes son las mismas para ambos grupos, y otro permitiendo diferentes pendientes.

```
ortho1=lme(distance~Sex + age,random=I(age-11) | Subject,data=Orthodont,
           method="ML")
ortho2= update(ortho1,fixed = distance~Sex * age)
```

y utilizamos la funcion ANOVA para compararlos (por eso hemos usado `method ="ML"`):

```
anova(ortho1,ortho2)
       Model df      AIC      BIC    logLik   Test L.Ratio p-value
ortho1     1  7 446.8352 465.6101 -216.4176
ortho2     2  8 443.8060 465.2630 -213.9030 1 vs 2 5.02921  0.0249
```

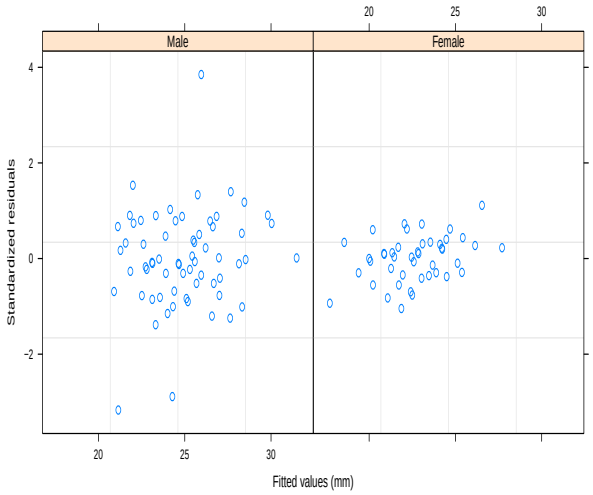
y volvemos a ajustar el modelo mediante REML: y volvemos a ajustar el modelo mediante REML:

```
ortho3=update(ortho2,method="REML")
```

Además podemos contrastar si las ordenadas en el origen y las pendientes son independientes:

```
ortho4=lme(distance~Sex + age,random = pdDiag(~age)),data=Orthodont)
anova(ortho3,ortho4)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
ortho3      1  8 448.5817 469.7368 -216.2908
ortho4      2  6 448.6453 464.5691 -218.3227 1 vs 2 4.063644  0.1311
```

Es importante recordar que el LRT es correcto, aunque hayamos usado REML, ya que los efectos fijos son los mismos en ambos modelos, y la hipótesis altenativa permite que el parámetro de la varianza tome cualquier valor. El gráfico de los residuos frente a los datos ajustados indica que se deberían utilizar diferentes varianzas para ambos sexos.



Podemos conseguirlo mediante la función de varianza `varIdent`

```
 ortho5=update(ortho4, weights = varIdent(form =~1|Sex))
Random effects:
 Formula: ~age | Subject
 Structure: Diagonal
        (Intercept)       age Residual
StdDev:    1.387321 0.1194852 1.731504
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | Sex
 Parameter estimates:
     Male    Female
1.0000000 0.4133617
Number of Observations: 108
Number of Groups: 27
```

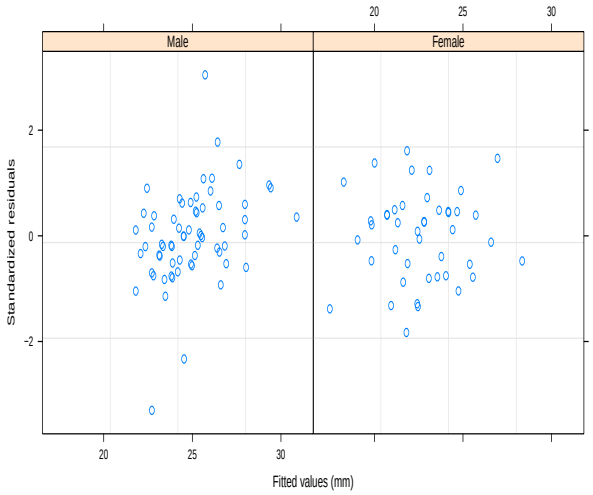
En este caso:

$$Var(\epsilon_{ij}) = \sigma^2 \theta^2_{s_{ij}}$$

por motivos de identificabilidad $\theta^2_{s_{ij}}$ es la razón entre cada varianza y la correspondiente al primer nivel de factor, de modo que:

$$\hat{\sigma}_{nino} = 1{,}73 \quad \hat{\sigma}_{nina} = 1{,}73 \times 0{,}41 = 0{,}71$$

El gráfico de residuos es ahora:



2. Correlación

La idea general es que dos observaciones "intra-individuos" están correlacionadas, y la correlación depende de la distancia temporal o espacial entre ellas. Además, se asume que la estructura de correlación es isotrópica: solo depende de la distancia relativa, no de la posición temporal o espacial de las dos observaciones:

$$Corr(\epsilon_{ij}, \epsilon_{ik}) = h\left(d(p_{ij}, p_{ik}), \boldsymbol{\rho}\right)$$

donde

$$\begin{aligned} \boldsymbol{\rho} &= \text{un vector de parámetros de correlación} \\ h() &= \text{función de correlación conocida} \\ p_{ij}, p_{ik} &= \text{posición de las observaciones } y_{ij},\ y_{ik} \\ d() &= \text{función distancia conocida} \end{aligned}$$

Además, asumimos que $h()$ es continua en $\boldsymbol{\rho}$, toma valores en $[-1, 1]$, y $h(0, \boldsymbol{\rho}) = 1$, de modo que las obsetvaciones a una distancia 0 están perfectamente correladas.

Funciones disponibles en el paquete `nlme`

Hay bastantes funciones disponibles, aquí mostraremos las de uso más frecuente.

- `corAR1`. Estructura autorregresiva de orden 1. Es apropiada para observaciones temporales equiespaciadas.

$$Corr(\epsilon_{ij}, \epsilon_{ik}) = \rho^{|i-j|}$$

  Por ejemplo, para $t = 5$:

$$corr(\boldsymbol{\epsilon}_i) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{pmatrix}$$

- `corCAR1` Es una versión en tiempo continuo de la anterior. Su especificación es la misma, pero el índice temporal de las observaciones puede ser cualquier valor no-negativo.

- `corARMA1`. Corresponde a un proceso $ARMA(p, q)$. Los procesos $AR(p)$ y $MA(q)$ se pueden especificar con esta función.

- `corCompSymm` La estructura es:

$$Corr(\epsilon_{ij}, \epsilon_{ik}) = \begin{cases} 1 & \text{if } j = k \\ \rho & \text{if } j \neq k \end{cases}$$

  es la misma estructura de correlación implcada por un modelo con ordenada en el origen aleatoria por clúster con errores independientes.

- `corSymm` especifica una estructura completamente general con un parámetro para cada posición. Por ejemplo:

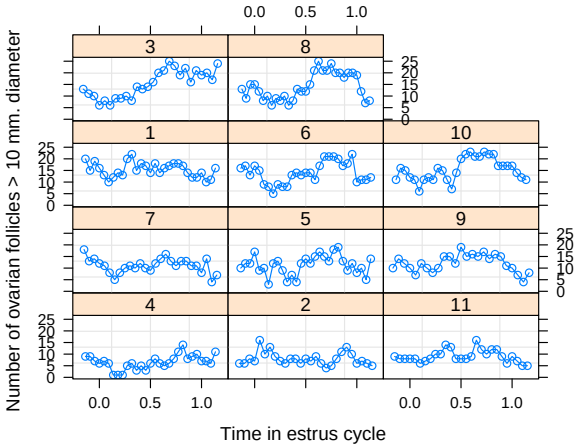
$$corr(\boldsymbol{\epsilon}_i) = \begin{pmatrix} 1 & \rho & \rho_2 & \rho_3 & \rho_4 \\ & 1 & \rho_5 & \rho_6 & \rho_7 \\ & & 1 & \rho_8 & \rho_9 \\ & & & 1 & \rho_{10} \\ & & & & 1 \end{pmatrix}$$

- `corExp` es una estructura espacial, donde

$$Corr(\epsilon_{ij}, \epsilon ik) = exp(-s/\rho) \quad s = d(p_{ij}, p_{ik})$$

Ejemplo: Ovary Data

Los datos con los que vamos a trabajar corresponden a un estudio sobre el número de folículos mayores de 10mm en ovarios de 11 yeguas (mare). Estos datos se registraron diariamente desde 3 días antes de la ovulación hasta 3 días después de la próxima ovulación (un total de 7 veces):



El gráfico muestra un comportamiento cíclico del número de folículos a lo largo del tiempo, por lo que un modelo inicial sería:

$$y_{ij} = (\beta_0 + u_i) + sin(2\pi t_{ij}) + \beta_2 cos(2\pi t_{ij}) + \epsilon_{ij}$$

```
ovary0=lme(follicles~sin(2*pi*Time)+cos(2*pi*Time),random=~1|Mare, data=Ovary)
```

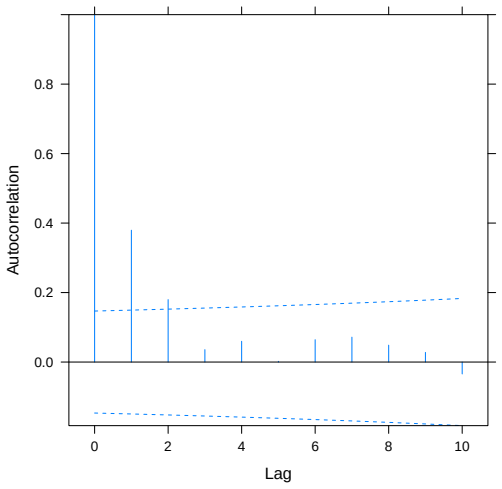
- ¿Es necesario un modelo con ciclos?
- ¿Qué hace el siguiente código?

```
ovary1=lme(follicles~sin(2*pi*Time)+cos(2*pi*Time),random=pdSymm(~sin(2*pi*Time)),
data=Ovary)
ovary2=lme(follicles~sin(2*pi*Time)+cos(2*pi*Time),random=pdDiag(~sin(2*pi*Time)),
data=Ovary)
anova(ovary2,ovary1)
```

Para ver si los herrores están correlados, utilizamos la función de autocorrelación parcial:

Vemos que las autocorrelaciones son significativas en los dos primeros lags, esto sugiere que un modelo AR(1) podría ser apropiado.

```
ovary3=update(ovary2,correlation=corAR1())
Random effects:
 Formula: ~sin(2 * pi * Time) | Mare
 Structure: Diagonal
        (Intercept) sin(2 * pi * Time) Residual
StdDev:    2.858385           1.257977 3.507053
Correlation Structure: AR(1)
 Formula: ~1 | Mare
 Parameter estimate(s):
      Phi
```




```
0.5721866
anova(ovary2,ovary3)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
ovary2     1  6 1638.082 1660.404 -813.0409
ovary3     2  7 1563.448 1589.490 -774.7240 1 vs 2 76.63382  <.0001
intervals(ovary3)
 Correlation structure:
        lower      est.     upper
Phi 0.4325123 0.5721866 0.6850416
```

El patrón de autocorrelaciones mostrado en la figura anterior, también es consistente con un modelo MA(2) (en el que sólo las dos primeras autocorrelaciones son significativas):

```
ovary4=update(ovary2, correlation=corARMA(q=2))
 anova(ovary3,ovary4,test=F)
       Model df      AIC      BIC    logLik
ovary3     1  7 1563.448 1589.490 -774.7240
ovary4     2  8 1571.231 1600.993 -777.6154
```

A la hora de comparar un modelo con estructura AR(1) con otro de estructura MA(2), no podemos utilizar el LRT, ya que los modelos no están anidados. En este caso utilizamos algún criterio de información:

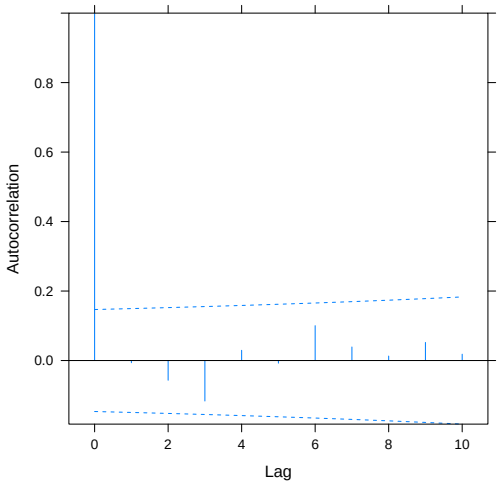
```
ovary4=update(ovary2, correlation=corARMA(q=2))
 anova(ovary3,ovary4,test=F)
       Model df      AIC      BIC    logLik
ovary3     1  7 1563.448 1589.490 -774.7240
ovary4     2  8 1571.231 1600.993 -777.6154
```

A la vista de estos resultados, el modelo AR(1) se ajusta mejor a los datos. Un modelo intermedio entre un AR(1) y un MA(2) es un modelo ARMA(1,1), el cual tiene autocorrelaciones que decaen exponencialmente a partir del lag 2, pero deja más libertad para la correlación al lag 1:

```
ovary5=update(ovary2,correlation=corARMA(p=1,q=1))
```

- ¿Cuál es el mejor modelo?

Una vez elegido el mejor modelo, volvemos a dibujar la función de autocorrelación, pero hemos de especificar que utilice los resudos normalizados:



3. Modelos lineales mixtos generalizados (GLMM)

Antes de hablar de los GLMMs vamos a hacer un breve recordtorio de los que son los modelos lineales generalizados (GLMs).

3.1. Modelos lineales generalizados

El objetivo de cualquier análisis es responder a la pregunta: ¿puede ser la variable de interés predicha por un conjunto de variables explicativas?, esta es la misma pregunta que nos hacemos cuando utilizamos un modelo de regresión lineal, entonces, ¿por qué es necesario utilizar otro tipo de modelos?. La razón fundamental es que para poder utilizar regresión lineal es necesario que la variable respuesta sea continua, y cumpla las hipótesis estándar del modelo lineal (datos Normales, varianza constante, etc.) Si la variable de interés es, por ejemplo binaria e ignoramos este hecho, lo que hacemos es ajustar este modelo:

$$p = Pr(\text{ocurra algo}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

si estimamos los parámetros utilizando el procedimiento `lm()` de R podríamos estar cometiendo dos graves errores:

1. Los valores predichos de la probabilidad podrían estar fuera del intervalo $(0, 1)$.
2. Los intervalos de confianza y los test para ver que variables son significativas están basados en la hipótesis de que los datos viene de una distribución Normal, cosa que no es cierta con datos binarios

Los modelos lineales generalizados (GLMs) extienden el modelo lineal para acomodar las variables respuestas que no siguen una distribución Normal, bajo un enfoque unificado. Es bastante común encontrarse en siatuaciones en las que la variable resuesta no cumple las hipótesis estándar del modelo lineal (datos Normales, varianza constante, etc.), por ejemplo: datos de conteo, datos dicotómicos, datos truncados, etc. Los GLMs se basan en la teoría de Nelder (1962) y Nerder (19989), desde entonces, con los avances del software estadístico, estos modelos se han convertido en una herramienta básica para muchos investigadores.
Hay dos temas fundamentales en la noción de los modelos lineales generalizados: la distribución de la variable respuesta, y cómo el modelo establece la relación entre la media de la variable respuesta y las variables explicativas.

Un concepto importante que unifica todos los GLms es la **familia exponencial de distribuciones** . Todas las distribuciones pertenecientes a la familia exponencial tiene una función de densidad (o de probabilidad) que se puede expresar de la siguiente forma:

$$f(y;\boldsymbol{\theta},\phi) = exp\left\{\frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi)\right\} \tag{5.1}$$

donde, en cada caso, $a(.)$, $b(.)$ y $c(.)$ serán funciones específicas. El parámetro $\boldsymbol{\theta}$ es lo que se llama *parámetro canónico de localización* y ϕ es un *parámetro de dispersión*. La distribucion Binomial, Poisson y Normal (entre otras) son miembros de la familia exponencial.

Componentes de un modelo lineal generalizado

En un modelo de regresión estándar:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(0,\sigma^2\boldsymbol{I}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$$

donde $\boldsymbol{X\beta}$ es una combinación lineal de las variables predictoras llamada *predictor lineal* (el cual se representa como $\boldsymbol{\eta}$), en este caso la media $\boldsymbol{\mu}$ está directamente relacionada con el predictor lineal, ya que en este caso $\boldsymbol{\mu} = \boldsymbol{\eta}$. Usando este modelo sencillo, podemos ver que hay dos componentes en el modelo: la función de probabilidad de la variable respuesta y la estructura lineal del modelo. En general, un modelo lineal generalizado tendra los siguientes componentes:

1. **Componente aleatorio**: $\boldsymbol{y}$ es un vector aleatorio procedente de una distribución que pertenece a la familia exponencial y cuya media es $\boldsymbol{\mu}$.

2. **Componente sistemático**: es el predictor lineal $\boldsymbol{\eta} = \boldsymbol{X\beta}$.

3. **La función *link***: es una función monótona, derivable que establece la relación entre la media y el predictor lineal

   $$\boldsymbol{\eta} = g(\boldsymbol{\mu}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \tag{5.2}$$

   En el caso del modelo de regresión ordinaria, $\boldsymbol{\mu} = \boldsymbol{\eta}$, por lo tanto la función link es la identidad. Hay muchas opciones par la función link. La función **link canonica** es una función que transforma la media en el parámetro canónico $\boldsymbol{\theta}$

   $$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\theta} \Rightarrow g \quad \text{es una función link canónica}$$

Hay muchas opciones par la función link.

La estimación de estos modelos se hace mediante *Iterative Reweighted Least Squares*.

Distribución	Link
Normal	$\boldsymbol{\eta} = \boldsymbol{\mu}$ (identidad)
Binomial	$\boldsymbol{\eta} = \ln\left(\frac{P}{1-P}\right)$ (logistística)
Poisson	$\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$ (logarítmica)
Exponential	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (recíproca)
Gamma	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (recíproca)

Cuadro 1: Funciones link más usadas en los GLMs

4. Conceptos básicos en GLMMs

Los GLMMs son una extensión de los GLMs en los que se añaden efectos aleatorios. Estos modelos son, sin embargo, bastante más difíciles de ajustar que los modelos mixto para datos normales, tanto desde el punto de vita computacional como de la interpretación.

En un modelo GLM:

$$\boldsymbol{\eta} = \boldsymbol{X\beta} \Rightarrow \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

En un MM:

$$\boldsymbol{u} \sim N(0, \boldsymbol{G}) \quad \boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \boldsymbol{R}) \Rightarrow \boldsymbol{y}|\boldsymbol{u} \sim B(m, p(\boldsymbol{u}));\ P(\lambda(\boldsymbol{u}))$$

Para poder ajustar el modelo necesitamos calcular la función de verosimilitud de $\boldsymbol{y}$ y esto supone el cálculo de una integral que no se puede resolver de forma analítica. Hay tres formas de ajustar un GLMM:

1. Máxima verosimilitud
2. Métodos bayesianos (MCMC)
3. Generalized Estimating Equations (GEE)

Nosotros nos vamos a centrar en el enfoque mediante máxima verosimilitud. Este método utiliza una aproximación numérica para calcular la integral (mediante cuadratura Gaussiana o aproximación de Laplace) o también puede eviar el cálculo de la integral mediante el uso de la quasi-verosimilitud.

En `R` la función que ajusta un GLMM es la función `glmer` del paquete `lme4`, funciona como `lmer`, pero se añade el argmento `family=` y `link`, al igual que en la función `glm`.

4.1. GLMMs para datos binarios: Cuidados prenatales en Bangladesh

Los datos con los que vamos a ilustrar el uso de GLMMs para datos binarios, corresponde a una encuesta de salud de mujeres en edad reproductiva (13-49 años) en Bangladesh. La variable respuesta es binaria y corresponde a si la mujer ha recibido o no cuidados prenatales por parte de personal cualificado al menos una vez antes del nacimiento más reciente (la encuesta se realizó sólo a mujeres que había tenido su último hijo en los últimos 5 años). Los datos tienen una estructura multinivel ya que las 5366 mujeres están anidadas en 361 comunidades. En areas rurales, la comunidad corresponde al pueblo, y en areas urbanas a areas censales.
Se consideran variables explicativas tanto a nivel 1 como 2.

- `comm`: identificador de la comunidad
- `wonid`: identificador de la mujer
- `antemed`: Tooma valor 1 si la mujer ha recibido cuidados prenatales, y 0 en otro casp.
- `bord`: El orden que ocupa el último hijo
- `mage`: Edad de la madre en el nacimento del último hijo.
- `urban`: Toma valor 1 si la comunidad es urbana y 0 en otro caso.
- `meduc`: Nivel educativo de la madre (1 = ninguno, 2 = primaria, 3 = secundaria o más).
- `islam`: Toma valor 1 si la religión es el Islam y 0 en otro caso.
- `wealth`: Nivel de riqueza del hogar, toma valores de 1 a 5 (siendo 1 el más pobre).

En el caso de un modelo con ordenada en el origen aleatoria para datos normales, tenemos:

$$E[y_{ij}|u_i] = \beta_0 + u_i + \epsilon_{ij} \quad u_i \sim N(0, \sigma_u^2) \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

En el caso de una variable respuesta binaria, $E[y_{ij}|u_i] = p_{ij} = P[y_{ij} = 1]$, de modo que:

$$logit(p_{ij}) = log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + u_i$$

En este caso e^{β_0} es el odds de que $y = 1$ para $u = 0$, y u_i es el efecto aleatorio de grupo cuya varianza representa la varianza resual entre grupos.

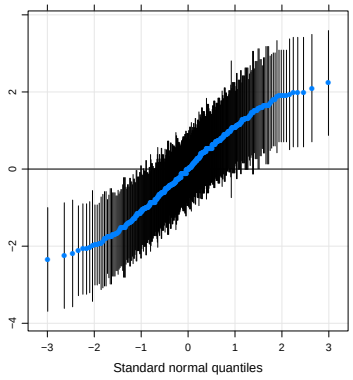
```
fit <- glmer(antemed ~ (1 | comm), family = binomial("logit"), data = mujeres)
summary(fit)
Formula: antemed ~ (1 | comm)
   Data: mujeres
  AIC  BIC logLik deviance
 6640 6653  -3318     6636
Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.4644   1.2101
Number of obs: 5366, groups: comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14811    0.07136   2.075   0.0379 *
```

En este caso el log-odds de recibir cuidado prenatal en una comunidad promedio (con $u_i = 0$) se estima como $\hat{\beta}_0 = 0{,}148$, y el intercept para cada comunidad es $0{,}148 + u_i$, y la varianza estimada es $\sigma_u^2 = 1{,}464$. Para contrastar si es necesario el efecto aleatorio:

```
fit0=glm(antemed~1,family = binomial("logit"), data = mujeres)
test=-2*logLik(fit0) +2*logLik(fit)
mean(pchisq(test,df=c(0,1),lower.tail=F))
0
```

Para examinar los residuos a nivel 2, es decir, $\hat{u}_i$, vamos a dibujarlos:

```
library(lattice)
qqmath(ranef(fit, postVar = TRUE), strip = FALSE)$comm
```



El gráfico muestra los residuos estimados para las 361 comunidades en la muestra. Para un alto número de comunidades, el intervalo de confianza no corta a la línea horizontal, indicando que los cuidados prenarales en esas comunidades está significativamente por encima o por abajo de la media, además los intervalos de confianza son bastante anchos, esto es debido a que las muestra dentro de las comunidades no son muy grandes.

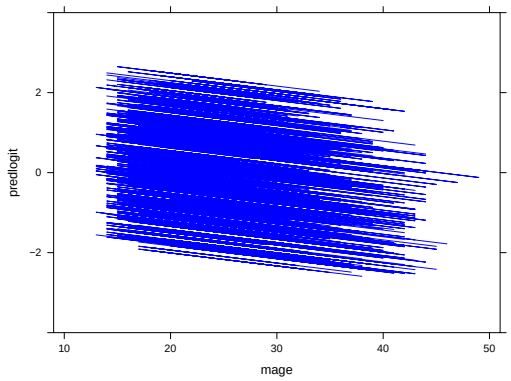
A continuación incluimos la variable correspondiente a la edad de la madre, pero antes la centramos:

```
magec=mujeres$mage-mean(mujeres$mage)
mujeres=cbind(mujeres,magec)
fit2=glmer(antemed~magec+(1|comm), family=binomial("logit"),data=mujeres)
summary(fit2)
Formula: antemed ~ magec + (1 | comm)
   Data: mujeres
  AIC  BIC logLik deviance
 6603 6623  -3299     6597
Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.4622   1.2092
Number of obs: 5366, groups: comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.144680   0.071365   2.027   0.0426 *
magec       -0.032394   0.005163  -6.275 3.51e-10 ***
```

Observamos que la varianza entre comunidades cambia poco, lo que significa que la distribución de la edad materna en las comunidades es similar. La ecuación de la recta de regresión que expresa la relación entre el log-odds de tener cuidado prenatal y la edad materna viene dadda por:

$$log\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = 0{,}144 - 0{,}032\texttt{magec}_{ij}$$

La linea ajustada para una comunidad dada se diferencia de la linea media en una cantidad u_i, de modo que las rectas para las comunidades son paralelas entre si.



De modo que para una mujer de 22 años el log-odds de recibir cuidados prenatales se encuantra entre -2.2 y 2.5 dependiendo de en que comunidad se encuentre. Esto se traduce en términos de probabilidades a $e^{-2,2}/(1+e^{-2,2}) = 0{,}10$ hasta $e^{2,5}/(1+e^{2,5}) = 0{,}92$.

Ahora incluimos las variables correspondientes al nivel de riqueza de las familias (centrada) y la educación:

```
mujeres$meduc=factor(mujeres$meduc)
wealthc=mujeres$wealth-mean(mujeres$wealth)
mujeres=cbind(mujeres,wealthc)

fit3=glmer(antemed~magec+wealthc+meduc+(1|comm), family=binomial("logit"),data=mujeres)
summary(fit3)
  AIC  BIC logLik deviance
 5993 6033  -2991     5981
Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 0.86794  0.93163
Number of obs: 5366, groups: comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.490564   0.079530  -6.168 6.90e-10 ***
magec       -0.005136   0.005662  -0.907    0.364
wealthc      0.402846   0.029408  13.699  < 2e-16 ***
meduc2       0.544873   0.084291   6.464 1.02e-10 ***
meduc3       1.305848   0.097153  13.441  < 2e-16 ***
```

En este caso el coeficiente de correlación intra-clase (ICC) es:

$$\frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

En el caso de la distribución logística estandard la varianza es $p^2/3 = 3{,}29$, de modo que en este caso $ICC = 0{,}868/(0{,}868 + 3{,}29) = 0{,}21$, es decir, el 21 % de la variabilidad residual en la propensión a tener cuidados prenatales es atribuible a características no observadas de la comunidad (cuando no incluimos variables explicativas esa cantidad se eleva al 31 %).

En el caso de un GLMM para datos binarios:

$$p_{ij} = \frac{exp(\beta_0 + \ldots + \beta_k x_k + u_i)}{1 + exp(\beta_0 + \ldots + \beta_k x_k + u_i)}$$

Si calculamos las probabilidades cuando $u_i = 0$ lo que obtenemos es el valor mediano de la probabilidad cuando las variables explicativas toman el valor medio a lo largo de los grupos (si las probabilidades están entre 0.2 y 0.8, el valor medio y mediano se parecerá mucho). En concreto:

$$p_{ij} = \frac{exp(z_{ij})}{1 + exp(z_{ij})}$$

donde

$$z_{ij} = -0{,}491 - 0{,}005\texttt{magec}_{ij} + 0{,}403\texttt{wealthc}_{ij} + 0{,}545\texttt{meduc2}_{ij} + 1{,}306\texttt{meduc3}_{ij}$$

Hasta ahora hemos considerado que todos los coeficientes de las variables explicativas eran similares para los distintos grupos, ahora vamos a suponer que la pendiente de `wealthc` es distinta para cada comunidad

```
fit4=glmer(antemed ~ magec+meduc+wealthc+(1+wealthc|comm),data=mujeres,
     family = binomial("logit"))
summary(fit4)
Random effects:
 Groups Name        Variance Std.Dev. Corr
 comm   (Intercept) 0.847061 0.92036
        wealthc     0.015466 0.12436  -0.951
Number of obs: 5366, groups: comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.491733   0.079255  -6.204 5.49e-10 ***
magec       -0.005178   0.005669  -0.913    0.361
meduc2       0.542194   0.084655   6.405 1.51e-10 ***
meduc3       1.300867   0.096648  13.460  < 2e-16 ***
wealthc      0.409055   0.030136  13.573  < 2e-16 ***
```

Ahora contrastamos si la ordenada en el origen y la pendiente son independientes:

```
fit5=glmer(antemed~magec+meduc+wealthc+(1|comm)+(0+wealthc|comm),data=mujeres,
     family=binomial("logit"))
anova(fit5,fit4)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit5  7 5995.2 6041.3 -2990.6
fit4  8 5987.3 6040.0 -2985.7 9.8704      1    0.00168 **
```

Por lo tanto no lo son, y es necesario incluir una pendiente para cada comunidad. El efecto de la riqueza en el log-odds de recibir cuidados prenatales en la comunidad i es $0{,}407 + \hat{v}_i$ (donde v_i es el efecto aleatorio que interacciona con la riqueza), y la varianza entre comunidades en el efecto de la riqueza es 0,015. El que la covarianza entre la ordenada en el origen y la pendiente sea negativa, $\hat{\sigma}_{uv} = -0{,}951 * 0{,}124 * 0{,}92 = -0{,}108$ implica que la comunidades con un nivel de cuidados prentales superiores a la media tienden a tener un efecto de la riqueza inferior a, es decir, en comunidades con un alto porcentaje de cuidados prenatales, este tiende a descender con la riqueza.

La ecuación de la recta de regresión para la comunidad i, para una mujer con edad media (`magec`=0), sin estudios (`meduc2`=`meduc3`=0) es:

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = (-0{,}492 + \hat{u}_i) + (0{,}409 + \hat{v}_i)\texttt{weathc}_{ij}$$

Para obtener la recta para mujeres con diferentes edades o niveles de educación, solo la ordenada en el origen cambiaría, por ejemplo para una mujer con educación primaria la ordenada en el origen pasaría de $-0{,}492$ a $-0{,}492 + 0{,}5421 = 0{,}05$.

Por último nos quedaría incluir las variables a nivel 2 (contextuales), en este caso hay solo una, `urban`:

```
fit6=glmer(antemed ~ magec+meduc+wealthc+urban+(1+wealthc|comm),data=mujeres,
     family = binomial("logit"))
summary(fit6)
Random effects:
 Groups Name        Variance Std.Dev. Corr
 comm   (Intercept) 0.649855 0.80614
        wealthc     0.027159 0.16480  -1.000
Number of obs: 5366, groups: comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.863744   0.086411  -9.996  < 2e-16 ***
magec       -0.004909   0.005678  -0.864    0.387
meduc2       0.569806   0.084905   6.711 1.93e-11 ***
meduc3       1.351318   0.096925  13.942  < 2e-16 ***
wealthc      0.355048   0.030905  11.489  < 2e-16 ***
urban        1.032669   0.112939   9.144  < 2e-16 ***
```

Concluimos que las mujeres que viven en comunidades urbanas son más propensas a buscar cuidados prenatales que las que viven en zonas rurales. Además la inclusión de esta variable ha descendido la variabilidad en la ordenada en el origen entre comunidades.

4.2. Ejemplo: Ciervos

Vamos a trabajar con datos que se encuantran en el archivo `ciervos.txt`. El objetivo es relacionar a longitud del ciervo (macho) `longi`con la probabilidad de infección por un parásito. Los datos fueron recogidos de entre 22 `Granjas` en España, y el número de ciervos por granja varía entre 3 y 83, con un total de 447 ciervos.

Podemos empezar ajustando un modelo glm:

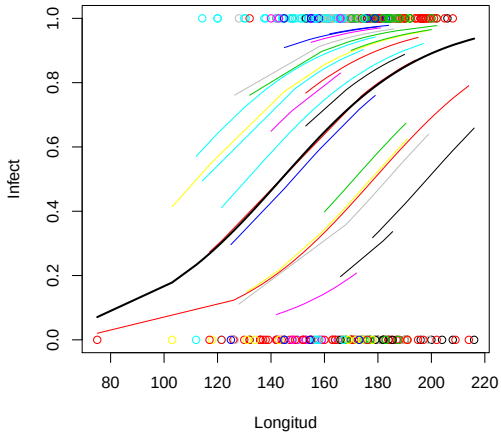
```
ciervos1=glm(infect~Longi+Granga,family=binomial,data=ciervos)
summary(ciervos1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.983e+00  1.594e+00  -5.008 5.49e-07 ***
Longi        3.907e-02  7.586e-03   5.151 2.59e-07 ***
GranjaAU     3.216e+00  7.971e-01   4.034 5.47e-05 ***
GranjaBA     4.294e+00  1.195e+00   3.593 0.000326 ***
GranjaBE     2.023e+01  1.855e+03   0.011 0.991297
GranjaCB     2.826e+00  7.803e-01   3.622 0.000292 ***
```

$\hat{\beta}$ para `Longi` es 0.0391, es decir, por cada 10cm adicionales que mida el ciervo, la posibilidad (odds) de infección se multiplica por $e^{10\times 0{,}0391} = 1{,}47$, para ciervos de una granja en particular. El modelo nos da una estimación de $Pr[infect|longi]$ para cada una de las 24 granjas, si queremos saber la $P[infect]$ necesitamos conocer el efecto de la granja en la que está el ciervo, entonces, ¿qué podemos decir de las granjas que no están en el estudio?: nada, a menos que supongamos que las 24 granjas del estudio son una muestra de la población de granjas:

$$\begin{aligned} y_{ij} &\sim B(1, p_{ij}) \quad i = 1, \ldots, 24. \quad j = 1, \ldots, n_i \\ log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 + \beta_1 L_{ij} + u_i \\ u_i &\sim N(0, \sigma_u^2) \end{aligned}$$

En R:

```
ciervos2=glmer(infect~Longi+(1|Granja),data=ciervos,family=binomial)
ciervos2
Random effects:
 Groups Name        Variance Std.Dev.
 Granja (Intercept) 3.0698   1.7521
Number of obs: 447, groups: Granja, 22

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.032335   1.272936  -3.953 7.71e-05 ***
Longi        0.037390   0.007199   5.194 2.06e-07 ***
```



La figura muestra las probabilidades predichas para cada granja (en color) y la probabilidad mediana (negro). Si queremos contrastar si la longitud es significativa:

```
ciervos3=update(ciervos2,.~.-Longi,ciervos)
anova(ciervos3,ciervos2)
         Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
ciervos3  2 464.26 472.46 -230.13
ciervos2  3 436.81 449.12 -215.41 29.443      1  5.758e-08 ***
```

Bibliografía

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods.* Sage Publications

Dansereau, F., Cho, J., and Yammarino, F.J. (2006). Avoiding the "Fallacy of the Wong Level". A within and between analysis approach. *Group Organization Management*, 31:536–577.

Goldstein, H. (2002). *Multilevel Statistical Models.* New York: John Wiley & Sons

Jargowsky, P.A. (2005). The ecological fallacy. *Encyclopedia of Social Measurement*, 1:715–722.

McCulloch, C.E., Searle, S.R. and Neuhaus,J.M. (2008). *Generalized, linear, and mixed models.* New York: John Wiley & Sons

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman & Hall

Nelder, J.A. and Wedderburn, R.W.M (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135:370-385

Patterson, H.D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58:545–554

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-effects models in S and S-Plus.* Springer

Rabe-Hesketh, S. and Skrondal, A. (2005). *Multilevel and Longitudinal Modeling Using Stata.* Stata Press

Satterthwaite, F. E., (1941). Synthesis of variance. *Psychometrika,*, 6:309–316.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components.* New York: John Wiley & Sons

Self, S.G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, 82:605–610.

Singer, J.D. and Willet, J.B. (2003). *Applied Longitudinal Data Analysis.* Oxford University Press

Snijers, T.A.B. (2003). Fixed and random effects. *Encyclopedia of Statistics in Behavioral Sciences*, 2:664–665.

Subramanian, S.V., Hones, K. Kaddour, A. and Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Jorunal of Epidemiology*, 38:342–360.

Tinklin, T. (2000). The influence of social background on application and entry to higher educationin Scotland: a multilevel analysis. *Higher Education Quarterly*, 54(4):343–385.

West, B.T., Welch, K.B. and Galecki, A.T. (2007). *Linear mixed models: a practical guide using statistical software* . CRC Press

Wretenberg, P. and Arborelius, U.P. and Lindberg, F (1993). The effects of a pneumatic stool and a one-legged stool on lower limb joint load and muscular activity during sitting and rising. *Ergonomics*, 36:519–535.