

Fundamentos de ciencia de datos con R

Gema Fernández-Avilés y José-María Montero

2023-06-02

Índice general

Prefacio	5
¡Hola mundo!	5
¿Por qué este libro?	6
¿A quién va dirigido?	7
El paquete CDR	8
¿Por qué R?	8
Agradecimientos	9
1. Modelos mixtos	11
1.1. Conceptos básicos	11
1.2. Formulación del modelo con efectos aleatorios o modelos mixtos	15
1.3. Procedimiento con R para ajustar modelos mixtos	17
1.4. Caso práctico	17

Prefacio

¡Hola mundo!

El siglo XXI está siendo testigo de grandes cambios vertiginosos en el contexto social y tecnológico, entre otros. Los tiempos han cambiado, la sociedad se ha globalizado y “exige” respuestas inmediatas a problemas muy complejos. Vivimos en el mundo de la **información**, de los **datos**, o mejor, de las **bases de datos masivas**, y los ciudadanos y, sobre todo, las empresas y los gobiernos, dirigen su mirada hacia el mundo científico para que les ayude a “**oír las historias**” que cuentan esos datos acerca de la realidad de la que han sido extraídos. Y dado su enorme volumen y sofisticación (en el nuevo mundo las imágenes y los textos, por ejemplo, también son datos), exigen algoritmos de nueva generación en el campo del *machine learning*, o incluso del *deep learning*, para “oír las historias” que cuentan. No parecen mirar al “antiguo” investigador científico, sino al “nuevo” *científico de datos*.

Ello, inevitablemente, se traduce en la necesidad de profesionales con una gran capacidad de adaptación a este nuevo paradigma: los científicos de datos, también llamados por algunos los “nuevos hombres del Renacimiento”, para lo cual las Universidades y demás instituciones educativas especializadas se apresuran a incluir el grado de Ciencia de Datos en su oferta educativa y a ofrecer seminarios de software estadístico de acceso abierto para sus estudiantes de primeros cursos.

Con la emergencia de la nueva sociedad, en la que el manejo de la ingente cantidad de información que genera se hace absolutamente necesario para circular por ella, la **Ciencia de Datos** ha venido para quedarse. Sin embargo, el mundo de la Ciencia de Datos es cualquier cosa menos sencillo. En él, cualquier ayuda, cualquier guía es bienvenida. Por ello, es muy recomendable que la persona que se quiera introducir en él, sea con fines de investigación o con fines profesionales, se agarre de la mano de un guía especializado que le lleve, de una manera amena, comprensible y eficiente, desde el planteamiento de su problema y la captura de la información necesaria para poderle dar una solución, hasta la redacción de las conclusiones finales que ha obtenido con los modernos informes reproducibles colaborativos. Y como en la parte central de ese camino tendrá que luchar con grandes gigantes (en la actualidad denominados técnicas estadísticas y algoritmos), el guía tendrá que explicarle, de manera sencilla y amena, en qué consiste la lucha (las técnicas y los algoritmos) y cómo llegar a la victoria lo más rápido posible, enseñándole a moverse por el mundo del software estadístico, en nuestro caso **R**, que le permitirá realizar los cálculos necesarios para vencer al problema planteado a una velocidad vertiginosa.

En resumen, la información masiva y el moderno tratamiento estadístico de la misma son la “mano invisible” que gobierna la sociedad del siglo XXI, y este manual pretende ser el guía anteriormente mencionado que le llevará de la mano cuando quiera caminar por ella.

¿Por qué este libro?

Lo dicho anteriormente ya justifica por sí solo la aparición de este manual. Afortunadamente, no es el primero en la materia, pues son ya bastantes los materiales de calidad publicados sobre Ciencia de Datos. Sin embargo, quizás, éste pueda ser considerado el más completo. Y ello por varias razones.

La primera es su **completitud**: este manual lleva de la mano al lector desde el planteamiento del problema hasta el informe que contiene la solución al mismo; o desde no saber qué hacer con la información de la que dispone, hasta ser capaz de transformar tales bases de datos masivas, y casi imposibles de manejar, en respuestas a problemas fundamentales de una empresa, institución o cualquier agente social.

La segunda es su **amplitud temática**:

- (i) Parte de las dos primeras preguntas que un neófito se puede hacer sobre esta temática: ¿qué es eso de la Ciencia de Datos que está en boca de todos? Y, ¿qué diablos es **R** y cómo funciona?
- (ii) Enseña cómo moverse en la jungla del *Big Data* y de los “nuevos” tipos de datos, siempre bajo el paraguas de la ética de los datos y del buen gobierno de dichos datos.
- (iii) Muestra al lector cómo obtener conocimiento de la oscuridad del enorme banco de información a su disposición, que no sabe cómo abordar ni manejar.
- (iv) No deja a nadie atrás, y de forma previa al contenido central del manual (las técnicas de Ciencia de Datos), incluye unas breves, pero magníficas, secciones sobre los rudimentos de la probabilidad, la inferencia estadística y el muestreo, para aquéllos no familiarizados con estas cuestiones.
- (v) Aborda una treintena de técnicas de Ciencia de Datos en el ámbito de la modelización, análisis de datos cualitativos, discriminación, *machine learning* supervisado y no supervisado, con especial incidencia en las tareas de clasificación y clusterización -así como, en el caso no supervisado, de reducción de la dimensionalidad, escalamiento multidimensional y análisis de correspondencias-, *deep learning*, análisis de datos textuales y de redes, y, finalmente, ciencia de datos espaciales (desde las perspectivas de la geoestadística, la econometría espacial y los procesos de punto).
- (vi) Hace especial hincapié en la reproducibilidad en tiempo real (o no) entre los distintos miembros de un equipo (sea universitario, empresarial, o del tipo que sea) y en la difusión de los resultados obtenidos, enseñando al lector cómo generar informes reproducibles mediante RMarkdown y documentos Quarto o en otros modernos formatos.
- (vii) Dedica un capítulo a la creación de aplicaciones web interactivas (con Shiny).

Índice general

7

- (viii) Para aquéllos con pasión por la codificación, y que quieran compartir código y colaborar con otros desarrolladores, este manual aborda la gestión rápida y eficaz de proyectos (del tamaño que sean) mediante Git, un sistema de control de versiones distribuido, gratuito y de código abierto, y GitHub, un servicio de alojamiento de repositorios Git del cual, aquellos no familiarizados con la cuestión de la codificación, o con aversión a ella, podrán tomar el código que necesitan.
- (ix) Muestra al lector los primeros pasos para iniciarse en el geoprocесamiento en la nube.
- (x) Y, finalmente, aborda más de una docena de casos de uso (en medicina, periodismo, economía, criminología, marketing, moda, demanda de electricidad, cambio climático, reconocimiento de patrones en la forma de tuitear...) que ilustran la puesta en práctica de todos los conocimientos anteriormente adquiridos.

La cuarta razón es que todo lo que el lector aprende en este manual lo puede reproducir y poner en práctica inmediatamente con **R**, puesto que el manual está trufado de *chunks* (o trozos de código **R**) que no tiene más que cortar y pegar para reproducir los ejemplos que se muestran en el libro, cuyos datos están en el paquete CDR; o utilizar dichas *chunks* para abordar el problema que le ocupa con los datos que tenga a su disposición. Una buena razón, sin duda. Por consiguiente, el manual es una buena combinación “teoría-práctica-software” que permite abordar cualquier problema que el científico de datos se plante en cualquier disciplina o situación empresarial, médica, periodística...

La quinta es su **variedad de perspectivas**. Son **más de 40 los participantes** en este manual. Algunos de ellos, prestigiosos profesores universitarios; otros, destacados miembros de instituciones públicas; otros, CEOs de empresas en la órbita de la ciencia de datos; otros, *big names* del mundo de **R** software... El manual es, sin duda, un magnífico ejemplo de colaboración Universidad-Empresa para buscar soluciones a los problemas de las sociedades modernas.

¿A quién va dirigido?

Fundamentos de ciencia de datos con R está dirigido a todos aquellos que desean desarrollar las habilidades necesarias para abordar proyectos complejos de Ciencia de Datos y “pensar con datos” (como lo acuñó Diane Lambert, de Google). El deseo de resolver problemas utilizando datos es su piedra angular. Por tanto, como se avanzó anteriormente, este manual no deja a nadie atrás, y lo único que requiere es “el deseo de resolver problemas utilizando datos”. No excluye ninguna disciplina, no excluye a las personas que no tengan un elevado nivel de análisis estadístico de datos, no excluye a nadie. Se ha procurado una combinación de rigor y sencillez, y de teoría y práctica, todo ello con sus correspondientes códigos en **R**, que satisfaga tanto a los más exigentes como a los principiantes.

También está destinado a todos aquellos que quieran sustituir la navegación por la web (la búsqueda del video, publicación de blog o tutorial *online* que solucione su problema –frustración tras frustración por la falta de consistencia, rigor e integridad de dichos materiales, así como por su sesgo hacia paquetes singulares para la implementación de las cuestiones que tratan–), por

una “**biblia de la ciencia de datos**” rigurosa pero sencilla, práctica y de aplicación inmediata sin ser ni un experto estadístico ni un experto informático.

Pero si a alguien está destinado especialmente, es a la comunidad hispano hablante. Este manual es un guiño a dicha comunidad, para que tenga a su disposición, en su lengua nativa, uno de los mejores manuales de Ciencia de Datos de la actualidad.

El paquete CDR



El paquete **CDR** contiene la mayoría de conjuntos de datos utilizados en este libro que no están disponibles en otros paquetes. Para instalarlo use la función `install_github()` del paquete `remotes`.

```
# este comando sólo necesita ser ejecutado una vez
# si el paquete remotes no está instalado, descomentar para instalarlo

# install.packages("remotes")
remotes::install_github("cdr-book/CDR")
```

La lista de todos los conjuntos de datos puede obtenerse haciendo `data()`.

```
library('CDR')
data(package = "CDR")
```

Este paquete ayudará al lector a reproducir todos los ejemplos del libro. De acuerdo con las mejores prácticas en **R**, el paquete **CDR** sólo contiene los datos utilizados en el libro.

¿Por qué R?

R es un lenguaje de código abierto para computación estadística que se ha consolidado entre la comunidad científica internacional, en las últimas dos décadas, como una herramienta de primer

Índice general

9

nivel, consolidándose como líder permanente en el ámbito de la implementación de metodologías estadísticas para el análisis de datos. La utilidad de **R** para la Ciencia de Datos deriva de un fantástico ecosistema de paquetes (activo y en crecimiento), así como de un buen elenco de otros excelentes recursos: libros, manuales, *blogs*, foros y *chats* interactivos en las redes sociales, y una gran comunidad dispuesta a colaborar, a orientar y a resolver diferentes cuestiones relacionadas con **R**.

Por otra parte, **R** es el lenguaje estadístico y de análisis de datos más utilizado en la mayoría de los entornos académicos y, cómo no, por una larga lista de importantes empresas, entre las que se cuentan Facebook (análisis de patrones de comportamientos relacionado con actualizaciones de estado e imágenes de perfil), Google (para la efectividad de la publicidad y la previsión económica), Twitter (visualización de datos y agrupación semántica), Microsoft (adquirió la empresa Revolution R), Uber (análisis estadístico), Airbnb (ciencia de datos), IBM (se unió al grupo del consorcio R), New York Times (visualización)...

La comunidad **R** también es particularmente generosa e inclusiva, y hay grupos increíbles, como *R-Ladies* y *Minority R Users*, diseñados para ayudar a garantizar que todos aprendan y usen las capacidades de **R**.

Agradecimientos

No queremos dar por finalizado este prefacio sin agradecer a los 44 autores participantes en esta obra su esfuerzo por condensar, en no más de 20 páginas, la teoría, práctica y tratamiento informático de la parte de la Ciencia de Datos que les fue encargada. Y no sólo eso; el “más difícil todavía” fue que debían dirigirse a un abanico de potenciales lectores tan grande como personas haya con “el deseo de resolver problemas utilizando datos”. Era misión imposible. Sin embargo, a la vista del resultado, ha sido misión cumplida. El esfuerzo mereció la pena.

Además, nos gustaría agradecer el apoyo incondicional recibido por (en orden alfabético): Itzcoatl Bueno, Ismael Caballero, Emilio L. Cano, Diego Henangómez, Ricardo Pérez, Manuel Vargas y Jorge Velasco.

También queremos poner de manifiesto que la edición de este texto ha sido financiada por diversos entes de la Universidad de Castilla-La Mancha. En su mayor parte, por el **Máster en Data Science y Business Analytics (con R software)** (a través de la orgánica: 02040M0280), pero también por la Facultad de Ciencias Jurídicas y Sociales de Toledo (a través de su contrato programa: orgánica 00440710), el Departamento de Economía Aplicada I (mediante sus fondos departamentales, DEAI 00421I126) y el Grupo de Investigación Economía Aplicada y Métodos Cuantitativos (que ha dedicado parte de sus fondos a la edición de esta obra, orgánica 01110G3044-2023-GRIN-34336).

A todos, eternamente agradecidos por ayudarnos en este reto de transformar la oscuridad en conocimiento, de convertir en una ciencia y en un arte la difícil tarea de sacar valor de los datos, el petróleo del futuro. Quizás en este momento no seamos conscientes de que hemos puesto nuestro granito de arena a la ciencia que, a buen seguro, juegue uno de los papeles más importantes de este siglo, caracterizado por el predominio de la información. Una ciencia, la Ciencia de Datos, que combina el análisis estadístico de datos, la algoritmia y el conocimiento del

negocio para sacar valor del bien más abundante de la sociedad en la que vivimos: la información. Una disciplina cuyo dominio caracteriza a los científicos de datos (también denominados los nuevos personajes del Renacimiento), profesión que ya fue calificada hace más de veinte años en la *Harvard Business Review* y en *The New York Times*, entre otros, como la “más sexy del siglo XXI”.

Nota

Este manual está publicado por [McGraw Hill](#). Las copias físicas están disponibles en [McGraw Hill](#). La versión *online* se puede leer de forma gratuita en <https://cdr-book.github.io/> y tiene la [licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional](#).

Si tiene algún comentario o sugerencia, no dude en contactar con los editores y los autores. ¡Gracias!

Capítulo 1

Modelos mixtos

María Durbán^a y Víctor Casero-Alonso^b

^aUniversidad Carlos III de Madrid ^bUniversidad de Castilla-La Mancha

1.1. Conceptos básicos

Los **modelos mixtos** (MM) para variables de respuesta continuas son modelos estadísticos en los que los residuos siguen una distribución Normal pero puede que no sean independientes o no tengan varianza constante. Son necesarios 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 elemento o individuo es medido repetidamente a lo largo del tiempo o bajo condiciones distintas.

Este tipo de estudios se pueden encontrar en diferentes áreas como la medicina, biología, ciencias experimentales y sociales.

1.1.1. Tipo y estructura de los datos

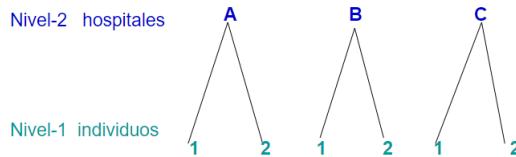
La estructura de los datos con la que se trabaja es el factor determinante para saber si se han de utilizar modelos mixtos, y en su caso, qué tipo de modelo.

1.1.1.1. Datos jerárquicos (o agrupados)

En este tipo de datos, la variable dependiente (de respuesta, de interés) se mide una sola vez en cada unidad de análisis (individuos, objetos, elementos ...), y los individuos¹ están agrupados (o anidados) en unidades mayores. Muchos tipos de datos tienen una estructura jerárquica: alumnos en escuelas, personas en municipios, 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. Por ejemplo, supóngase que se quiere hacer un estudio sobre el tiempo de recuperación en pacientes hospitalizados por COVID-19 en diferentes hospitales. Se tiene la siguiente estructura con dos niveles:

- Muchos individuos en el nivel 1 (pacientes).
- Agrupados en unas pocas unidades en el nivel 2 (hospitales).



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

En cada nivel de la jerarquía se pueden medir variables. Algunas estarán medidas en su nivel *natural*; por ejemplo, en el nivel “hospital” se podría medir el tamaño y en el nivel “pacientes” situación socio-económica. Además, se pueden mover las variables de un nivel a otro mediante agregación o desagregación:

- **Agregación:** la variable correspondiente al nivel más bajo se mueve a un nivel más alto; por ejemplo, se puede asociar a cada hospital la media del nivel socioeconómico de sus pacientes.
- **Desagregación:** mover las variables a un nivel más bajo; por ejemplo, asignarle a cada paciente una variable que indique el tamaño de su hospital de referencia.

1.1.1.2. Medidas repetidas y datos longitudinales

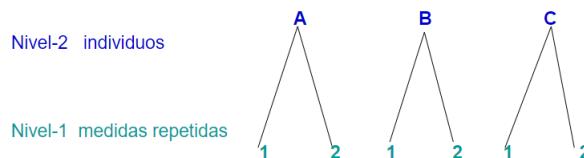
En este tipo de datos, la variable dependiente se mide más de una vez en un mismo individuo ([Singer and Willet, 2003](#)). Por ejemplo, se miden los niveles de glucosa de un enfermo antes y después de haberle inyectado insulina. Este tipo de datos también puede ser considerado como

¹En adelante nos referiremos a las unidades de análisis como individuos.

1.1. Conceptos básicos

13

datos multinivel (o jerárquicos) donde el nivel 2 representa a los individuos y el nivel 1 a las diferentes medidas tomadas. Dado que las medidas se toman a un mismo individuo, es probable que no sean independientes, por lo que utilizar un modelo lineal ordinario no sería apropiado.



Por **datos longitudinales** se entienden 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 MM 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 no son independientes.

1.1.2. ¿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 ambos efectos es distinta.

Los **efectos fijos** son variables en las cuales el investigador ha incluido sólo los niveles (o tratamientos) que son de su interés. Por ejemplo, en un experimento se puede estar interesado en comparar dos grupos, uno al que se le aplica un tratamiento y otro de control. En este caso, el estudio compara los grupos y no interesa generalizar los resultados a otros tratamientos que podrían haber sido incluidos. Otro ejemplo sería el caso en el que se hace una encuesta y se eligen 10 ciudades. Si sólo interesan los resultados para esas 10 ciudades y no se quieren generalizar al resto de ciudades que podrían haber sido seleccionadas, la variable ‘ciudad’ es un efecto fijo. Si se eligen las ciudades de forma aleatoria de una población grande de ciudades, la variable ‘ciudad’ es un **efecto aleatorio**.

Una cantidad se considera aleatoria cuando cambia sobre las unidades de una población. Cuando un efecto en un modelo estadístico es considerado aleatorio, se está asumiendo que se quieren extraer conclusiones sobre la población de la cual se han elegido las unidades observadas, y no se tiene interés en esas unidades en particular. En este contexto se habla de **intercambiabilidad**, en el sentido de que se podría cambiar una unidad de la muestra por otra de la población y sería indiferente. Este es el caso de los factores de agrupamiento o diseño, como las parcelas en un experimento agrícola, o los días cuando un experimento se lleva a cabo en días distintos, o el 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, pues en este último caso 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, normalmente interesa comparar los resultados de la variable dependiente para los distintos niveles de la variable explicativa, es decir, interesa la diferencia entre las medias. En el caso de efectos aleatorios, no interesa específicamente comparar si las medias son distintas, sino 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.

La Fig. 1.1 puede ayudar a determinar si un efecto es fijo o aleatorio:

1. ¿Cuál es el número de niveles?

Pequeño	Fijo
Grande o infinito	Possiblemente aleatorio

2. ¿Son los niveles repetibles?

Sí	Fijo
No	Aleatorio

3. ¿Hay, conceptualmente, un número infinito de niveles?

No	Possiblemente fijo
Sí	Possiblemente aleatorio

4. ¿Se necesitan realizar inferencias para niveles no incluidos en el muestreo?

No	Possiblemente fijo
Sí	Possiblemente aleatorio

Figura 1.1: Cuestiones para determinar si un efecto es fijo o 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 el efecto ‘empresa’. Si esta variabilidad fuera próxima a cero, no sería necesario incluir el efecto aleatorio asociado con la empresa.

¿Por qué hay que utilizar modelos mixtos?

Cuando las observaciones están agrupadas en niveles o siguen una cierta jerarquía, las unidades se ven afectadas por el grupo al que pertenecen. Las jerarquías (o niveles) 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 se hubieran seleccionado aleatoriamente de entre toda la población de alumnos. Los modelos mixtos permiten tener en cuenta que las observaciones no son independientes.

1.2. Formulación del modelo con efectos aleatorios o modelos mixtos

El nombre de *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.

1.2.1. Formulación general

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim N(0, \mathbf{G}), \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{R}), \quad (1.1)$$

donde:

- \mathbf{X} es una matriz $n \times k$ (k es el número de efectos fijos).
- \mathbf{Z} es una matriz $n \times p$ (p es el número de efectos aleatorios).
- $\boldsymbol{\beta}$ es el vector de efectos fijos y \mathbf{u} el de efectos aleatorios.
- \mathbf{G} es la matriz de varianzas-covarianzas de los efectos aleatorios, con dimensión $p \times p$.
- \mathbf{R} es la matriz de varianzas-covarianzas del error.

1.2.1.1. Estimación de $\boldsymbol{\beta}$ y \mathbf{u}

Se hace mediante las llamadas **ecuaciones de Henderson** (Henderson, 1953). Permiten obtener el mejor estimador lineal insesgado de $\boldsymbol{\beta}$ y el mejor predictor lineal insesgado de \mathbf{u} . Se obtienen maximizando la densidad conjunta de \mathbf{y} y \mathbf{u} :

$$f(\mathbf{y}, \mathbf{u}) = f(\mathbf{y}|\mathbf{u})f(\mathbf{u}), \quad \mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \quad \mathbf{u} \sim N(0, \mathbf{G}). \quad (1.2)$$

Derivando con respecto a $\boldsymbol{\beta}$ y \mathbf{u} e igualando a cero se obtienen las ecuaciones de Henderson, cuya solución es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (1.3)$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (1.4)$$

donde $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. Sin embargo, \mathbf{V} depende de los parámetros de la varianza en el modelo, que forman parte de \mathbf{G} y \mathbf{R} y que es necesario estimar, como se muestra a continuación.

1.2.1.2. Estimación de los componentes de la varianza

Los métodos más comunes para la estimación de los parámetros de las matrices de covarianzas son: máxima verosimilitud (ML) y máxima verosimilitud restringida (REML). No existe una solución cerrada para los estimadores, y se estiman de forma numérica o mediante algoritmos iterativos. REML tiene en cuenta los grados de libertad utilizados para estimar los efectos fijos en el modelo. Si n es pequeño, REML dará mejores estimaciones que ML; si n es grande, no habrá prácticamente ninguna diferencia. El método preferido es REML.

1.2.2. Inferencia y selección del modelo

1.2.2.1. Contrastes de hipótesis para los efectos fijos, β

Utilizando la distribución aproximada:

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

- Si se contrastan parámetros individuales, se utiliza el t-test para un solo efecto.
- Si se contrasta un conjunto de parámetros, se utiliza el F-test para más de un efecto.
- También se pueden comparar modelos usando el test de la razón de verosimilitud, LRT por sus siglas en inglés:

$$LRT = -2 [\ln(l_{H_0}) - \ln(l_{H_1})] \approx \chi^2_{df}. \quad (1.5)$$

Nota

En este caso hay que utilizar ML para estimar los parámetros de la varianza.

1.2.2.2. Contrastes de hipótesis para los parámetros de varianza

Al usar el test LRT (Eq. (1.5)) se ha de tener en cuenta que la distribución asintótica del estadístico del test depende de si el valor del parámetro bajo la hipótesis nula (H_0) está o no en la frontera del espacio paramétrico²

- **Caso 1:** El valor de los parámetros de varianza bajo H_0 no está en la frontera del espacio paramétrico (por ejemplo, al contrastar si los parámetros de varianza de dos efectos aleatorios son iguales o no). En ese caso se utiliza el test normalmente.
- **Caso 2:** El valor de los parámetros de varianza bajo H_0 está en la frontera del espacio paramétrico (por ejemplo, si se quiere contrastar si la varianza del efecto aleatorio es cero o no). La distribución asintótica del estadístico del test es una mixtura entre χ^2_p y χ^2_{p-1} , concretamente $0,5\chi^2_p + 0,5\chi^2_{p-1}$, donde p es el número de parámetros de la varianza que se hacen cero bajo la H_0 .

²Es espacio paramétrico es el conjunto de posibles valores del parámetro. Los valores que están en la frontera son los valores que están en el límite inferior (el mínimo) o el superior (máximo) del conjunto de valores posible. Dado que la varianza es positiva, si se contrasta si el valor es cero, estaría tomando un valor en la frontera.

1.2.3. Diagnosis del modelo

En el caso de modelos mixtos, se ha de contrastar la hipótesis de normalidad tanto para los residuos al nivel más bajo como para los efectos aleatorios, así como también las hipótesis de independencia **LOS SUPUESTOS DE INDEPENDENCIA DEBERÍAN PONERSE EN LA FORMULACIÓN GENERAL DEL MODELO.**

En este tipo de modelos se utilizan los residuos condicionales, que son la diferencia entre los valores observados y el valor predicho condicional:

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

Estos residuos tienden a estar correlados y sus varianzas pueden cambiar de un grupo a otro, aunque en el verdadero modelo los residuos están incorrelados y tienen 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 estándar son conocidas), o a los **residuos studentizados** (si son desconocidas y se utilizan estimaciones de las mismas). Con estos residuos se hace un análisis similar al caso de los modelos de regresión lineal.

1.3. Procedimiento con **R** para ajustar modelos mixtos

Hay varios paquetes de **R** para el ajuste de modelos mixtos. Los más usados son **nlme** y **lme4**. El segundo es una versión del primero que incluye modelos más generales y mejora los gráficos. A continuación se describe la función principal del paquete **lme4**.

1.3.1. La función **lmer()**

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

También se puede especificar el método de estimación: “REML” o “ML”.

La parte aleatoria del modelo se incluye entre paréntesis en la ecuación y “|” separa las variables de agrupamiento de las predictoras. Si no hay variables predictoras para la parte aleatoria se pone un 1.

La función **VarCorr()** aplicada a un objeto **lmer** proporciona información sobre la estructura de componentes de varianza.

1.4. Caso práctico

En esta sección se comienza viendo cómo construir diferentes modelos con efectos aleatorios según a qué nivel estén medidas las variables explicativas y se termina dando una guía de cons-

trucción de estos modelos en la práctica. Los datos con los que se va a trabajar se encuentran en el dataframe `Hsb82` del paquete `mlmRev` y provienen de un estudio titulado *High School and Beyond*. Los datos corresponden a 7.185 estudiantes repartidos en 160 escuelas, el número de alumnos por escuela varía entre 14 y 67. La variable de interés, `mAch`, es el nivel estandarizado alcanzado en matemáticas. Una cuestión inicial que se puede plantear es si el nivel socioeconómico (`cse`) del alumno predice las diferencias en el nivel de matemáticas. Para ello se ajusta el modelo:

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

que 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).

```
library("mlmRev")
Hsb82$school = factor(Hsb82$school, ordered=F)
multi0 <- lm(mAch ~ cses, data = Hsb82)
summary(multi0)

#>
#> Call:
#> lm(formula = mAch ~ cses, data = Hsb82)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -17.8660  -5.1165   0.2966   5.3880  14.8705
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 12.74785   0.07933 160.69 <2e-16 ***
#> cses         2.19117   0.12010   18.24 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.725 on 7183 degrees of freedom
#> Multiple R-squared:  0.04429,    Adjusted R-squared:  0.04415
#> F-statistic: 332.8 on 1 and 7183 DF,  p-value: < 2.2e-16
```

La ordenada en el origen es 12.75 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 se puede ver que el coeficiente es significativo.

Supóngase que ocurre la situación mostrada en la Fig. 1.2:

Los alumnos de la escuela A (rombos negros) sacan, en promedio, mejores notas que las que le asigna el modelo ajustado; con la escuela B (círculos azules) ocurre lo contrario. El gráfico indica que la ordenada en el origen (el intercepto) 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 hacia abajo. Eso se puede conseguir permitiendo que cada escuela tenga su propia ordenada en el origen:

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

1.4. Caso práctico

19

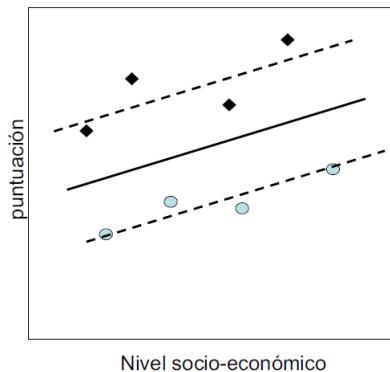


Figura 1.2: Ilustración de posibles escenarios para dos escuelas

Este modelo es similar al anterior añadiendo el subíndice i para identificar el centro al que pertenece cada alumno. En realidad, se utiliza una variable categórica con tantas categorías como escuelas.

```
multi1 <- lm(mAch ~ cses + school, data = Hsb82)
```

Se están considerando las escuelas como un efecto fijo y no aleatorio, es decir, implícitamente se está suponiendo que solo interesan estas escuelas en particular.

La situación 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 pueda dar lugar a un aumento distinto en la nota del test en cada centro. En la Fig. 1.3 se ve como la pendiente de la recta para la escuela C es distinta a la dos anteriores.

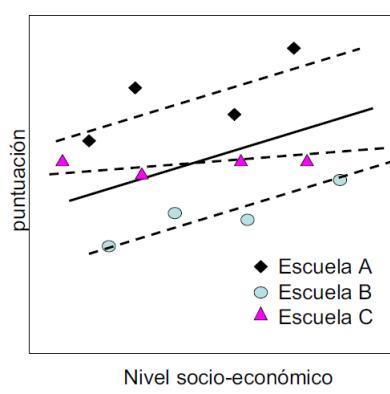


Figura 1.3: Ilustración de posibles escenarios para tres escuelas

El modelo que permite tener en cuenta esta situación es:

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

donde aparece ahora el sub-índice i también en la pendiente, lo que indica que cada centro tiene una pendiente diferente.

El código

```
multi2 <- lm(mAch ~ cses*school, data = Hsb82)
```

generaría 159 coeficientes más (uno por cada escuela), que son los que se incluirían con la interacción. Pero no interesan estas escuelas en concreto, sino la población de la que estas escuelas son una muestra.

Con un modelo con efectos aleatorios, sin embargo, se pueden contestar preguntas como: ¿Cuáles son las causas de esta variabilidad? ¿Qué variables pueden explicarla?

1.4.1. Modelo con ordenada en el origen aleatoria

Es el modelo mixto más sencillo. Se considera que los datos tienen una estructura con dos niveles: los alumnos están en el nivel 1 y están agrupados en escuelas, nivel 2. Se empieza suponiendo que no se dispone de ninguna variable explicativa, y que por lo tanto el único interés es la diferencia entre las notas medias del test de matemáticas en los distintos centros.

Los dos niveles del modelo son:

$$\text{Nivel 1: } y_{ij} = \beta_{0i} + \epsilon_{ij}$$

- El subíndice j corresponde a alumnos y el i a escuelas, si se considera a las escuelas como un efecto aleatorio,
- β_{0i} (la media de cada escuela) vendría dada por:

$$\text{Nivel 2: } \beta_{0i} = \beta_0 + u_i,$$

- β_0 es la media de todos los alumnos,
- u_i es la desviación de la media de la escuela i respecto de la media de todas las escuelas.

Poniendo las dos ecuaciones juntas:

$$y_{ij} = \beta_0 + u_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_m \quad (1.6)$$

- La media de y para el grupo i es $\beta_0 + u_i$,
- 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,
- $u_i \sim N(0, \sigma_u^2)$, $\epsilon_{ij} \sim N(0, \sigma_e^2)$, y ambos son independientes, es decir, las observaciones que provienen de distintas escuelas son independientes.

En el ejemplo de las escuelas:

1.4. Caso práctico

21

```
library("lme4")
Modelo0 <- lmer(mAch ~ 1+(1 | school), data = Hsb82)
Modelo0
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ 1 + (1 | school)
#> Data: Hsb82
#> REML criterion at convergence: 47116.79
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.935
#> Residual           6.257
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)
#>          12.64
```

- La media total estimada es 12.64,
- La media estimada para la escuela i es: $12.64 + \hat{u}_i$, donde \hat{u}_i es el efecto aleatorio estimado de dicha escuela.

Para obtener los valores predichos de los efectos aleatorios se utiliza la función `ranef()`.

La Fig. 1.4 permite ver los efectos aleatorios junto con sus intervalos de confianza (las escuelas han sido ordenadas atendiendo a su media para apreciar mejor la variabilidad entre las mismas).

Debajo de la chunck se dice que \hat{u}_i es el efecto aleatorio ESTIMADO, y justo debajo se dice ‘para obtener los valores PREDICHIOS...’ habría que poner el mismo término, ¿no?

Poner la caption de la figura

Se dibujan los efectos aleatorios para ver si siguen una distribución Normal; para eso se ajusta el modelo con la función `lmer()`:

Cuando se dice se dibujan los efectos aleatorios para ver si siguen una distribución normal: los efectos se dibujaron junto con sus intervalos para poderlos ver en un gráfico; ha estaban dibujados, no?

```
library("lattice")
qqmath(ranef(Modelo0, condVar = TRUE))$school
```

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

```
confint(Modelo0)
#>              2.5%    97.5%
#> .sig01      2.594729  3.315880
#> .sigma       6.154803  6.361786
#> (Intercept) 12.156289 13.117121
```

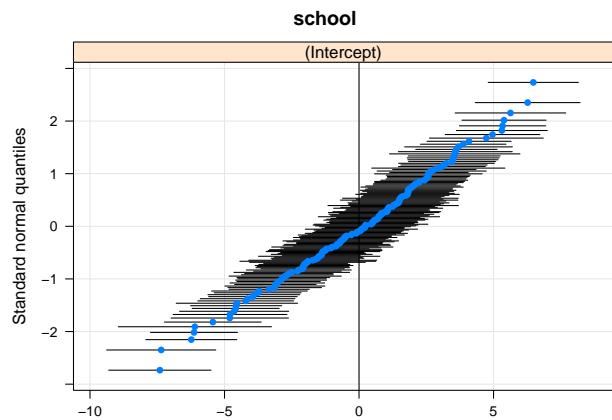


Figura 1.4: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

pudiéndose apreciar que el intervalo para sig01 no contiene al cero. Sin embargo, la forma más correcta de hacerlo sería utilizando el LRT (véase Eq. (1.5)) con:

$$\begin{aligned} H_0 : \quad \sigma_u^2 = 0 &\Rightarrow y_{ij} = \beta_0 + \epsilon_{ij} \\ H_1 : \quad \sigma_u^2 \neq 0 &\Rightarrow y_{ij} = \beta_0 + u_i + \epsilon_{ij}. \end{aligned}$$

El resultado del test, en este caso, se compara con el valor de una mixtura de distribuciones Chi-cuadrado $0,5\chi_0^2 + 0,5\chi_1^2$.

Decir que significa el 0 y el 1 de las chi cuadrado (bajo H_0 y H_1) La gente puede pensar que la primera tiene cero grados de libertad y la segunda 1 Otra opción es decirlo en los términos en los que os habéis referido a ello al final de la página 6: ‘La distribución asintótica del estadístico del test es una mixtura entre χ_p^2 y χ_{p-1}^2 , concretamente $0,5\chi_p^2 + 0,5\chi_{p-1}^2$, donde p es el número de parámetros de la varianza que se hacen cero bajo la H_0 .’

```
Modelo_NULL <- lm(mAch ~ 1, data = Hsb82)
test = -2*logLik(Modelo_NULL, REML = T) + 2*logLik(Modelo0, REML = T)
mean(pchisq(test, df = c(0, 1), lower.tail = F))
#> [1] 9.320673e-217
```

Conclusión: el efecto aleatorio es necesario en el modelo.

El siguiente paso sería introducir las variables explicativas (en este caso solo hay una), ya estén al nivel 1 o al 2.

1.4.1.1. Variables explicativas en el nivel 1 (individuos)

Como la variable explicativa está medida al nivel 1, se introduce en la ecuación del nivel 1:

1.4. Caso práctico

23

- Nivel 1: $y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \epsilon_{ij}$
- Nivel 2: $\beta_{0i} = \beta_0 + u_i$

En el nivel 2 he cambiado μ_i por β_{0i} . Confirmadme que es correcto

Si 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:

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

En este modelo, la relación global entre Y y 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 una determinada escuela i viene dada por $\beta_0 + u_i$. Será mayor o menor que la ordenada en el origen global β_0 en una cantidad u_i . Aunque la ordenada en el origen varía de grupo a grupo, la pendiente es la misma para todos los grupos. Todas las líneas rectas ajustadas para cada grupo son paralelas.

En el ejemplo de las escuelas, se introduce como variable explicativa `cse`s (nivel socioeconómico centrado en su media):

```
Modelo1 <- lmer(mAch ~ cses+(1 | school), data = Hsb82)
Modelo1
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses + (1 | school)
#>   Data: Hsb82
#> REML criterion at convergence: 46724
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.945
#> Residual           6.084
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)      cses
#>          12.636     2.191
```

Ahora se tienen dos efectos fijos:

$$\begin{aligned}\hat{\beta}_0 &= 12,64 \\ \hat{\beta}_1 &= 2,19\end{aligned}$$

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

$$E[y|cses] = 12,64 + 2,19 cses$$

Para contrastar si la variable `cse`s es significativa se utiliza el LRT (Eq. (1.5)). Primero se tienen que ajustar de nuevo los modelos que se quieren comparar usando máxima verosimilitud (en vez de REML). Si se utiliza la función `lmer()` para ajustar el modelo no es necesario reajustar con

ML pues la función `anova` lo hará automáticamente, mientras que si se usa la función `lme()` sí será necesario hacerlo.

```
anova(Modelo0, Modelo1)
#> Data: Hsb82
#> Models:
#> Modelo0: mAch ~ 1 + (1 | school)
#> Modelo1: mAch ~ cses + (1 | school)
#>          npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo0     3 47122 47142 -23558      47116
#> Modelo1     4 46728 46756 -23360      46720 395.4  1 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por lo tanto, el nivel socioeconómico afecta a los resultados escolares. Comparado con el modelo sin la variable explicativa (Modelo0), la inclusión del nivel socio-económico (Modelo1) ha reducido la variabilidad a nivel del alumno en un 2.8% $(6.084 - 6.257)/6.257 = -0.028$.

1.4.1.2. Variables explicativas en el nivel 2 (grupos)

Si las variables explicativas se miden al nivel 2, entonces:

$$\begin{aligned} \text{Nivel 1: } y_{ij} &= \beta_{0i} + \epsilon_{ij} \\ \text{Nivel 2: } \beta_{0i} &= \beta_0 + \beta_2 s_i + u_i \end{aligned}$$

$$y_{ij} = \underbrace{\beta_0 + \beta_2 s_i}_{\text{efectos fijos}} + \underbrace{u_i + \epsilon_{ij}}_{\text{efectos aleatorios}} .$$

En ambos niveles he cambiado μ_i por β_{0i} . Confirmadme que es correcto. Lo he hecho también en los demás modelos que vienen

En nuestro ejemplo, la variable utilizada es `sector` (público o privado):

$$mAch = \beta_0 + \beta_2 \text{sector} + u_i + \epsilon_{ij}$$

Se ajusta el modelo usando la función `lmer()`:

```
Modelo2 <- lmer(mAch ~ sector + (1 | school), data = Hsb82)
Modelo2
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ sector + (1 | school)
#>   Data: Hsb82
#> REML criterion at convergence: 47080.13
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.584
```

1.4. Caso práctico

25

```
#> Residual           6.257
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#>   (Intercept) sectorCatholic
#>             11.393          2.805
```

$$E[y|sector] = 11,39 + 2,8 \text{ sector},$$

o equivalentemente

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

La nota de un alumno en una escuela privada se espera que sea 2.8 unidades mayor que la de un alumno en una escuela pública (se puede generalizar *¿A qué? podíais poner una nota a pie de página*, pues se asume que las escuelas son un efecto aleatorio). La varianza del efecto aleatorio de nivel 2, σ_u^2 , ha descendido: $(2,935^2 - 2,584^2)/2,935^2 = 0,22$, es decir, al introducir la variable sector se ha reducido en un 22% la variabilidad *no explicada entre los centros* *¿no será la variabilidad a nivel de escuela?* Lo he escrito así (a nivel de) para utilizar los mismos términos que cuando lo habéis dicho en el caso de meter una variable explicativa en el nivel 1, pero deberíais decirlo en términos "de clase" porque habrá mucha gente que verá esto aquí por primera vez.

2,584 y 2,935 son desviaciones típicas, no varianzas

En la salida pone sectorCatholic, cuando se supone que los niveles eran público y privado

Por cierto, veo que arriba a la varianza del efecto aleatorio de nivel 2 le llamáis σ_u^2 . Pero en el final de la página 10 decís $u_i \sim N(0, \sigma_i^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$. Habría que unificar la notación. Yo creo que lo del final de la página 10 habría que ponerlo bien

Para contrastar si la variable sector es significativa se usa de nuevo el test LRT:

```
anova(Modelo0, Modelo2)
#> Data: Hsb82
#> Models:
#>   Model0: mAch ~ 1 + (1 | school)
#>   Model2: mAch ~ sector + (1 | school)
#>             npar  AIC  BIC logLik deviance Chisq Df Pr(>Chisq)
#>   Model0     3 47122 47142 -23558      47116
#>   Model2     4 47087 47115 -23540      47079 36.705  1  1.374e-09 ***
#>   ---
#>   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por lo tanto, el hecho de que la escuela sea pública o privada influye en el resultado académico de los alumnos.

1.4.2. Modelo con pendiente aleatoria

En este tipo de modelos se supone que la relación entre la variable respuesta y las variables explicativas es distinta para las distintas unidades de nivel 2, es decir, la relación puede cambiar de un centro educativo a otro. Por ejemplo, el efecto del nivel socioeconómico en las notas puede ser distinto en distintos centros, de modo que se puede 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.

$$\begin{aligned} \text{Nivel 1: } & y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij} \\ \text{Nivel 2: } & \beta_{0i} = \beta_0 + u_i \\ & \beta_{1i} = \beta_1 + v_i \end{aligned}$$

Poniendo las dos ecuaciones juntas:

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

donde σ_{uv} es la covarianza entre las ordenadas en el origen y las pendientes de los grupos (β_{0i} y β_{1i} , respectivamente). Un valor positivo de la covarianza implica que los grupos con un valor del efecto de grupo u_i elevado tienden a tener valores elevados de v_i , o equivalentemente, centros educativos con ordenada en el origen alta, tienen pendiente alta.

El modelo en R sería:

```
Modelo3 <- lmer( mAch ~ cses + (cses | school), data = Hsb82)
Modelo3
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses + (cses | school)
#>   Data: Hsb82
#> REML criterion at convergence: 46714.23
#> Random effects:
#> Groups      Name        Std.Dev. Corr
#> school     (Intercept) 2.9464
#>           cses         0.8331  0.02
#> Residual            6.0581
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)      cses
#>       12.636      2.193
```

El efecto del nivel socioeconómico en la escuela i se estima como $2,19 + \hat{u}_i$, y la varianza de las pendientes entre No sería mejor 'para las'? escuelas es $0,833^2 = 0,694$. Para la escuela promedio se predice un aumento de 2,19 en la puntuación cuando el nivel socioeconómico aumenta en una unidad.

Ahora se tienen las siguientes estimaciones de los parámetros de varianza:

$$\hat{\sigma}_u^2 = 8,68 \quad \hat{\sigma}_v^2 = 0,694 \quad \hat{\sigma}_{uv} = \rho\sigma_u\sigma_v = 0,051 \quad \hat{\sigma}^2 = 36,7$$

1.4. Caso práctico

27

La varianza de la ordenada en el origen estimada, 8,68, se interpreta como la variabilidad (de la nota) entre las escuelas para un nivel socioeconómico medio (valor nulo de la variable por estar centrada).

El parámetro de covarianza estimado es $\sigma_{uv} = 0,051$, por lo que se puede plantear si es necesario o no.

Para comprobarlo, el contraste de hipótesis sería en este caso:

$$H_0 : \sigma_{uv} = 0 \quad \text{y} \quad H_1 : \sigma_{uv} \neq 0$$

```
Modelo3.1 <- lmer(ses ~ cses + (cses || school), data = Hsb82)
```

Cuando se quiere que haya un efecto aleatorio para la ordenada en el origen y para la pendiente pero que estén incorrelados, en la función solo hay que poner doble barra en vez de simple.

En este caso no es necesario utilizar la mixtura de distribuciones Chi-cuadrado para contrastar $H_0 : \sigma_{uv} = 0$, pues σ_{uv} puede tomar cualquier valor.

```
anova(Modelo3.1, Modelo3)
#> Data: Hsb82
#> Models:
#> Model03.1: ses ~ cses + ((1 | school) + (0 + cses | school))
#> Model03: mAth ~ cses + (cses | school)
#>          npar      AIC      BIC logLik deviance Chisq Df Pr(>Chisq)
#> Model03.1     5 -226164 -226129 113087   -226174
#> Model03       6   46723   46764 -23355    46711      0  1             1
```

Por lo tanto, se puede suponer que la covarianza es 0.

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 se necesita la aproximación:

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

Por lo tanto, la pendiente es diferente en las distintas escuelas.

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

```
AIC(logLik(Modelo3))
#> [1] 46726.23
AIC(logLik(Modelo3.1))
#> [1] -226113.6
AIC(logLik(Modelo1))
#> [1] 46732
```

A veces la covariable medida a nivel 2 (a nivel de grupo, en este caso escuelas) puede explicar tanto la variabilidad de la ordenada en el origen como de la pendiente:

$$\text{Nivel 1: } y_{ij} = \beta_{i0} + \beta_{1i}x_{ij} + \epsilon_{ij}$$

$$\text{Nivel 2: } \beta_{i0} = \beta_0 + \beta_{2i}s_i + u_i$$

$$\beta_{1i} = \beta_1 + \beta_{3i}s_i + v_i$$

$$y_{ij} = \underbrace{\beta_0 + \beta_1x_{ij} + \beta_{2i}s_i + \beta_{3i}x_{ij}s_i}_{\text{efectos fijos}} + \underbrace{u_i + v_ix_{ij} + \epsilon_{ij}}_{\text{efectos aleatorios}}.$$

CUIDADO¡¡¡ EL MODELO FINAL OS LO HE CORREGIDO; EN MI OPINIÓN ESTABAN VARIAS COSAS MAL; CONFIRMADME QUE ESTE ES EL CORRECTO

Al introducir la variable medida al nivel 2, la parte fija se modifica (con respecto al Modelo 3), pero no la parte aleatoria:

- β_2 No hay β_2 en el modelo; hay β_{2i} indica si los centros privados son diferentes de los públicos en cuanto a su nota media.
- β_3 No hay β_e en el modelo; hay β_{3i} indica si los centros privados difieren de los públicos en cuanto a la relación entre el nivel socio-económico y la nota en matemáticas.

```
Modelo4 <- lmer(mAch ~ cses*sector + (cses || school),
                 data = Hsb82)
Modelo4
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses * sector + ((1 | school) + (0 + cses / school))
#>   Data: Hsb82
#> REML criterion at convergence: 46648.85
#> Random effects:
#> Groups      Name        Std.Dev.
#> school      (Intercept) 2.5971
#> school.1    cses        0.5182
#> Residual    6.0580
#> Number of obs: 7185, groups:  school, 160
#> Fixed Effects:
#>             (Intercept)          cses       sectorCatholic
#>             11.393            2.784            2.805
#> cses:sectorCatholic
#>             -1.346
```

Los centros privados tienen una nota media más alta que los públicos (2.81 puntos más), y una pendiente más suave que la de dichos centros públicos (-1.35). Esto último indica que en un colegio privado la mejora de la nota con respecto al nivel socio-económico es más lenta que un colegio público.

1.4.3. ¿Cómo construir el modelo en la práctica?

1. Se ajusta el modelo con todos los efectos fijos y aleatorios posibles.

1.4. Caso práctico

29

```
Modelo5 <- lmer(mAch ~ cses * sector+(cses | school), data = Hsb82)
```

2. Se contrasta qué efectos aleatorios son significativos, sin mover los efectos fijos.

Primero se contrasta si la covarianza entre efectos fijos y aleatorios es cero o no:

```
#Se ajusta el modelo con covarianza = 0
Modelo5.1 <- lmer(ses ~ cses * sector+(cses || school),
                   data = Hsb82)
anova(Modelo5.1, Modelo5)
#> Data: Hsb82
#> Models:
#> Modelo5.1: ses ~ cses * sector + ((1 | school) + (0 + cses | school))
#> Modelo5: mAch ~ cses * sector + (cses | school)
#>          npar    AIC    BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo5.1     7 -220069 -220021 110042   -220083
#> Modelo5      8  46650  46705 -23317    46634      0  1
```

Como lo es, no se tendría que contrastar nada más. Los efectos aleatorios son los que se han incluido en el Modelo5. Si no hubiera sido significativa, se continuaría contrastando si la pendiente aleatoria es significativa y si la ordenada en el origen lo es.

3. Una vez elegidos los efectos aleatorios que se mantienen en el modelo, se eligen los efectos fijos:

```
Modelo6 = update(Modelo5, . ~ .-cses:sector)
anova(Modelo6, Modelo5)
#> Data: Hsb82
#> Models:
#> Modelo6: mAch ~ cses + sector + (cses | school)
#> Modelo5: mAch ~ cses * sector + (cses | school)
#>          npar    AIC    BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo6     7 46678 46726 -23332    46664
#> Modelo5     8 46650 46705 -23317    46634 29.983  1  4.358e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como la interacción es significativa ya no se tendría que hacer ningún test más y este sería el modelo final.

¿Se pueden poner unas palabritas intuitivas sobre el hecho de que la significatividad de la interacción implica que este es el modelo final?

Resumen

En este capítulo se introducen los modelos mixtos o modelos con efectos aleatorios. En particular:

- Se dan las claves para distinguir entre efectos fijos y aleatorios.
- Se presenta la formulación del modelo y indica cómo llevar a cabo la estimación del mismo.
- Se explican las etapas del proceso a seguir para el ajuste de este tipo de modelos.
- Se muestra el uso de **R** para ajustar estos modelos.
- Se ilustra el análisis de modelos multinivel como caso particular de un modelo con efectos aleatorios.

Bibliografía

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.

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

Snijers, T. (2003). Fixed and random effects. 2:664–665.

Índice alfabético

- agregación, 12
- datos
 - jerárquicos, 12
 - longitudinales, 13
- desagregación, 12
- ecuaciones de Henderson, 15
- efectos
 - aleatorios, 13
 - fijos, 13
- intercambiabilidad, 13
- modelos mixtos lineales, 15
- residuos
 - estandarizados, 17
 - studentizados, 17