Trabajo Fin de Máster

Análisis longitudinal en estudios farmacogenéticos

Beatriz Piñeiro Lamas

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Análise lonxitudinal en estudos farmacoxenéticos

Título en español: Análisis longitudinal en estudios farmacogenéticos

English title: Longitudinal analysis in pharmacogenetic studies

Modalidad: Modalidad B

Autor/a: Beatriz Piñeiro Lamas, Universidade de Santiago de Compostela

Director/a: Rosa M. Crujeiras Casais, Universidade de Santiago de Compostela
Alberto Rodríguez Casal, Universidade de Santiago de Compostela

Tutor/a: Raquel Cruz Guerrero, CiMUS

Breve resumen del trabajo:
La farmacogenética es una disciplina científica que estudia la base genética de la variabilidad en la respuesta a fármacos y el riesgo de efectos adversos. En este contexto, debido al seguimiento a largo plazo de los pacientes, existe mucha información de registros de variables. Dichos registros pueden ser considerados como mediciones longituinales, y podrían ser analizados en el contexto de un modelo lineal mixto. El objetivo fundamental de este trabajo es el análisis de la aplicabilidad, ventajas y limitaciones de los modelos mixtos a un estudio de farmacogenética y su comparación con aproximaciones estadísticas más tradicionales en este contexto.

Otras observaciones:
Este trabajo se ha realizado a propuesta de la estudiante Beatriz Piñeiro Lamas.

Doña Rosa M. Crujeiras Casais, Profesora Titular del Departamento de Estadística, Análisis Matemático y Optimización de la Universidade de Santiago de Compostela, don Alberto Rodríguez Casal, Profesor Titular del Departamento de Estadística, Análisis Matemático y Optimización de la Universidade de Santiago de Compostela, y doña Raquel Cruz Guerrero, Investigadora de CiMUS, informan que el Trabajo Fin de Máster titulado

Análisis longitudinal en estudios farmacogenéticos

fue realizado bajo su dirección por doña Beatriz Piñeiro Lamas para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 09 de Julio de 2020.

La directora:

CRUJEIRAS CASAIS ROSA MARIA - 52931780V

Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2020.07.09 23:22:35 +02'00'

Doña Rosa M. Crujeiras Casais

El director:

RODRIGUEZ CASAL ALBERTO - 34896829X

Firmado digitalmente por RODRIGUEZ CASAL ALBERTO - 34896829X
Fecha: 2020.07.09 19:03:20 +02'00'

Don Alberto Rodríguez Casal

La tutora:

CRUZ GUERRERO, RAQUEL (AUTENTICACIÓN)

Firmado digitalmente por CRUZ GUERRERO, RAQUEL (AUTENTICACIÓN)
Nombre de reconocimiento (DN): c=ES, serialNumber=12383132R, sn=CRUZ, givenName=RAQUEL, cn=CRUZ GUERRERO, RAQUEL (AUTENTICACIÓN)
Fecha: 2020.07.10 11:05:35 +02'00'

Doña Raquel Cruz Guerrero

La autora:

PIÑEIRO LAMAS BEATRIZ - 79345883T

Firmado digitalmente por PIÑEIRO LAMAS BEATRIZ - 79345883T
Fecha: 2020.07.09 18:44:07 +02'00'

Doña Beatriz Piñeiro Lamas

Agradecimientos

Antes de comenzar me gustaría dar las gracias a mis directores, Alberto Rodríguez y Rosa Crujeiras, por todo lo que me enseñaron tanto en la elaboración de este trabajo como en los estudios de grado y máster.

A Alberto, por guiarme y aconsejarme tan bien desde hace ya varios años, y por ayudarme a introducirme en el mundo de la investigación y a luchar por lo que me gusta.

A Rosa, por su dedicación y por la pasión y motivación con la que hace su trabajo, y por haberse convertido para mí en un ejemplo a seguir.

Gracias también al CiMUS, ya que en sus instalaciones he realizado la mayor parte de este trabajo. A mi tutora Raquel, por todo lo que pude aprender de ella y por darme la oportunidad de integrarme en su grupo.

Finalmente, a mi familia, pareja y amigos. En especial a Borja y a Paula, por todos los buenos momentos que compartimos durante nuestra etapa universitaria y por seguir juntos a pesar de haber tomado distintos caminos.

Índice general

Resumen

Resumen en español

El Trastorno por Déficit de Atención e Hiperactividad (TDAH) es uno de los trastornos psiquiátricos infantiles más prevalentes. El fármaco administrado con mayor frecuencia para su tratamiento es el metilfenidato. A pesar de que dicho fármaco es efectivo en la mayoría de los casos, existe una proporción de pacientes que no responden bien o que sufren efectos adversos al recibirlo. Debido a esto, el tratamiento del TDAH se determina a menudo mediante el *método de prueba y error*, probando diferentes dosis y medicamentos hasta conseguir el efecto terapéutico deseado con el menor número de efectos adversos. Esto retrasa el control de los síntomas de la enfermedad, por lo que un mejor conocimiento del trastorno podría permitir la toma de decisiones con más fundamento para ofrecer a cada individuo una medicina personalizada, administrándole el tratamiento y la dosis más adecuada. En este contexto son de gran utilidad los estudios de asociación de genoma completo, que rastrean el genoma con la finalidad de encontrar variaciones genéticas que nos permitan adelantarnos al fracaso terapéutico o a la aparición de efectos adversos. Además, es conveniente estudiar la respuesta al tratamiento a largo plazo para obtener así resultados más completos.

English abstract

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common childhood psychiatric disorders. Methylphenidate is the most frequently administered drug for its treatment. Despite the effectiveness of this drug in most cases, there is a proportion of patients who do not respond well or experience adverse effects. That is the reason why ADHD treatment is often determined by the *trial and error method*, testing different doses and treatments until the desired therapeutic effect is obtained. This method delays the control of the disorder's symptoms, and for this reason a better knowledge of ADHD could allow to make more informed decisions in order to offer a personalized medicine, giving to each patient the most appropriate treatment and dose. In this context, genome wide association studies try to find genetic variations that could allow us to anticipate therapeutic failure or the appearance of adverse effects. In addition, it is convenient to study the long-term effects of medication to obtain more complete results.

Capítulo 1

Introducción

La farmacogenética es una disciplina científica que estudia el efecto de la variabilidad genética de un individuo en su respuesta a fármacos y su riesgo de sufrir efectos adversos. Esta disciplina es de vital importancia en el caso de enfermedades graves y/o crónicas, como pueden ser el cáncer y el SIDA, debido a la agresividad y/o duración de los tratamientos. El objetivo principal de la farmacogenética es la predicción del riesgo de toxicidad o fracaso terapéutico al administrar un determinado medicamento a una cierta persona. Cada persona es genéticamente diferente de las demás y, por tanto, los medicamentos no actúan de la misma forma para todos. La toxicidad de algunos medicamentos es actualmente un obstáculo importante para el éxito del tratamiento en una proporción significativa de individuos; identificar a aquellos que tienen un mayor riesgo de desarrollar toxicidad podría tener una importancia relevante en la práctica clínica. Lo ideal sería contar con una medicina personalizada con el fin de mejorar la calidad de vida de los pacientes, proporcionando el tratamiento y la dosis más adecuados a cada uno de ellos.

Con frecuencia, los estudios en farmacogenética se basan en el análisis de la variabilidad genética presente en un grupo de pacientes -de una misma patología y bajo el mismo tratamiento- en relación con una serie de variables resumen que sirven de indicadoras de la respuesta al tratamiento o bien de la existencia de toxicidades específicas. Habitualmente estas variables se encuentran dicotomizadas: es común tener una variable que exprese la presencia o ausencia de un determinado efecto adverso, codificada de forma binaria. En estos estudios, por lo tanto, se dispone de información de individuos afectados y no afectados por tal efecto adverso.

El fundamento genético de la variabilidad en la respuesta a los fármacos hay que buscarlo en los polimorfismos genéticos. Un polimorfismo genético es una variación en la secuencia del ADN entre individuos de la misma especie que se encuentra con una frecuencia superior al 5 % (por debajo de este porcentaje se denomina variación rara). Dicha variación puede ser de varios tipos, entre los cuales se encuentran los SNPs. Los SNPs (*single nucleotide polymorphisms*) son los polimorfismos más frecuentes y consisten en la sustitución de una única base nitrogenada (adenina (A), timina (T), citosina (C) o guanina (G)) en un punto concreto del genoma.

Pero, ¿cuántos SNPs influyen sobre la variable respuesta de interés (funcionamiento del medicamento, presencia o ausencia de un efecto adverso)? ¿Cómo se distribuyen dichos SNPs por el genoma? ¿Interactúan entre sí?

En este trabajo nos centraremos en la influencia de los SNPs en la presencia o ausencia de efectos adversos en el tratamiento del Trastorno por Déficit de Atención e Hiperactividad. A continuación presentaremos las características principales del trastorno, con la intención de describir después nuestros datos. Posteriormente introduciremos los estudios de asociación de genoma completo, que son estudios

de rastreo genómico que nos permitirán estudiar el posible efecto de una gran cantidad de SNPs sobre una cierta variable respuesta de interés, como puede ser la presencia de un determinado efecto adverso.

1.1. Trastorno por Déficit de Atención e Hiperactividad

TDAH son las siglas de Trastorno por Déficit de Atención e Hiperactividad. Se trata de un trastorno de carácter neurobiológico originado en la infancia que implica un patrón de déficit de atención, hiperactividad y/o impulsividad (Lange et al., 2010). En muchas ocasiones no se presenta solo, sino que aparece junto a otros trastornos psiquiátricos como son el trastorno negativista desafiante, el trastorno de conducta, el trastorno de ansiedad o el trastorno del ánimo (Larson et al., 2011). Cuando el TDAH se asocia a otros trastornos normalmente se complica el diagnóstico, empeora la evolución y la respuesta al tratamiento es peor.

El TDAH suele aparecer en la infancia, generalmente a partir de los 7 años. Es frecuente que se reconozca en los niños cuando comienza la educación primaria, coincidiendo con dificultades en el rendimiento escolar. Para su diagnóstico es fundamental evaluar que el déficit de atención, hiperactividad e impulsividad se presenten desde una edad temprana (antes de los 12 años) y con una intensidad y frecuencia superior a la normal para la edad y la etapa de desarrollo del niño. Además, se debe evaluar si tales síntomas conllevan problemas en el rendimiento escolar, así como en sus relaciones personales con los amigos, los profesores y la propia familia (es decir, se debe evaluar si les afecta en el ámbito escolar, familiar y social).

No todos los niños con el trastorno manifiestan los mismos síntomas ni con la misma intensidad. En función de los síntomas de cada paciente se han establecido tres subtipos de TDAH:

- Presentación predominante de falta de atención.
- Presentación predominante de hiperactividad/impulsividad.
- Presentación combinada de déficit de atención e hiperactividad/impulsividad.

El subtipo más frecuente es el combinado, que aparece en un 60 % de los casos, seguido del subtipo de falta de atención en un 30 % y del subtipo hiperactivo/impulsivo en un 10 % (Faraone et al., 1998). Existen diferentes patrones de comportamiento en cuanto al sexo: las niñas presentan más frecuentemente comportamientos de falta de atención, mientras que los niños muestran más comportamientos hiperactivos/impulsivos (Biederman et al., 2002).

El TDAH es uno de los trastornos psiquiátricos infantiles más frecuentes, situándose por encima de otros como la esquizofrenia o el trastorno bipolar. El estudio de meta-análisis descrito en Catalá-López et al. (2012) indica que la prevalencia del TDAH en niños y adolescentes en España es de un 6.8 %. Pese a la alta prevalencia del TDAH, nos encontramos ante una realidad social de desconocimiento sobre el trastorno. Además, la cantidad de niños que reciben tratamiento para el TDAH ha aumentado. No está claro si hay más niños que realmente tienen TDAH o si simplemente hay más niños que reciben un diagnóstico de TDAH (aunque no sea un diagnóstico correcto).

El TDAH tiene un componente genético. En cuanto a su heredabilidad, diversos estudios han demostrado que los familiares de personas con TDAH tienen un riesgo significativamente mayor de padecer el trastorno que las personas sin antecedentes familiares; de hecho, se estima una heredabilidad del 76 % (Faraone et al., 2005). Además de los factores genéticos, también hay factores ambientales que afectan al TDAH, como por ejemplo los traumatismos craneoencefálicos en la infancia, la prematuridad, la encefalopatía hipóxico-isquémica, el bajo peso al nacimiento, el consumo de tóxicos como el alcohol o

el tabaco durante el embarazo, el maltrato, el trauma emocional y los abusos sexuales.

1.2. Descripción de los datos

En el presente trabajo, a modo de ilustración, analizaremos los datos que disponemos de 199 pacientes procedentes de las consultas del servicio de neuropediatría infantil y del servicio de psiquiatría infantil, ambos pertenecientes al Hospital Universitario Fundación Jiménez Díaz de Madrid. Los criterios de inclusión en el estudio fueron ser caucásico, tener entre 6 y 18 años y haber sido diagnosticado de TDAH. Por otra parte, los criterios de exclusión fueron no cumplir los criterios de inclusión, presentar enfermedades psiquiátricas mayores (como trastorno bipolar o esquizofrenia) o presentar afectación neurológica o trastorno del desarrollo intelectual (coeficiente intelectual inferior a 70) (Gómez-Sánchez, 2017).

En la Tabla 1.1 se muestra un resumen de los datos. De los 199 pacientes, 158 son niños y 41 son niñas, lo cual se traduce en un 79.4 % y en un 20.6 %, respectivamente. Por lo tanto, se observa una mayor tasa de diagnóstico en niños que en niñas, con un ratio de aproximadamente 4:1. Este resultado concuerda con el ratio global presentado en Novik et al. (2006), que varía desde 3:1 hasta 16:1. En cuanto a la edad, la mínima son 6 años y la máxima 18, siendo la media 11 y la mediana 10.57 años. Con respecto al subtipo de TDAH, 75 de los pacientes presentan el subtipo de falta de atención, 9 el subtipo de hiperactividad/impulsividad y 115 el subtipo combinado. Los porcentajes en los que se presenta cada subtipo se asemejan bastante a los porcentajes globales estimados en Faraone et al. (1998). Se registraron también los antecedentes del tratamiento mediante una variable binaria que indica si el paciente ha sido o no tratado con anterioridad; de los 199 pacientes, 142 ya habían recibido previamente un tratamiento para el TDAH.

Edad	Mínima	6
	Máxima	18
	Media	11
	Mediana	10.57
Sexo	Femenino	41
	Masculino	158
Subtipo	Déficit atención	75
	Hiperactividad	9
	Combinado	115
Tratamiento previo	No	57
	Sí	142

		3 meses	6 meses	12 meses
Fármaco	MTF-LI	16	12	10
	MTF-LP	183	151	164
	NA	0	36	25
Dosis (mg/día)	Mínima	5	10	10
	Máxima	81	81	81
	Media	35.48	39.66	39.31

Tabla 1.1: Tabla resumen de los datos de los 199 niños diagnosticados con TDAH.

El fármaco más utilizado en España para el tratamiento del TDAH es el metilfenidato (MTF), que es un fármaco psicoestimulante. Existen diferentes fórmulas de presentación disponibles: MTF de liberación inmediata (MTF-LI) y MTF de liberación prolongada (MTF-LP), siendo esta última la más utilizada. Cada uno de los niños recibe una de estas dos formulaciones del fármaco en una determinada dosis diaria, medida en miligramos/día.

Con el fin de evaluar la respuesta farmacológica de los pacientes se recogió una serie de información en las revisiones llevadas a cabo a los 3, 6 y 12 meses de seguimiento. En concreto, en esos tres instantes temporales se recogió información de la formulación del fármaco suministrado a cada paciente, de la dosis y de la presencia o ausencia de una serie de efectos adversos. Los efectos adversos considerados fueron insomnio, falta de apetito, alteraciones gastrointestinales, cefaleas, alteraciones emocionales, alteraciones conductales y alteraciones cognitivas. Se trata de variables binarias, codificadas como 0 (ausencia) y 1 (presencia).

Todas las variables de las que hemos hablado hasta ahora, debido a su naturaleza, son fenotípicas. No debemos olvidar que la finalidad de nuestro análisis es estudiar el efecto de la variabilidad genética en la respuesta a fármacos y el riesgo de sufrir efectos adversos. Por esta razón es necesario tener en cuenta también variables genotípicas. Para cada individuo se sabe cuál es su combinación de alelos (esto es, su genotipo) para un determinado número de SNPs. De todos los SNPs de los cuales tenemos información, en nuestro estudio nos vamos a centrar únicamente en aquellos que están en el cromosoma 22. El cromosoma 22 es el cromosoma no sexual o autosoma más pequeño; está compuesto por alrededor de 51 millones de pares de bases, representando así entre el 1.5 y 2 % del total de ADN. De esta forma, disponemos de la información de 17377 SNPs.

1.3. Estudios de asociación de genoma completo

Un estudio de asociación de genoma completo (comúnmente conocido como GWAS por sus siglas en inglés, *genome wide association study*) es un estudio de asociación de rastreo genómico que se basa en la utilización de un número considerable de SNPs, con el propósito de descubrir si están asociados a la variable respuesta de interés. A pesar de todos los esfuerzos realizados la realidad es que todavía no se conoce bien la base genética de muchas enfermedades complejas. Gracias a los GWAS se han identificado asociaciones estadísticamente significativas entre cientos de SNPs y enfermedades complejas comunes, como son la diabetes tipo 2, el cáncer de próstata, el cáncer de mama y la esquizofrenia (Xue et al., 2018; Chung et al., 2010; Dennison et al., 2020).

Para estudiar el posible efecto de un SNP en un determinado efecto adverso, teniendo en cuenta además ciertas covariables como pueden ser el sexo del paciente, la edad o la dosis de tratamiento que recibe, nos interesarán modelos del siguiente tipo:

$$\text{Efecto adverso} = \text{SNP} + \text{covariables}.$$

Además, al hacer un GWAS no solo estaremos interesados en comprobar la influencia de un único SNP, sino de una colección. Nos interesarán entonces los siguientes modelos:

$$\text{Efecto adverso} = \text{SNP}_m + \text{covariables},$$

donde $m = 1, \ldots, M$, siendo M un número elevado. Dependiendo de la naturaleza de la variable respuesta, así como de la estructura de los datos, nos convendrá ajustar un tipo de modelo de regresión u otro. En algunas ocasiones convendrá ajustar modelos de regresión lineales (por ejemplo, cuando la variable respuesta de efecto adverso es continua), mientras que en otras será más adecuado utilizar modelos de regresión logística (cuando la variable respuesta es dicotómica, por ejemplo indicando simplemente la presencia o ausencia del efecto adverso). Por otra parte, ante determinadas estructuras de

dependencia en los datos será preferible el uso de modelos mixtos, que permiten incluir tanto efectos fijos como efectos aleatorios.

En cuanto a los SNPs, se trata de polimorfismos mayoritariamente bialélicos. Sea B el alelo[1] más frecuente y b el alelo menos frecuente, con $B, b \in \{\text{A, C, G, T}\}$. Dependiendo de la combinación del alelo heredado de la madre y del alelo heredado del padre (es decir, del genotipo) un individuo puede ser homocigoto para el alelo B si en ambos cromosomas aparece el alelo B, homocigoto para el alelo b si en ambos cromosomas aparece el alelo b, y heterocigoto cuando en un cromosoma aparece el alelo B y en el otro el alelo b. Típicamente en un GWAS la asociación entre cada SNP y la variable respuesta de interés se prueba bajo un modelo genético específico que puede asumir un determinado modo de herencia. El modelo codominante asume que puede haber un riesgo distinto en los individuos heterocigotos Bb y homocigotos bb respecto a los homocigotos BB (cada genotipo proporciona un riesgo de enfermedad diferente y no aditivo). Para algunas enfermedades complejas se puede asumir que el efecto para los individuos heterocigotos Bb es la mitad que para los homocigotos bb. Este modelo se conoce como modelo aditivo y lo que hace es considerar que el heterocigoto está justo a medio camino, lo cual muchas veces tiene sentido biológico. El modelo aditivo se conoce también como test alélico porque los números 0, 1 y 2 indican el número de veces que aparece el alelo menos frecuente. Sin embargo, algunas veces se espera que nuestro SNP siga otros patrones de herencia como el dominante o el recesivo. En otras palabras, a veces se asume que basta con tener un alelo b para conferir riesgo (modelo dominante, tanto Bb como bb tienen el mismo riesgo) o que es necesario tener las dos copias del alelo b (modelo recesivo). Para especificar el modo de herencia en un modelo de regresión tan solo necesitamos codificar adecuadamente la información genotípica tal y como se muestra en la Tabla 1.2.

	BB	Bb	bb	tipo de variable
codominante	0	1	2	factor
aditivo	0	1	2	numérica
dominante	0	1	1	factor
recesivo	0	0	1	factor

Tabla 1.2: Codificación de los modos genéticos de herencia.

Existen evidencias empíricas que sostienen la idea de que tanto los factores genéticos como ambientales afectan a las enfermedades comunes. Por esta razón, muchas veces interesa realizar los análisis de asociación entre el SNP y el efecto adverso teniendo en cuenta otras covariables. Así, en cada uno de los modelos de regresión considerados en un GWAS es común considerar otras variables explicativas además de la información genotípica del SNP.

Tras el ajuste de los M modelos de regresión del GWAS (uno por cada SNP de interés) podemos obtener los M p-valores relativos al contraste de significación de los M SNPs. Puesto que estamos ante un caso de test múltiples no es adecuado utilizar los niveles de significación usuales para determinar si un SNP es estadísticamente significativo o no, sino que es necesario corregirlos. Debido a esta multiplicidad, una de las limitaciones de los GWAS sin corrección se encuentra en la acumulación de falsos positivos (un falso positivo se da cuando se rechaza la hipótesis nula de ausencia de asociación siendo en realidad cierta). Además, otros problemas que pueden presentar son la falta de información

[1] Alelo: variante de una secuencia de ADN en un cierto locus (el locus indica la posición en el cromosoma). Un individuo hereda dos alelos, uno del padre y el otro de la madre.

de la función de algunos genes, los posibles sesgos debido a la mala selección de casos y controles y los errores de genotipificación. También es difícil identificar las interacciones gen-ambiente. Hay que tener en mente que en un GWAS se consideran solo algunos SNPs de todos los que hay en nuestro genoma; en la actualidad se estima que tenemos aproximadamente 3 millones, pero el caso es que todavía queda mucho trabajo por hacer en el estudio del genoma humano.

Es importante destacar que los GWAS solo nos proporcionan SNPs sospechosos o candidatos, que deben confirmarse mediante otros experimentos. Esto quiere decir que es posible que encontremos algún SNP estadísticamente significativo pero que no dispongamos de una explicación biológica que nos permita entender dicha relación. De nuevo, esto podría ser reflejo del conocimiento todavía limitado que tenemos sobre el genoma humano.

Además, en el caso de encontrar un SNP estadísticamente significativo lo ideal sería poder utilizarlo en la práctica clínica como biomarcador, de modo que antes de darle un tratamiento a un paciente fuese aconsejable realizarle un estudio genético para conocer qué información genotípica posee en dicha posición genómica. De ese modo, se podría decidir si el tratamiento es adecuado para él o si por el contrario tendrá una elevada probabilidad de sufrir efectos adversos al recibirlo. La realidad es que el paso del descubrimiento de un SNP significativo a la práctica clínica es complicado; se trata de una transición compleja y larga. Las agencias reguladoras de medicamentos (la *European Medicament Agency* o EMA en Europa y la *Food and Drug Administration* o FDA en Estados Unidos) se encargan de la validación y aprobación de los biomarcadores. Dicha información se puede consultar a través de la página del *Pharmacogenetics and Pharmacogenomics Knowledge Database* (PharmGKB), donde es posible consultar las fichas técnicas de los medicamentos y las recomendaciones de dosis propuestas para cada uno de ellos[2]. Podemos ver que, en función de su impacto, los biomarcadores se clasifican como obligatorios, recomendables, justificables e informativos.

1.4. Objetivos y organización del trabajo

En este trabajo se llevará a cabo un GWAS con los datos que disponemos de niños diagnosticados con TDAH, con el objetivo de encontrar algún SNP que pueda estar asociado a los efectos adversos que puede desencadenar el tratamiento prescrito con más frecuencia para dicho trastorno, que es el metilfenidato.

A pesar de que la condición clínica de la mayoría de los pacientes tratados con metilfenidato mejora, alrededor de un 35 % no responde al tratamiento o sufre efectos adversos (Hodgkins at al., 2012; Johnston et al., 2015). Entre los efectos adversos, los más frecuentes son la falta de apetito (que puede incluso desencadenar anorexia) y el insomnio. Si fuésemos capaces de encontrar alguna variación genética que esté relacionada con esos efectos adversos, quizás podríamos llegar a predecir el riesgo que tiene cada niño de sufrirlos en caso de recibir dicho tratamiento. De este modo, en caso de detectar que un niño tiene una elevada probabilidad de responder de forma negativa, sería aconsejable buscar para él una alternativa terapéutica más adecuada.

Tal y como ya hemos comentado, el tipo de modelo de regresión utilizado en un GWAS depende de la naturaleza de la variable respuesta y de la estructura de los datos. En el Capítulo 2 nos centraremos en el caso de modelos de regresión logística, adecuados cuando la variable respuesta es binaria. Ajustaremos dichos modelos a los datos de TDAH para estudiar cada instante temporal (3, 6 y 12 meses de seguimiento) por separado. Para ello utilizaremos el software estadístico `R`, en concreto la función `mlreg` de la librería `GenABEL`. Esta función también nos permitirá representar un tipo de gráfico muy utilizado en estudios de asociación de genoma completo, conocido como *Manhattan plot*, que

[2] `www.pharmgkb.org/search/labelList.action`

nos permitirá ver fácilmente los SNPs para los cuales se ha obtenido significación estadística. Además, introduciremos los principales métodos de ajuste del nivel de significación para el caso de test múltiples y aplicaremos la metodología desarrollada a nuestros datos. Debido a que los datos con los que trabajamos presentan una estructura de dependencia que viene determinada por el seguimiento temporal de cada individuo, se hace adecuado recurrir a los modelos de regresión mixtos, que se desarrollan en el Capítulo 3 (lineales) y 4 (generalizados). Puesto que en nuestro conjunto de datos de TDAH no disponemos de ninguna variable respuesta continua, para ilustrar los contenidos del Capítulo 3 nos ayudaremos del conjunto de datos `sleepstudy` de la librería `lme4` de `R`. Para el ajuste de los modelos correspondientes consideraremos la función `lmer` de esa misma librería. En el Capítulo 4 retomaremos los datos de TDAH para ilustrar los conceptos introducidos y ajustaremos los modelos con la función `glmer`, también de la librería `lme4`. El Capítulo 5 se dedica a la aplicación de los modelos mixtos lineales generalizados a los datos de TDAH. Puesto que no existe ninguna función en `R` que represente los *Manhattan plots* en el caso de modelos mixtos, programaremos el código necesario para obtener dichos gráficos. Finalmente, el Capítulo 6 incluye las conclusiones del trabajo. Cabe destacar que todos nuestros estudios se han llevado a cabo considerando únicamente la información genética del cromosoma 22. Sería de gran utilidad hacer lo mismo teniendo en cuenta toda la información del genoma, pero debido al coste computacional que eso conlleva se necesitaría hacer uso de las instalaciones del Centro de Supercomputación de Galicia (CESGA).

Capítulo 2

Regresión logística en un GWAS

Los modelos de regresión logística se utilizan cuando la variable respuesta es dicotómica o binaria. Es habitual que las variables de efecto adverso estén codificadas de forma binaria, indicando simplemente su presencia o ausencia. Ya hemos comentado que en nuestro caso práctico las variables de efecto adverso, entre las cuales se encuentran insomnio y falta de apetito, son dicotómicas. Por lo tanto, parece que el ajuste de modelos logísticos en el GWAS podría ser adecuado.

En este capítulo revisaremos brevemente el modelo de regresión logística, para posteriormente ajustar este tipo de modelos en un GWAS usando nuestros datos. También introduciremos varios métodos de ajuste del nivel de significación para test múltiples.

2.1. Formulación del modelo de regresión logística

El modelo de regresión logística es un modelo de regresión que explica el comportamiento de una variable dependiente discreta dicotómica o binaria en función de una o más variables independientes. La variable respuesta Y toma dos posibles valores: 0 = fracaso y 1 = éxito. Este tipo de variables aparecen frecuentemente en el ámbito biomédico, indicando por ejemplo la presencia o ausencia de una patología o de un efecto adverso. Mediante el ajuste de un modelo de regresión logística se pueden identificar factores de riesgo y factores de protección para la patología o efecto adverso en cuestión, además de estimar cuánto aumenta la probabilidad de padecerlos si se dan una serie de características o condiciones.

Por ser la variable respuesta binaria, su distribución es Bernoulli y su media la probabilidad de éxito:

$$E(Y) = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = P(Y = 1).$$

Al igual que en los modelos lineales de regresión, se supone que la media de Y depende de los valores de X, siendo X un vector columna con una o varias variables explicativas. Esto es, se plantea la construcción de un modelo para

$$\pi(x) = E(Y|X = x) = P(Y = 1|X = x).$$

Supondremos que π depende de un vector de parámetros β (vector columna de la misma dimensión que X) que podemos estimar, por lo que escribiremos $\pi(x, \beta)$ para destacar esta dependencia. Por ser una probabilidad, $\pi(x, \beta)$ toma valores en el intervalo $[0, 1]$.

Debido a la naturaleza binaria de la variable respuesta y a que $\pi(x, \beta)$ toma valores en el intervalo $[0, 1]$, no es posible construir un modelo lineal ni para Y ni para $\pi(x, \beta)$, ya que se inclumplirían las

suposiciones básicas de linealidad, homocedasticidad y normalidad. A continuación veremos que podemos solventar este problema mediante una transformación de la probabilidad de éxito que permita transofrmar el intervalo $[0, 1]$ en toda la recta real.

Para una variable binaria se define la Odds como el cociente entre la probabilidad de éxito y la probabilidad de fracaso. En nuestro caso,

$$\text{Odds}(Y|X = x) = \frac{\pi(x, \beta)}{1 - \pi(x, \beta)}.$$

Se trata de una forma alternativa de parametrizar el modelo de distribución de $Y|X = x$, y es una cantidad que se mueve en el intervalo $[0, +\infty)$.

- Odds $= 1$: corresponde a una probabilidad de éxito de 0.5.
- Odds > 1: la probabilidad de éxito es mayor que la de fracaso.
- Odds < 1: la probabilidad de éxito es menor que la de fracaso.

Si aplicamos un logaritmo sobre la Odds obtendremos una cantidad que puede tomar cualquier valor real, y que por lo tanto es susceptible de ser explicada mediante un modelo de regresión lineal.

En definitiva, el modelo logístico se basa en considerar un modelo lineal sobre una transformación de la probabilidad de éxito, de manera que se transforme el intervalo $[0, 1]$ en toda la recta real. Así, se tiene que

$$g(\pi(x, \beta)) = x'\beta,$$

donde g es una función link, que en el caso del modelo logístico es la función logit: $g(\pi(x, \beta)) = \log\left(\frac{\pi(x,\beta)}{1-\pi(x,\beta)}\right)$. Sin más que invertir la función logit podemos expresar la probabilidad de éxito como

$$\pi(x, \beta) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}.$$

2.1.1. Interpretación de los parámetros del modelo

Una de las ventajas del uso de modelos de regresión logística es la facilidad de interpretación de los parámetros en términos de las Odds y Odds Ratio (OR). Ya hemos visto que para una variable dicotómica la Odds se calcula como el cociente entre la probabilidad de éxito y la probabilidad de fracaso. De esta forma, es posible obtener el valor de la Odds a partir de la probabilidad de éxito y viceversa.

Supongamos que tenemos una única variable explicativa y que es discreta binaria ($X_1 = 0$ o $X_1 = 1$). En tal caso, $x = (1, x_1)'$ y $\beta = (\beta_0, \beta_1)$, siendo x_1 el valor que toma la variable X_1, luego $x'\beta$ se reduce a $\beta_0 + \beta_1 x_1$. De esta forma,

$$\pi(x, \beta) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}.$$

La Odds en el grupo de referencia ($X_1 = 0$) viene dada por $\exp(\beta_0)$:

$$\text{Odds}(Y|X_1 = 0) = \frac{P(Y = 1|X_1 = 0)}{P(Y = 0|X_1 = 0)} = \frac{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}}{1 - \frac{\exp(\beta_0)}{1+\exp(\beta_0)}} = \exp(\beta_0).$$

Además, $\exp(\beta_1)$ es la OR, que representa el cociente de las Odds entre la población $X_1 = 1$ y la población $X_1 = 0$. Esto es, la OR indica por cuánto hay que multiplicar la Odds del grupo de referencia para obtener la Odds del otro grupo.

$$\frac{\text{Odds}(Y|X_1 = 1)}{\text{Odds}(Y|X_1 = 0)} = \frac{\frac{P(Y=1|X_1=1)}{P(Y=0|X_1=1)}}{\exp(\beta_0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

En caso de ser X_1 continua la Odds en el grupo de referencia ($X_1 = 0$) es $\exp(\beta_0)$, mientras que en la OR las poblaciones que se consideraran en el cociente corresponden a valores de X_1 separados por una unidad. Dicho con otras palabras, $\exp(\beta_1)$ representa la OR debida a haber incrementado la variable explicativa en una unidad.

2.1.2. Estimación de los parámetros del modelo

La estimación de los parámetros del modelo de regresión logística se hace mediante el método de máxima verosimilitud. Este método da lugar a unas ecuaciones de verosimilitud que en general no tienen solución explícita, por lo que se requiere de métodos numéricos para el cálculo de las estimaciones.

Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra aleatoria simple de (X, Y), con $Y|X = x \sim Ber(\pi(x, \beta))$. Se tiene que y_i es el valor observado en la variable Y para el i-ésimo individuo y x_i es un vector que contiene sus valores en las $p-1$ variables explicativas. Esto es, ahora x_1 ya no denota el valor que toma la primera variable explicativa, como en el apartado anterior, sino que $x_1 = (1, x_{11}, x_{12}, \ldots, x_{1p-1})'$ contiene los valores observados en las $p - 1$ variables explicativas para el individuo 1. Nótese que se añade un 1 al principio del vector para permitir una constante como parámetro del modelo. La función de verosimilitud, suponiendo que los y_i son independientes, adopta la forma

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \Big[\pi(x_i, \beta)^{y_i}(1 - \pi(x_i, \beta))^{1-y_i}\Big] \tag{2.1}$$

y su logaritmo es

$$\log \mathcal{L}(\beta) = l(\beta) = \sum_{i=1}^{n} \Big[y_i \log \pi(x_i, \beta) + (1 - y_i) \log(1 - \pi(x_i, \beta))\Big].$$

El gradiente del logaritmo de la verosimilitud respecto de β es

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \pi(x_i, \beta)}{\partial \beta} \cdot \frac{y_i - \pi(x_i, \beta)}{\pi(x_i, \beta)(1 - \pi(x_i, \beta))}.$$

Nótese que estamos abusando de la notación, ya que $\beta = (\beta_0, \ldots, \beta_{p-1})$ es un vector. Lo que realmente se hace es la derivada parcial con respecto a cada una de las componentes de β.

Puesto que $\pi(x, \beta) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$, se tiene que

$$\frac{\partial \pi(x, \beta)}{\partial \beta} = x'\pi(x, \beta)(1 - \pi(x, \beta)),$$

y de esta forma

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i'[y_i - \pi(x_i, \beta)] = \sum_{i=1}^{n} x_i' y_i - x_i'\pi(x_i, \beta).$$

Si igualamos la expresión anterior a cero se obtienen las ecuaciones de verosimilitud. Se trata de ecuaciones que no admiten solución explícita, ya que $\pi(x_i, \beta)$ no es función lineal de β. Por tanto,

son necesarios métodos numéricos para su resolución, como por ejemplo Newton-Raphson. En `R`, la función habitual para ajustar este tipo de modelos considera el método IRLS (*iteratively reweighted least squares*) para el cálculo de los estimadores. Para más información sobre este método, consultar Venables y Ripley (2002, página 185).

2.1.3. Contraste de los parámetros del modelo

Sea $\hat{\beta}$ el estimador de máxima verosimilitud de β. La teoría asintótica sobre los estimadores de máxima verosimilitud nos permite aproximar la distribución del estimador mediante una distribución normal de la siguiente manera:

$$\hat{\beta} - \beta \sim N_p(0, I(\hat{\beta})^{-1}),$$

donde $I(\hat{\beta})$ es la matriz de información de Fisher evaluada en $\hat{\beta}$:

$$I(\hat{\beta}) = E\left(- \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\Big|_{\beta=\hat{\beta}} \right).$$

Una vez estimado el vector $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$, nos puede interesar realizar contrastes de hipótesis relativos a alguna de las p componentes. Por ejemplo, nos podría interesar contrastar si la componente β_1 se puede considerar como nula (lo que se traduciría en que se podría prescindir de la covariable X_1 en el modelo). Para realizar este contraste de hipótesis podemos considerar el estadístico

$$\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \sim N(0,1),$$

que se conoce como estadístico de Wald. El error típico de $\hat{\beta}_1$, $\hat{\sigma}(\hat{\beta}_1)$, se puede obtener directamente a partir de la diagonal de la matriz $I(\hat{\beta})^{-1}$. En particular, $\hat{\sigma}(\hat{\beta}_1)$ es la raíz cuadrada del segundo elemento de dicha diagonal. En base a este estadístico podemos ofrecer el nivel crítico para el coeficiente y su correspondiente p-valor. Esto lo aplicaremos en la siguiente sección, en la que nos centraremos en estudiar la significación de cada uno de los SNPs en los modelos ajustados.

2.2. Aplicación a datos reales

Retomando los datos de niños diagnosticados con TDAH, el interés se centra en el ajuste de los siguientes M modelos de regresión logística:

$$\text{logit}(Y) = \beta_0 + \beta_1 \text{ SNP}_m + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1},$$

donde Y es la variable respuesta de efecto adverso y $X_2, \ldots, X_{p-1}$ son las covariables incluidas en el modelo, además del m-ésimo SNP, con $m = 1, \ldots, M$. Esto es, el SNP_m juega el papel de la primera covariable introducida en el modelo (X_1).

Más que centrarnos en la interpretación de la Odds y OR, que es lo que se hace habitualmente tras ajustar un modelo de regresión logística, nos vamos a centrar en los contrastes de significación de los parámetros del modelo. En concreto, para los M modelos ajustados nos interesa contrastar en cada uno de ellos si el coeficiente asociado al SNP, β_1, es significativamente distinto de cero.

Para comenzar, consideraremos como variable respuesta el efecto adverso insomnio a los 3 meses. Tenemos información de 38 niños con insomnio y 126 niños sin tal efecto adverso, como podemos ver

en la Tabla 2.1. Además, hay 35 datos faltantes de los cuales no tenemos información sobre esta variable.

		3 meses	6 meses	12 meses
Insomnio	No	126	144	148
	Sí	38	29	20
	NA	35	26	31
Falta apetito	No	124	119	139
	Sí	40	53	29
	NA	35	27	31

Tabla 2.1: Datos correspondientes a los efectos adversos insomnio y falta de apetito.

En función del modo de herencia que consideremos (codominante, aditivo, dominante o recesivo), podremos ajustar diferentes modelos. En este trabajo nos centraremos en el modelo aditivo, que es el más utilizado. Además, recordemos que podemos incluir distintas covariables de interés en los modelos. En la Tabla 1.1 podemos ver que disponemos de información sobre las siguientes variables: edad, sexo, subtipo de TDAH, antecedente de tratamiento, fármaco y dosis. Puesto que nos estamos centrando en el insomnio a los 3 meses, para fármaco y dosis tendremos en cuenta las medidas tomadas a los 3 meses de seguimiento. En este contexto surge de forma natural la pregunta de si todas esas covariables deben entrar en el modelo. En caso negativo, el interés se centraría en saber cuáles deben entrar y cuáles no. El objetivo es conseguir un buen ajuste considerando el menor número de variables posible. Por lo tanto, debemos hacer una selección de variables para posteriormente considerar los modelos más adecuados y proceder a los contrastes de significación de los SNPs, que es el objetivo final. El método que vamos a utilizar para la selección de variables es el de selección por pasos. Existen tres alternativas distintas de este método:

- Selección progresiva (*forward*). Se parte de una situación en la que no hay ninguna covariable y en cada paso se incluye una aplicando un cierto criterio de entrada, hasta que ninguna de las restantes lo verifican.
- Eliminación progresiva (*backward*). Se parte del modelo con todas las covariables y en cada paso se elimina una aplicando un criterio de salida, hasta que ninguna de las incluidas lo verifican.
- Paso a paso (*stepwise*). Combina un criterio de entrada y uno de salida. Normalmente se parte de un modelo sin covariables y en cada paso puede haber una inclusión o una exclusión (*forward/backward*).

En nuestro caso hemos decidido optar por la selección paso a paso o *stepwise*, y en la práctica lo hemos llevado a cabo con ayuda de `R`. Es importante tener en cuenta cuál es la condición considerada para suprimir o incluir una variable en cada paso. El criterio que nosotros hemos utilizado es el criterio global de la información de Akaike o AIC, que se basa en la verosimilitud y en el número de parámetros del modelo. El primer paso consiste en partir del modelo más simple, que en nuestro caso es el modelo que incluye como única variable explicativa la información genotípica del SNP. A partir de aquí, se añadirá a dicho modelo aquella covariable que consiga que el AIC disminuya. De esta forma en cada paso se estudia la introducción progresiva de variables, planteándose también si todas las covariables

añadidas hasta ese momento deben permanecer en el modelo. En la Tabla 2.2 podemos ver el número de modelos (de los 17377 totales) en los que aparece cada una de las covariables, tras haber hecho la selección. En base a esa información podemos concluir que en el caso de insomnio a los 3 meses parece que lo más adecuado es el modelo simple, esto es, el que incluye como única covariable la información genotípica del SNP. Hay que tener en cuenta que no solo disponemos de información de insomnio a los 3 meses, sino que también se tienen medidas correspondientes a los 6 meses y a los 12 meses. En ambos escenarios parece que el mejor modelo, además de la información genotípica, ha de incluir la covariable fármaco (considerando el fármaco que el individuo recibe a los 6 y a los 12 meses, respectivamente). Además, se dispone de una variable resumen de insomnio, que nos indica la presencia o ausencia del efecto adverso en alguno de los tres instantes temporales estudiados. En tal caso lo más adecuado es el modelo simple, como en el caso de insomnio a los 3 meses. Nótese que cuando consideramos el efecto adverso insomnio global no tenemos en cuenta el fármaco ni la dosis, pues para ambas covariables la información que tenemos corresponde a los instantes temporales 3, 6 y 12 meses de seguimiento. Si bien uno puede plantearse hacer una selección de variables más general (incondicional al SNP, véase Apéndice A para este enfoque), puesto que nuestro interés se centra en estudiar la significación de cada SNP resulta más adecuada la selección condicional que hemos hecho.

	SNP (como única covariable)	edad	sexo	subtipo	tratamiento previo	fármaco	dosis
Insomnio 3 meses	17022	0	1	174	0	1	249
Insomnio 6 meses	295	0	0	0	6	17081	2
Insomnio 12 meses	91	23	0	1085	0	17284	5
Insomnio global	17375	0	0	2	0	-	-

Tabla 2.2: Número de modelos (de un total de 17377) en los que aparece cada una de las covariables tras haber hecho la selección variables paso a paso.

Una vez seleccionadas las covariables para cada uno de los cuatro escenarios, podemos proceder al ajuste de los modelos. Nos seguiremos centrando en insomnio a los 3 meses. En este caso recordemos que en el GWAS consideraremos modelos de regresión logística bajo un modelo de herencia aditivo e incluyendo como única covariable el SNP. Tras el ajuste de los M modelos de regresión correspondientes, que se ha llevado a cabo utilizando `R`, podemos obtener los M p-valores relativos al contraste de significación de los M SNPs. La complejidad de los GWAS se encuentra aquí, ya que se trata de un estudio de múltiples test. Tal y como ya sabemos, si al hacer un contraste de hipótesis consideramos un nivel de significación $\alpha = 0.05$ esto quiere decir que 5 de cada 100 veces podríamos rechazar la hipótesis nula de ausencia de asociación siendo en realidad la hipótesis nula cierta. A esto se le conoce como error tipo I o probabilidad de falso positivo.

	No rechazamos H_0	Rechazamos H_0
H_0 verdadera	Decisión correcta	Error tipo I (falso positivo)
H_0 falsa	Error tipo II (falso negativo)	Decisión correcta

Tabla 2.3: Tipos de errores en contrastes de hipótesis.

En la Tabla 2.3 se muestran los posibles resultados al hacer un contraste de hipótesis. Si se considera un único SNP, y por tanto se realiza un único contraste, la probabilidad de falso positivo es relativamente baja. Sin embargo, si consideramos 1 millón de SNPs podemos encontrarnos con 50000 SNPs asociados a una enfermedad o efecto adverso que en realidad pueden ser falsos positivos. Cuando se contrastan simultáneamente M hipótesis nulas, en cada una de las pruebas pueden ocurrir los dos tipos de errores de la Tabla 2.3. Teniendo en cuenta los M contrastes de forma conjunta, se obtienen los resultados de la Tabla 2.4.

	No rechazamos H_0	Rechazamos H_0	
H_0 verdadera	U	V	M_0
H_0 falsa	T	S	$M - M_0$
Total	$M - R$	R	M

Tabla 2.4: Tipos de errores en contrastes de hipótesis múltiples.

Debido al elevado número de test llevados a cabo en un GWAS es necesario el uso de un umbral de significación adecuado que tenga en cuenta dicha multiplicidad, con el fin de controlar el número de falsos positivos, que se denota por V (número de hipótesis nulas verdaderas que se rechazan a lo largo de los M contrastes). Los métodos de ajuste para test múltiples se centran en controlar alguna de las siguientes tasas de error: la tasa de error global (*family-wise error rate* o FWER) o la tasa de falsos descubrimientos o falsos positivos (*false discovery rate* o FDR).

El FWER se define como la probabilidad de cometer por lo menos un error tipo I en los M test de hipótesis, esto es, FWER $= P(V \geq 1)$. El método más conocido para controlar el FWER es el de Bonferroni, que consiste en tomar para cada test un nivel corregido α/M, siendo M el número de test. Si no usamos ninguna corrección, y suponiendo que $M = M_0$, la probabilidad de cometer al menos un falso positivo, considerando un nivel de significación α para cada uno de los M test y asumiendo independencia entre ellos, es $1 - (1 - \alpha)^M$, que es aproximadamente igual a 1 cuando M es grande. Sin embargo, usando la corrección de Bonferroni se asegura que la probabilidad de obtener al menos un falso positivo es menor o igual que α. El problema de este método es que puede ser demasiado conservador y cuando la cantidad de test es grande los niveles corregidos resultan demasiado bajos, y en consecuencia los p-valores demasiado altos. Esto significa que se seleccionarán pocos (o ninguno) SNPs como candidatos. En nuestro caso $M = 17377$; considerando $\alpha = 0.05$, el nivel de significación corregido sería $\frac{0.05}{17377} = 2.88 \times 10^{-6}$. Es importante destacar que el método de Bonferroni funciona aún cuando no hay independencia, ya que acota la probabilidad de falso positivo en el caso más desfavorable (es decir, cuando todas las hipótesis nulas son ciertas) mediante la cota de la unión: la probabilidad de que exista al menos un falso positivo es menor o igual que la suma de las probabilidades de que en cada uno de los test ocurra un falso positivo, que son todas menores o iguales que $\frac{\alpha}{M}$ y por tanto sumadas no superan el valor α.

Cuando no nos centramos en un solo cromosoma y consideramos el genoma completo el número de test es todavía mayor, y por lo tanto el nivel de significación corregido por el método de Bonferroni es muy pequeño. Con la finalidad de ofrecer una mejora a este método surgió la idea de tener en cuenta el número de SNPs independientes, y de calcular el nivel corregido dividiendo el nivel de significación α entre dicha cantidad. En 2005 el Consorcio Internacional HapMap estimó que el número de SNPs independientes es de 150 por cada 500000 pares de bases en la población europea. El genoma humano está formado por aproximadamente 3300 millones de pares de bases, lo cual se traduce en alrededor de 1000000 de SNPs comunes independientes. Por lo tanto, para $\alpha = 0.05$ y usando la corrección de Bon-

ferroni teniendo en cuenta dicha cantidad, obtenemos un nivel corregido de $0.05/1000000 = 5 \times 10^{-8}$. El estudio de meta-análisis descrito en Jannot et al. (2015) indica que dicho valor se ha convertido en el umbral estándar en GWAS en los que se trabaja con SNPs para poblaciones de ancestralidad europea. No obstante, hay que tener en cuenta que el uso de este umbral tiene importantes limitaciones, ya que se usa independientemente del tamaño muestral y de las frecuencias alélicas. Puesto que en nuestro caso el tamaño muestral es pequeño ($n = 199$), se podrá ver afectada la potencia estadística. Es importante destacar que esta limitación en el tamaño muestral es a veces imposible de corregir, sobre todo en el caso de enfermedades raras, en las cuales el número de afectados es muy reducido. Los datos con los que trabajamos en la práctica son de niños españoles, y además no se han considerado variantes poco frecuentes o raras, por lo que parece que el umbral estándar se adapta a nuestro caso. Sin embargo, hay que tener en cuenta que dicho umbral se utiliza cuando se considera el genoma completo, y nosotros estamos considerando únicamente el cromosoma 22. El cromosoma 22 tiene alrededor de 51 millones de pares de bases, lo cual se traduce en alrededor de 15300 SNPs independientes. Utilizando la corrección de Bonferroni en tal caso, el nivel corregido sería $\frac{0.05}{15300} = 3.27 \times 10^{-6}$. Este nivel es simplemente una adaptación del umbral estándar a nuestro caso, esto es, al cromosoma 22. De este modo tenemos tres niveles de significación para test múltiples: el nivel de Bonferroni (2.88×10^{-6}), el nivel estándar (5×10^{-8}) y el nivel estándar adaptado al cromosoma 22 (3.27×10^{-6}). Es importante destacar que cada uno de estos niveles son globales, en el sentido de que se utiliza un mismo nivel para los M test. Esto no ocurre en el caso de FDR, como veremos un poco más adelante. El hecho de que se use el mismo nivel para todos los test resultará de gran utilidad cuando realicemos una representación gráfica de los M p-valores, ya que simplemente graficando esos tres umbrales podremos ver qué SNPs resultan significativos y cuáles no.

Por otro lado, el FDR se define como la proporción de hipótesis nulas verdaderas que resultan rechazadas dentro del total de hipótesis rechazadas, esto es, $\text{FDR} = E\left(\frac{V}{R}\right)$, donde R es el número total de hipótesis nulas rechazadas. Por lo tanto, se centra en la proporción de veces que un SNP fue declarado significativo cuando en realidad no lo era. FDR es menos exigente que Bonferroni. Benjamini y Hochberg (1995) proponen un ajuste de los p-valores para controlar el FDR, que está basado en el supuesto de independencia de los p-valores. Bajo esta hipótesis, si $p^{(m)}$ denota el m-ésimo p-valor ordenado, $p^{(1)} < \ldots < p^{(M)}$, se rechaza $H_0^{(m)}$ cuando $p^{(m)} < \frac{m}{M}\alpha$. En este enfoque se pide que $p^{(1)} < \frac{\alpha}{M}$, al igual que en Bonferroni, pero el umbral se flexibiliza para los siguientes p-valores. Es decir, en este caso ya no se tiene un único umbral común para los M test. Bajo la hipótesis nula de no existencia de significación estadística $p^{(m)}$ debería de comportarse como $U_{(m)}$, donde $U_{(1)}, \ldots, U_{(M)} \sim U[0,1]$. El ajuste propuesto por Benjamini y Hochberg supone la independencia de los p-valores; debido al desequilibrio de ligamiento, fenómeno por el cual sabemos que hay SNPs que están correlacionados, no es correcto asumir que los test son independientes. Por esta razón es importante tener en cuenta la posible dependencia espacial de los SNPs. El método propuesto por Benjamini y Heller (2007) permite controlar el FDR teniendo en cuenta dicha dependencia espacial, considerando que se conoce a priori una forma de agrupar los datos (en nuestro caso, que somos capaces de establecer diferentes grupos homogéneos de SNPs, no necesariamente del mismo tamaño). El método propuesto consiste en un procedimento jerárquico, de modo que el primer paso se basa en obtener un único p-valor para cada uno de los grupos. Después, para aquellos grupos para los que se rechace la hipótesis nula de ausencia de asociación, se calculan los p-valores intra-grupo. Desafortunadamente, no disponemos de información suficiente para establecer grupos homogéneos de los SNPs considerados, luego no podremos utilizar este método en la práctica. No obstante, en caso de disponer de dicha información quizás sería el método más adecuado para trabajar con nuestros datos.

Tras el ajuste de los M modelos de regresión y de la obtención los M p-valores relativos al contraste de significación de los SNPs se suele utilizar un tipo de gráfico, denominado *Manhattan plot*, con el fin de mostrar aquellos SNPs significativos. En el eje de abscisas de un *Manhattan plot* se representan las coordenadas genómicas (es decir, la localización en el genoma de cada SNP), mientras que en el eje

de ordenadas se representan los $-\log_{10}(p\text{-valor})$, teniendo en cuenta el p-valor relativo al contraste de significación de cada SNP. Por lo tanto, cada punto de un *Manhattan plot* corresponde a un SNP. Tal y como ya hemos comentado, los tres umbrales globales que hemos calculado (2.33×10^{-6}, 5×10^{-8} y 3.27×10^{-6}) nos ofrecen una ventaja a nivel gráfico. Si trabajamos con el umbral de Bonferroni un SNP será significtivo cuando su p-valor asociado sea menor que 2.33×10^{-6} o, equivalentemente, cuando $-\log_{10}(p\text{-valor})$ sea mayor que 5.54. Si trabajamos con el umbral estándar el p-valor deberá ser menor que 5×10^{-8} o $-\log_{10}(p\text{-valor})$ mayor que 7.3. Por último, trabajando con el umbral estándar corregido para el cromosoma 22 el p-valor debe ser menor que 3.27×10^{-6} o $-\log_{10}(p\text{-valor})$ mayor que 5.49. Por tanto, una vez representado el *Manhattan plot*, consideraremos que la asociación de un SNP al efecto adverso de interés es significativa cuado el punto correspondiente toma un valor mayor que 5.54, 7.3 o 5.49 en el eje de ordenadas, dependiendo de la corrección múltiple utilizada. Cuanto más alto esté un punto en el *Manhattan plot*, mayor será la asociación del SNP en cuestión con la variable respuesta. Para estos tres casos que hemos comentado el umbral de significación es el mismo para todos los test; sin embargo, si utilizamos FDR el umbral será distinto para cada uno, de modo que en dicho caso ya no podremos trazar una línea horizontal en el *Manhattan plot* para saber si un SNP es o no significativo. En tal caso, para determinar si un SNP es o no significativo ya no nos será útil la observación directa de dicha representación gráfica, sino que deberemos calcular el nivel correspondiente a cada uno de los M test tal y como se indica en Benjamini y Hochberg (1995).

En la Figura 2.1 se muestran los *Manhattan plot* correspondientes al efecto adverso insomnio, teniendo en cuenta la información recogida a los 3, 6 y 12 meses y la información global. A simple vista podemos ver que, desafortunadamente, ninguno de los puntos supera el valor 5.49 en el eje de ordenadas, y en consecuencia tampoco supera los valores 5.54 ni 7.3. Por tanto, ningún SNP pasa ninguno de los tres umbrales de significación globales considerados. En cuanto al FDR, las conclusiones son las mismas: no se ha encontrado ningún SNP que sea significativo tras utilizar este tipo de corrección múltiple.

En la Tabla 2.5, se muestran los nombres de los cinco SNPs con p-valor más pequeño en cada uno de los cuatro casos, así como sus p-valores. Podemos ver que el p-valor más pequeño se obtiene para el SNP kgp10126194 en el caso del efecto adverso insomnio global, con un valor de 1.62×10^{-5}.

A modo de curiosidad cabe destacar que en los *Manhattan plot* podemos observar que hay una zona sin puntos, la cual corresponde al centrómero del cromosoma 22 (en los centrómeros no hay SNPs).

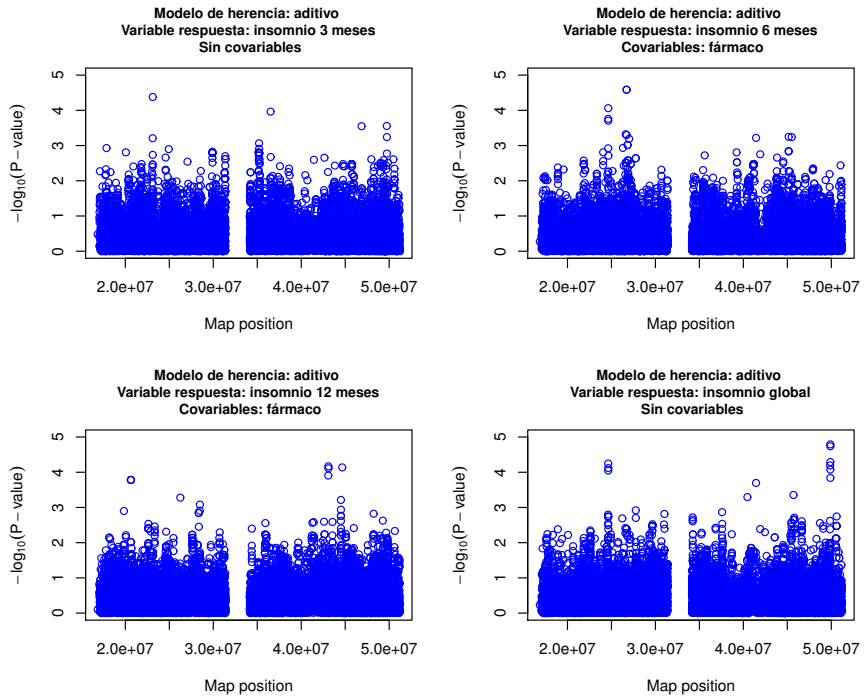


Figura 2.1: *Manhattan plot* de los modelos aditivos (tras la selección de covariables condicional al SNP) para el efecto adverso insomnio a los 3, 6 y 12 meses y para insomio global.

Insomnio 3 meses		**Insomnio 6 meses**		**Insomnio 12 meses**		**Insomnio global**	
SNP	p-valor	SNP	p-valor	SNP	p-valor	SNP	p-valor
kgp12564095	4.18×10^{-5}	kgp12038899	2.6×10^{-5}	kgp737339	6.82×10^{-5}	kgp10126194	1.62×10^{-5}
kgp8094871	1.09×10^{-4}	rs9613560	2.6×10^{-5}	kgp832782	7.28×10^{-5}	kgp9959058	1.83×10^{-5}
kgp12566813	2.78×10^{-4}	rs9613208	8.7×10^{-5}	rs130392	7.71×10^{-5}	kgp4888280	5.13×10^{-5}
kgp78329	2.83×10^{-4}	kgp2320320	1.74×10^{-4}	kgp1072134	1.24×10^{-4}	rs9613208	5.65×10^{-5}
rs8138195	5.74×10^{-4}	rs9613212	1.97×10^{-4}	rs3747084	1.65×10^{-4}	rs8136911	6.42×10^{-5}

Tabla 2.5: SNPs con los p-valores más pequeños de los modelos aditivos (tras la selección de covariables condicional al SNP) para el efecto adverso insomnio a los 3, 6 y 12 meses y para insomnio global.

Se ha repetido el mismo estudio, pero teniendo en cuenta ahora la variable respuesta de falta de apetito. Para comenzar, se ha hecho una selección de variables condicional al SNP, obteniendo los resultados de la Tabla 2.6.

	SNP (como única covariable)	edad	sexo	subtipo	tratamiento previo	fármaco	dosis
Falta apetito 3 meses	103	0	3	9	205	17257	85
Falta apetito 6 meses	0	607	0	0	17377	17365	207
Falta apetito 12 meses	1	17341	4	0	0	17372	6
Falta apetito global	194	17183	0	0	0	-	-

Tabla 2.6: Número de modelos (de un total de 17377) en los que aparece cada una de las covariables tras haber hecho la selección de variables paso a paso.

En base a esos resultados parece que lo más adecuado a los 3 meses es considerar la covariable fármaco, a los 6 meses fármaco y antecedente de tratamiento, y a los 12 meses edad y fármaco. En el caso de falta de apetito global, lo más adecuado es considerar la edad. Una vez ajustados los M modelos en cada uno de esos cuatro escenarios se han obtenido los *Manhattan plot* de la Figura 2.2, en la cual podemos ver que ningún SNP supera los umbrales globales que estamos considerando. Si tenemos en cuenta la corrección FDR la conclusión es la misma: no se encuentra ningún SNP estadísticamente significativo.

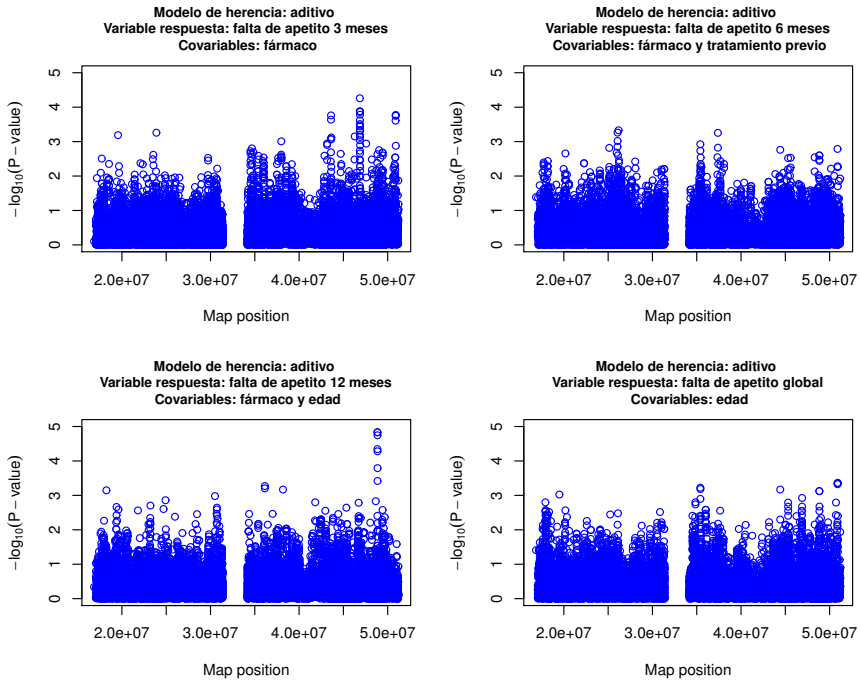


Figura 2.2: *Manhattan plot* de los modelos aditivos (tras la selección de covariables condicional al SNP) para el efecto adverso falta de apetito a los 3, 6 y 12 meses y para falta de apetito global.

A continuación podemos ver que el p-valor más pequeño, con un valor de 1.6×10^{-5}, se ha obtenido en el caso de falta de apetito a los 12 meses para el SNP rs12172263.

Falta apetito 3 meses		**Falta apetito 6 meses**		**Falta apetito 12 meses**		**Falta apetito global**	
SNP	p-valor	SNP	p-valor	SNP	p-valor	SNP	p-valor
kgp12106191	1.06×10^{-4}	kgp11624734	3.84×10^{-4}	rs12172263	1.60×10^{-5}	kgp22793223	3.83×10^{-4}
kgp4932021	1.78×10^{-4}	kgp3761176	4.55×10^{-4}	kgp2396555	1.60×10^{-5}	kgp22820108	3.93×10^{-4}
rs6520023	1.78×10^{-4}	rs9610624	7.23×10^{-4}	rs4823779	2.2×10^{-5}	rs2236030	4.13×10^{-4}
rs760991	1.8×10^{-4}	kgp10612999	7.93×10^{-4}	rs9615894	6.57×10^{-5}	kgp5293872	5.79×10^{-4}
kgp22793223	2.11×10^{-4}	rs12484788	9.59×10^{-4}	kgp7397507	6.74×10^{-5}	rs2008512	6.19×10^{-4}

Tabla 2.7: SNPs con p-valores más pequeños de los modelos aditivos (tras la selección de covariables condicional al SNP) para el efecto adverso falta de apetito a los 3, 6 y 12 meses y para falta de apetito global.

2.3. Limitaciones de la regresión logística en un GWAS

Tras llevar a cabo un GWAS con modelos de regresión logística, teniendo en cuenta 17377 SNPs y considerando las variables respuesta de insomnio y falta de apetito, hemos visto que no se ha encontrado ninguna asociación estadísticamente significativa. Recordemos que los 17377 SNPs con los que estamos trabajando corresponden al cromosoma 22. Podríamos repetir el mismo análisis considerando los SNPs de otros cromosomas con el fin de averiguar si se encuentran SNPs significativos en otras regiones genómicas.

Una limitación que nos encontramos en este contexto es que el hecho de estudiar los tres instantes temporales por separado no nos permite evaluar la progresión a lo largo del tiempo. No obstante, no es posible hacer un único GWAS considerando de forma conjunta los datos de los tres instantes temporales y ajustando modelos de regresión logística. En caso de hacerlo se estaría incumpliendo la hipótesis de independencia de observaciones, que es precisamente una de las hipótesis básicas del modelo de regresión logística (se utiliza en la ecuación (2.1), en el cáclulo de la verosimilitud). Esto es así ya que cabe esperar que las medidas de un mismo individuo a los 3, 6 y 12 meses no sean independientes entre sí, sino que parece razonable suponer que existe una cierta correlación entre ellas. Por otra parte, teniendo en cuenta la variable resumen del efecto adverso estamos perdiendo información que podría ser de mucha utilidad. De hecho, esto se ve claro en la Tabla 2.2, pues los modelos cambian al cambiar el instante temporal.

En el siguiente capítulo veremos que debido a la estructura de dependencia que presentan nuestros datos, que viene determinada por el seguimiento temporal de cada individuo, se hace adecuado recurrir a los modelos de regresión mixtos.

Capítulo 3

Modelos mixtos lineales

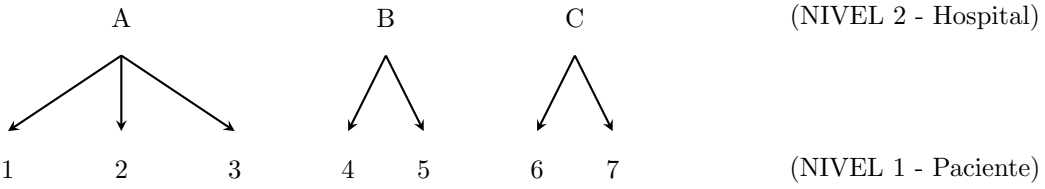
En muchas ocasiones, y muy frecuentemente en el ámbito de la biomedicina, aparecen de forma natural datos agrupados. La principal característica de este tipo de datos es la correlación existente entre las observaciones de cada grupo. Por esta razón los análisis que asumen la independencia entre las observaciones, como el modelo lineal o el modelo lineal generalizado, son inapropiados. En este contexto los modelos mixtos pueden ser útiles para modelar las estructura de correlación existente.

En este capítulo, tras introducir las principales estructuras de dependencia y la diferencia entre efectos fijos y efectos aleatorios, nos centraremos en los modelos mixtos lineales (denotados frecuentemente en la literatura por LMM por sus siglas en inglés, *linear mixed models*), en los que se considera una variable respuesta continua. Comenzaremos presentando el modelo más simple e iremos aumentando la complejidad progresivamente. Una vez familiarizados con los conceptos, introduciremos la formulación del modelo mixto lineal general.

3.1. Estructura de los datos

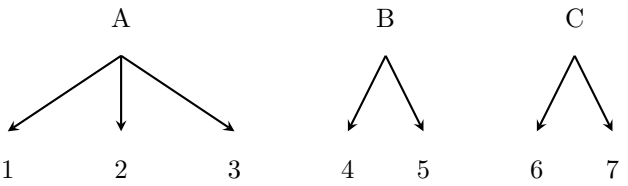
La estructura de los datos es precisamente la que nos va a indicar si debemos utilizar modelos mixtos, y en caso afirmativo nos informará sobre qué tipo de modelo mixto es el adecuado (Durbán, s.f.). A continuación presentaremos las dos estructuras de dependencia más usuales, incluyendo un ejemplo para cada una de ellas, y veremos a cuál se ajustan mejor nuestros datos de TDAH.

- Datos jerárquicos o anidados. La variable respuesta se mide una única vez en cada individuo, que es la unidad básica del análisis. Estos individuos están agrupados o anidados en unidades superiores. Por ejemplo, podemos clasificar alumnos en escuelas y pacientes en hospitales. En cada uno de estos niveles de jerarquía se pueden medir variables. Podría ser interesante, por ejemplo, tener en cuenta si el hospital es público o privado, así como la edad y el sexo del paciente.



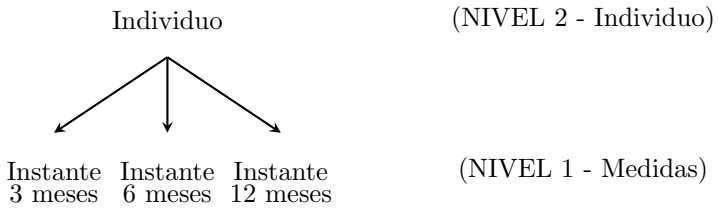
- Medidas repetidas y datos longitudinales. La variable respuesta se mide más de una vez sobre el mismo individuo. Por ejemplo, es común medir los niveles de glucosa de un enfermo antes y

después de haberle inyectado insulina, y también tomar varias medidas a lo largo del tiempo sobre un mismo paciente para saber si está respondiendo de forma satisfactoria a un tratamiento. Parece razonable pensar que las medidas tomadas sobre un mismo individuo no son independientes. Ante este tipo de estructura en los datos los individuos pasan a estar en el nivel superior, mientras que las medidas tomadas sobre cada uno de ellos conforman el nivel inferior.



En estudios de medidas repetidas la variable respuesta se mide más de una vez en cada individuo, no necesariamente a lo largo del tiempo. Por contra, en los análisis longitudinales la variable respuesta se mide en varios instantes temporales en cada individuo, abarcando normalmente un período relativamente largo de tiempo. En algunos casos puede ser complicado determinar si estamos ante medidas repetidas o datos longitudinales. No obstante, desde el punto de vista del análisis de los datos mediante modelos mixtos, se trabaja de la misma forma en ambos casos.

Nótese que en nuestro caso práctico estamos ante un estudio de medidas repetidas, donde cada individuo es medido en tres instantes temporales. Es de esperar que las medidas tomadas sobre un mismo niño estén correlacionadas, sobre todo si han sido registradas en momentos temporales próximos.



Hemos visto que cuando la estructura de los datos es jerárquica o cuando estamos ante un estudio de medidas repetidas o de datos longitudinales podemos considerar varios niveles de información. Además, en cada uno de estos niveles se pueden medir variables explicativas. En nuestro caso, genotipo, edad, sexo, subtipo y antecedentes de tratamiento son variables de segundo nivel, mientras que fármaco, dosis y efectos adversos son variables de primer nivel.

Partiendo de esta estructura en dos niveles, el objetivo es analizar la variabilidad entre grupos (en nuestro caso entre individuos) y dentro de los mismos (para un individuo en concreto). En muchos casos existe la posibilidad de que haya más de dos niveles de información. Por ejemplo, sería posible agrupar medidas en pacientes, pacientes en hospitales y hospitales en regiones.

3.2. Efectos fijos y efectos aleatorios

El término modelo mixto se refiere a la presencia tanto de efectos fijos como de efectos aleatorios en el mismo análisis. Para entender la diferencia entre efectos fijos y efectos aleatorios es importante tener clara cuál es la estructura de dependencia de los datos y cuál es la variable de agrupación. En el

caso de estudios de medidas repetidas o datos longitudinales la variable agrupadora que determina la estructura de dependencia es la variable que identifica al individuo.

Un efecto fijo es una constante desconocida que tratamos de estimar a partir de los datos. Esto es común en modelos lineales y en modelos lineales generalizados. Por ejemplo, en un modelo de regresión lineal simple con una variable explicativa continua el interés se centra en estimar la pendiente y la ordenada en el origen. Si introducimos una variable explicativa categórica tendremos que estimar también el efecto de cada uno de sus niveles. Supongamos que queremos estudiar el nivel de linfocitos en un conjunto de 1000 pacientes diagnosticados de una cierta enfermedad, para la cual se pueden suministrar cuatro fármacos distintos. Cada uno de los pacientes recibe uno de esos cuatro fármacos, y se le mide el nivel de linfocitos a lo largo del tiempo. Si consideramos como variable explicativa el fármaco, nuestro interés se centrará en estudiar cuál es el efecto de cada uno de ellos sobre el nivel de linfocitos. Es importante notar que los cuatro fármacos se han elegido de forma deliberada, y no al azar. Nos interesan esos cuatro en particular, que son los que se utilizan para tratar la enfermedad en cuestión, y no buscamos generalizar los resultados a otros fármacos. Por lo tanto, se considera un efecto fijo asociado a cada uno de los cuatro fármacos. Los efectos fijos corresponden a variables categóricas que tienen ciertos niveles prefijados, los cuales interesaría utilizar de nuevo si repitiésemos el experimento.

Por otra parte, un efecto aleatorio es una variable aleatoria, razón por la cual no tiene sentido estimarlo, sino que lo que se podrá hacer es predecirlo. De hecho, el interés se centra en estimar los parámetros que describen la distribución de tal efecto aleatorio. Retomando el ejemplo del nivel de linfocitos, nótese que tenemos información de 1000 pacientes. Normalmente no estimaremos un efecto fijo para cada uno de ellos, sino que se considerará que ese conjunto de individuos de los cuales tenemos información es simplemente una muestra representativa de todos los individuos diagnosticados de la enfermedad en cuestión. A diferencia del caso de la variable categórica que indica el fármaco suministrado a cada paciente, ahora hay más posibles niveles (más pacientes) y no nos interesa estudiar cada uno de ellos de forma precisa: no estamos particularmente interesados en los individuos del estudio sino en la entera población de los mismos. En el caso del fármaco sí que nos interesa saber si el 1 es mejor que el 2 en cuanto al nivel de linfocitos, pero con respecto a los pacientes no nos importa si el 1 progresa mejor que el 2 o si el 56 progresa peor que el 897. Por lo tanto, se considerará que el efecto individuo es un efecto aleatorio. Esto es, los efectos aletorios están asociados a una variable categórica cuyos niveles se pueden ver como una selección aleatoria de un conjunto general y más grande de todos los posibles niveles.

Relacionando estos conceptos generales con nuestros datos de TDAH, se tiene que los efectos de fármaco son efectos fijos (recordemos que el fármaco que se usa es el metilfenidato y que tiene dos formulaciones posibles: de liberación inmediata y de liberación prolongada). Sin embargo, el efecto de cada niño es aleatorio, pues aunque tenemos información de 199 niños no estamos interesados particularmente en cada uno de ellos sino que los consideramos como una muestra representativa de niños con TDAH, en este caso de España. Nótese que si volviésemos a replicar el estudio los niveles del fármaco no cambiarían, mientras que los niños considerados no tendrían que ser necesariamente los mismos. Por tanto, consideraremos un posible efecto aleatorio vinculado a la estructura de segundo nivel (nivel individuo). Estaremos interesados en ver cómo el efecto aleatorio atribuido al niño explica la variabilidad en la variable dependiente de efecto adverso. En el caso de que estuviésemos interesados en los resultados para esos 199 niños en concreto, sin intención de generalizar los resultados al resto de niños con TDAH, habría que estimar un efecto fijo para cada uno, lo cual supondría la estimación de 199 parámetros. El problema de este modelo, además de su coste computacional, es que no nos daría una estimación de la variabilidad entre niños, la cual puede ser muy informativa. En conclusión, en nuestro estudio consideraremos únicamente un efecto aleatorio asociado al individuo. Para el resto de covariables categóricas (SNP, sexo, antecedente de tratamiento, fármaco y subtipo) consideraremos efectos fijos con la finalidad de estudiar si el efecto del cambio de un nivel a otro es significativo.

Recordemos que en lo que resta de este capítulo nos vamos a centrar en los modelos mixtos lineales, en los cuales la variable respuesta debe ser continua. En nuestro caso práctico las variables respuesta de efecto adverso son todas binarias. Por esa razón, para ilustrar los conceptos que iremos introduciendo nos ayudaremos de los datos `sleepstudy`, disponibles en la librería `lme4` de `R` y procedentes del estudio llevado a cabo en Belenky et al. (2003). Dichos datos fueron registrados con la finalidad de investigar el efecto de la privación de sueño en los tiempos de reacción de camioneros de larga distancia. Se seleccionaron 18 camioneros, a los cuales se les permitió dormir su cantidad habitual de horas de sueño el día 0. Los 9 días siguientes solo se les permitió dormir 3 horas diarias, y se midió el tiempo de reacción a partir de una serie de test llevados a cabo a lo largo de cada día, a cada sujeto. Las variables que podemos encontrar en dicha base de datos son el tiempo de reacción (variable respuesta, medida en milisegundos), el día (indicando el número de días de privación de sueño) y el código de identificación de cada sujeto. Está claro que en este caso la variable agrupadora que da lugar a la estructura de dependencia es la que identifica al individuo. Podríamos pensar que estamos interesados en cada uno de los 18 participantes del estudio en particular, pero quizás sea más razonable pensar que esos 18 individuos simplemente han sido elegidos al azar de la población de camioneros de larga distancia y que no nos interesan de manera individual. Si se volviese a repetir el estudio, seguramente no sería necesario que los participantes fuesen los mismos. De este modo, parece adecuado considerar un efecto aleatorio asociado al individuo.

3.3. ANOVA con efectos aleatorios

Un caso particular de modelo lineal general es aquel donde únicamente se considera una variable explicativa categórica que divide la población en varios grupos (modelo de análisis de la varianza, ANOVA). Se trata de un modelo de regresión con una variable respuesta continua y una única variable explicativa discreta. Se consideran J muestras:

$$\begin{array}{cccccc} Y_{11} & Y_{21} & \dots & Y_{n_1 1} & \text{de una población} & N(\mu_1, \sigma^2) \\ Y_{12} & Y_{22} & \dots & Y_{n_2 2} & \text{de una población} & N(\mu_2, \sigma^2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Y_{1J} & Y_{2J} & \dots & Y_{n_J J} & \text{de una población} & N(\mu_J, \sigma^2) \end{array}$$

Cada una de las J muestras está formada por variables independientes y con la misma distribución. Se trata, por tanto, de J muestras aleatorias simples. Además, se supone que las J muestras son independientes entre sí. Nótese que a las medias se les permite ser distintas, pero las varianzas se suponen todas iguales.

Existen diferentes formas de parametrizar el ANOVA: por desviación respecto de la media global, a través de las medias locales o por desviación respecto de un grupo de referencia. Teniendo en cuenta la primera de ellas, el modelo se puede escribir como sigue:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \qquad j = 1, \dots, J, \quad i = 1, \dots, n_j,$$

siendo μ la media global, α_j la desviación del j-ésimo grupo a la media global y $\epsilon_{ij} \sim N(0, \sigma^2)$ independientes. Con esta parametrización se supone además que la suma de las desviaciones es igual a cero. Si lo expresamos en forma matricial como modelo lineal, quedaría así:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{n_J J} \end{pmatrix}$$

Los J grupos del ANOVA resultan de juntar los individuos que comparten la misma categoría en la variable explicativa discreta. Por ejemplo, si la variable explicativa discreta es el fármaco recibido por el paciente (supongamos que hay fármaco 1, 2, 3 y 4), el interés se centra en estudiar si pasar del fármaco 1 al 2 incrementa el nivel de linfocitos. Es importante destacar que los grupos están definidos de manera definitiva, unívoca, y el interés se centra en estudiarlos de manera precisa. Se trata por lo tanto de efectos fijos.

Centrándonos en el estudio de privación de sueño, recordemos que la variable categórica agrupadora es la que identifica al individuo. Si ajustamos un ANOVA nos encontramos con varios problemas. El primero y más notorio es que estaríamos considerando la independencia de las diez medidas tomadas sobre cada individuo, lo cual no parece muy realista. El segundo es que tendríamos que estimar un efecto para cada uno de los 18 individuos. En este caso no es un número extremadamente alto, pero hay que tener en cuenta que en un estudio de este tipo podríamos considerar incluso cientos de participantes. Además de la carga computacional que eso conlleva, es esencial darse cuenta de que nuestro interés no se centra en estudiar el efecto de pasar de un individuo a otro, luego la estimación de cada uno de esos efectos normalmente carece de utilidad. Veremos a continuación que la consideración de efectos aleatorios será una mejor alternativa.

En el modelo de análisis de la varianza con efectos aleatorios se piensa en una cantidad en principio ilimitada de grupos, como pueden ser los individuos. Ya hemos comentado que normalmente no interesa estudiar específicamente cada uno de los individuos de los que tenemos datos, sino que se consideran como si fuera una muestra aleatoria de todos los participantes potenciales del estudio. Se trata de estudiar qué proporción de variabilidad podemos atribuir a cada nivel de información: recordemos que el nivel 1 de información está formado por las medidas, mientras que el nivel 2 está formado por los individuos. El ANOVA con efectos aleatorios ya no busca estimar una media diferente para cada individuo, como haría el ANOVA con efectos fijos, sino que lo que hace es estimar una media global y predecir un efecto aleatorio para cada individuo. Esta aleatoriedad es la que hará que unos individuos tengan mayor o menor capacidad de reacción.

En la Figura 3.1 (izquierda) podemos ver el diagrama de puntos de los datos de privación de sueño. Se puede apreciar una tendencia creciente en el tiempo de reacción a medida que aumentan los días de privación, así como una variabilidad que parece aumentar a lo largo de los días. El gráfico resulta poco informativo, ya que no se tiene en cuenta a qué individuo pertenece cada dato. Lo ideal sería disponer de la trayectoria de cada individuo. Esto es lo que se muestra en la Figura 3.1 (derecha), que resulta mucho más útil.

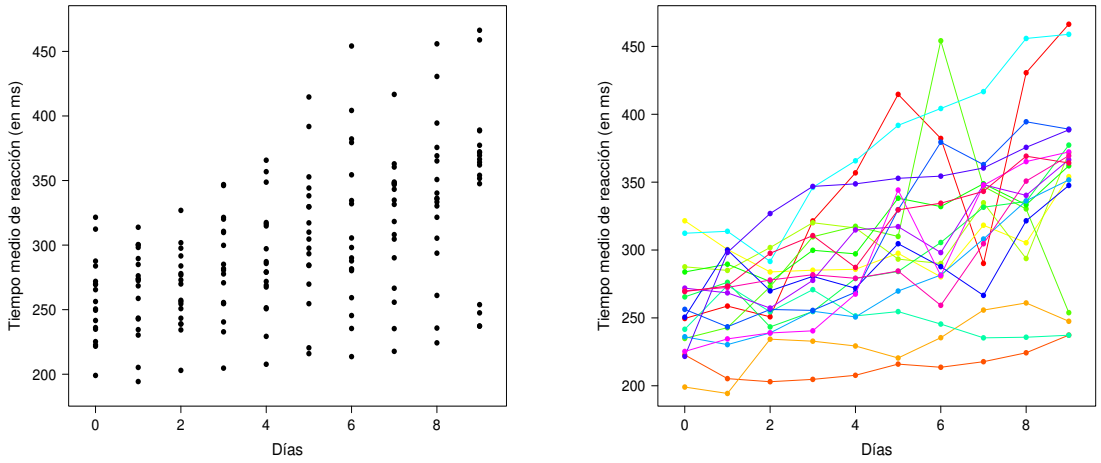


Figura 3.1: Diagrama de puntos de los datos de privación de sueño (izquierda) y trayectorias individuales (derecha).

El modelo de análisis de la varianza con efectos aleatorios (conocido como RANOVA o *random effects* ANOVA) se formula de la siguiente forma:

$$Y_{ij} = \mu + u_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_j, \quad j = 1, \ldots, J,$$

donde Y_{ij} denota en nuestro caso la i-ésima medida del tiempo de reacción del j-ésimo individuo, μ el tiempo medio global de reacción, u_j la desviación del j-ésimo individuo a la media global (esto es, el efecto aleatorio del individuo) y ϵ_{ij} el error. Además, J es el número de individuos (en nuestro caso 18) y se supone que las desviaciones debidas al grupo verifican $u_j \sim N(0, \sigma_u^2)$ variables independientes e indénticamente distribuidas (iid) e independientes de los errores, que cumplen $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ y son independientes entre sí. Está claro que con este modelo no se incluye la tendencia al cambiar el día.

La estimación de los efectos fijos y la predicción de los efectos aleatorios de un modelo mixto (en nuestro caso μ y u_j) se pueden obtener a partir de las ecuaciones de modelos mixtos de Henderson. Por otra parte, la estimación de la matriz de varianzas-covarianzas de Y (de la cual en nuestro caso, como veremos un poco más adelante, podemos extraer los valores de σ_u^2 y σ_ϵ^2) se obtiene por el método de máxima verosimilitud o máxima verosimilitud restringida. Trataremos este tema con más profundidad en la Sección 3.6. Con la ayuda de la función `lmer` de la librería `lme4` de `R` hemos ajustado el modelo, obteniendo una estimación de 298.51 para la media global, esto es, $\hat{\mu} = 298.51$. Esto significa que el tiempo medio global de reacción es de casi 300 milisegundos. Cada individuo tiene una desviación con respecto a dicho valor que viene determinada por su efecto aleatorio asociado. Estos efectos aleatorios tienen media 0 y desviación típica estimada $\hat{\sigma}_u = 35.75$. De esta forma, para el individuo j el modelo ajustado no es más que una recta horizontal que tiene en cuenta la media global y la desviación a ella correspondiente a dicho individuo. Con la ayuda de la función `ranef` de `R`, también perteneciente a la librería `lme4`, podemos obtener la predicción de cada uno de los 18 efectos aleatorios. Así, en la Figura

3.2 podemos ver los modelos ajustados para cada uno de los 18 individuos, representados en distintos colores. En trazo punteado se muestra la horizontal correspondiente a la media global, esto es, la recta $y = 298.51$.

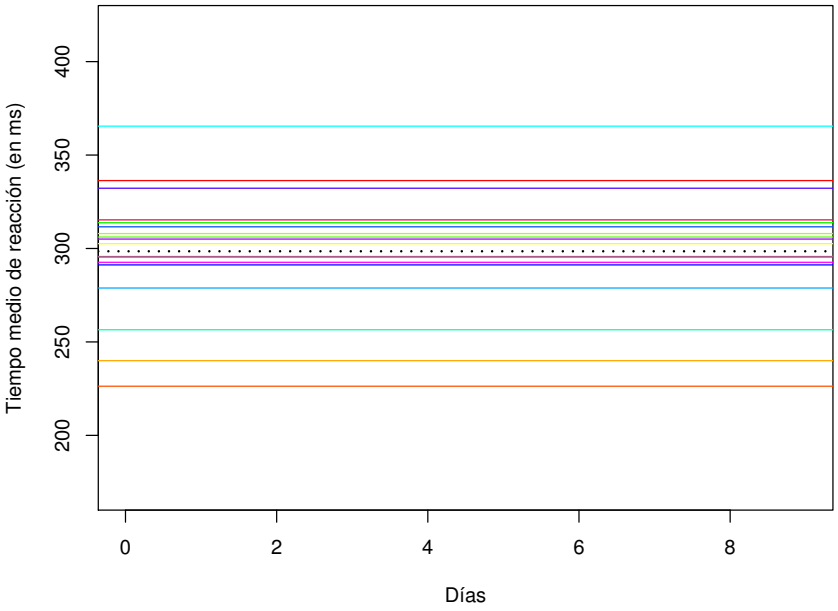


Figura 3.2: Modelo RANOVA para los datos de privación de sueño.

Es importante destacar que en la Figura 3.1 (derecha) podemos ver que claramente se aprecia una tendencia creciente a lo largo del tiempo para casi todos los individuos, de modo que al aumentar el número de días de privación de sueño también aumenta el tiempo medio de reacción. En esta sección hemos ignorado la información de los días de privación de sueño, ya que al ajustar un RANOVA únicamente consideramos la variable respuesta continua y la variable explicativa discreta agrupadora, que en nuestro caso es la variable de identificación de cada individuo. En la siguiente sección introduciremos la variable temporal en el modelo.

En cuanto a la notación del RANOVA, nótese que i representa la medida (nivel 1) y j denota al individuo (nivel 2). De esta forma siempre es posible añadir niveles de agrupación superiores con las letras siguientes, respetando el orden alfabético. Por ejemplo, podemos agrupar medidas (i) en individuos (j) e individuos en hospitales (k), surgiendo así una tercer nivel de información.

El modelo de análisis de la varianza con efectos aleatorios se puede reescribir de la siguiente forma:

$$Y_{ij} = \beta_{0j} + \epsilon_{ij}, \qquad i = 1, \ldots, n_j, \quad j = 1, \ldots, J,$$

donde $\beta_{0j} = \mu + u_j$ es un intercepto aleatorio, diferente para cada grupo, verificando $\beta_{0j} \sim N(\mu, \sigma_u^2)$. En la Figura 3.2 se puede ver que las 18 rectas ajustadas, una para cada individuo, tienen distinto intercepto. Se dice que dicho intercepto es aleatorio, pues su valor viene determinado por el efecto aleatorio asociado al individuo. Además, las rectas son paralelas: tienen la misma pendiente, en nuestro

caso nula porque no hemos considerado ninguna variable explicativa además de la agrupadora.

La varianza total del RANOVA se descompone en dos sumandos: la varianza entre grupos (entre individuos) y la varianza intra-grupo (intra-individuo). Además, se puede ver que la correlación existente entre dos medidas del mismo individuo, pongamos Y_{ij} e $Y_{i'j}$, con $i, i' \in \{1, \ldots, n_j\}$ para cierto $j \in \{1, \ldots, J\}$, es igual a $\frac{\sigma_u^2}{\sigma_u^2+\sigma_\epsilon^2}$:

$$\mathrm{Cor}(Y_{ij}, Y_{i'j}) = \frac{\mathrm{Cov}(Y_{ij}, Y_{i'j})}{\sqrt{\mathrm{Var}(Y_{ij})}\sqrt{\mathrm{Var}(Y_{i'j})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}.$$

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{i'j}) &= \mathrm{Cov}(\mu + u_j + \epsilon_{ij}, \mu + u_j + \epsilon_{i'j}) = \mathrm{Cov}(u_j + \epsilon_{ij}, u_j + \epsilon_{i'j}) \\ &= \mathrm{Cov}(u_j, u_j) + \mathrm{Cov}(u_j, \epsilon_{i'j}) + \mathrm{Cov}(\epsilon_{ij}, u_j) + \mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j}) = \mathrm{Var}(u_j) = \sigma_u^2.\end{aligned}$$

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\mu + u_j + \epsilon_{ij}) = \mathrm{Var}(u_j + \epsilon_{ij}) = \mathrm{Var}(u_j) + \mathrm{Var}(\epsilon_{ij}) = \sigma_u^2 + \sigma_\epsilon^2.$$

$$\mathrm{Var}(Y_{i'j}) = \mathrm{Var}(\mu + u_j + \epsilon_{i'j}) = \mathrm{Var}(u_j + \epsilon_{i'j}) = \mathrm{Var}(u_j) + \mathrm{Var}(\epsilon_{i'j}) = \sigma_u^2 + \sigma_\epsilon^2.$$

Por otra parte, las medidas de dos individuos distintos, pongamos Y_{ij} e $Y_{kj'}$, con $j, j' \in \{1, \ldots, J\}$, $j \neq j'$, $i \in \{1, \ldots, n_j\}$ y $k \in \{1, \ldots, n_{j'}\}$, son incorreladas:

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{kj'}) &= \mathrm{Cov}(\mu + u_j + \epsilon_{ij}, \mu + u_{j'} + \epsilon_{kj'}) = \mathrm{Cov}(u_j + \epsilon_{ij}, u_{j'} + \epsilon_{kj'}) \\ &= \mathrm{Cov}(u_j, u_{j'}) + \mathrm{Cov}(u_j, \epsilon_{kj'}) + \mathrm{Cov}(\epsilon_{ij}, u_{j'}) + \mathrm{Cov}(\epsilon_{ij}, \epsilon_{kj'}) = 0.\end{aligned}$$

Así, la matriz de covarianzas del vector de observaciones Y presenta una estructura diagonal en bloques. Dentro de cada bloque el valor de la diagonal es $\sigma_u^2 + \sigma_\epsilon^2$, mientras que fuera de la diagonal el valor es σ_u^2. En concreto, si consideramos $J = 2$, $n_1 = 3$ y $n_2 = 2$, la matriz de covarianzas de las observaciones tendría la siguiente forma:

$$\mathrm{Cov}(Y, Y) = \begin{pmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \sigma_u^2 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 \\ 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 \end{pmatrix}.$$

Este tipo de estructura en la matriz de covarianzas se conoce como simetría compuesta (*compound symmetry*). En la matriz se observa una estructura por bloques, asociada con la estructura de segundo nivel. Los elementos de la diagonal son los mismos para toda la población, y lo mismo ocurre con la correlación entre dos medidas cualesquiera, dentro de cada individuo. En el caso de los datos de privación de sueño la matriz de covarianzas sería de dimensión 180, formada por 18 bloques de dimensión 10.

Con el fin de determinar qué proporción de la varianza total es atribuible a la variación entre grupos se considera el coeficiente de partición de la varianza (VPC, *variance partition coefficient*):

$$\mathrm{VPC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{\text{varianza entre grupos}}{\text{varianza total}}.$$

Dicho de otra forma, VPC nos infroma sobre la proporción de varianza que queda explicada por el nivel superior de información. En nuestro caso se tiene que $\hat{\sigma}_u = 35.75$ y $\hat{\sigma}_\epsilon = 44.26$. Por lo tanto, el 39.48 % de la varianza total queda explicada por el nivel individuo, por lo que parece que la introducción del efecto aleatorio de individuo en el modelo es acertada.

3.4. Modelo mixto lineal con variables explicativas de primer y segundo nivel

En el RANOVA no se tiene en cuenta ninguna información adicional salvo la estructura de los datos. Esto es, se considera únicamente la variable agrupadora. El modelo RANOVA se puede extender sin más que introducir variables explicativas, que se pueden considerar en los dos niveles de información.

La variable respuesta se seguirá denotando por Y_{ij}, con $j = 1, \ldots, J$ e $i = 1, \ldots, n_j$. Para las covariables se utilizará una notación diferente dependiendo del nivel de información al que correspondan: las de primer nivel se denotarán por X_{ij}, mientras que las de segundo nivel se denotarán por W_j.

3.4.1. Modelo con intercepto aleatorio

Recordemos que el RANOVA se puede escribir como

$$Y_{ij} = \beta_{0j} + \epsilon_{ij}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_j,$$

siendo $\beta_{0j} = \mu + u_j$ un intercepto aleatorio verificando $\beta_{0j} \sim N(\mu, \sigma_u^2)$. Seguiremos considerando una estructura de dependencia de datos longitudinales o medidas repetidas, en donde en el nivel superior de información están los individuos y en el nivel inferior están las medidas tomadas sobre cada uno de ellos en distintos instantes temporales. Además de la variable respuesta, medida en cada uno de dichos instantes, podemos tener información adicional de otras variables. En el caso de los datos de privación de sueño tenemos la información adicional que nos da la variable temporal que indica los días. Se trata de una variable de primer nivel.

En esta situación se puede formular un modelo en cada grupo de medidas $j = 1, \ldots, J$ (para cada individuo) para explicar la variable respuesta Y en función de la explicativa X y teniendo además en cuenta la estructura de los datos:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_j. \tag{3.1}$$

En este modelo el intercepto es aleatorio, como en el caso del RANOVA. A la variable X_{ij} se le asocia un efecto fijo, β_1, que nos indica el impacto que tiene el aumento de un día de privación de sueño sobre tiempo de reacción. Así, la pendiente β_1 es la misma para todos los grupos, por lo que las rectas ajustadas para cada individuo serán paralelas, pero se le permite ser no nula. Nótese que en nuestro modelo realmente X_{ij} solo depende de i (que nos indica el día), y no de j (que nos indica el individuo). Esto es así porque todos los camioneros han sido evaluados en los mismos instantes temporales. No obstante, la formulación X_{ij} nos da mucha flexibilidad ya que permitiría incluir instantes de tiempo diferentes en el seguimiento de los individuos. Es importante destacar que, salvo excepciones, en la Figura 3.1 (derecha) se aprecia un patrón más o menos lineal para cada individuo, con tendencia creciente. Por lo tanto, parece que tiene sentido plantearse la formulación de estos modelos lineales.

Con la ayuda de la función `lmer` de `R` hemos ajustado este modelo a los datos de privación de sueño, obteniendo las siguientes estimaciones: $\hat{\mu} = 251.4051$, $\hat{\beta}_1 = 10.4673$, $\hat{\sigma}_u^2 = 1378.2$ y $\hat{\sigma}_\epsilon^2 = 960.5$. Esto quiere decir que el tiempo medio de reacción en condiciones normales (esto es, durmiendo la cantidad

habitual de horas) es de aproximadamente 250 milisegundos (alrededor de un cuarto de segundo). Por cada día de privación de sueño el tiempo de reacción se incrementa en 10.5 milisegundos. Además, `R` también nos permite predecir con la función `ranef` el efecto aleatorio u_j de cada individuo, y por lo tanto representar los modelos ajustados, que se muestran en la Figura 3.3.

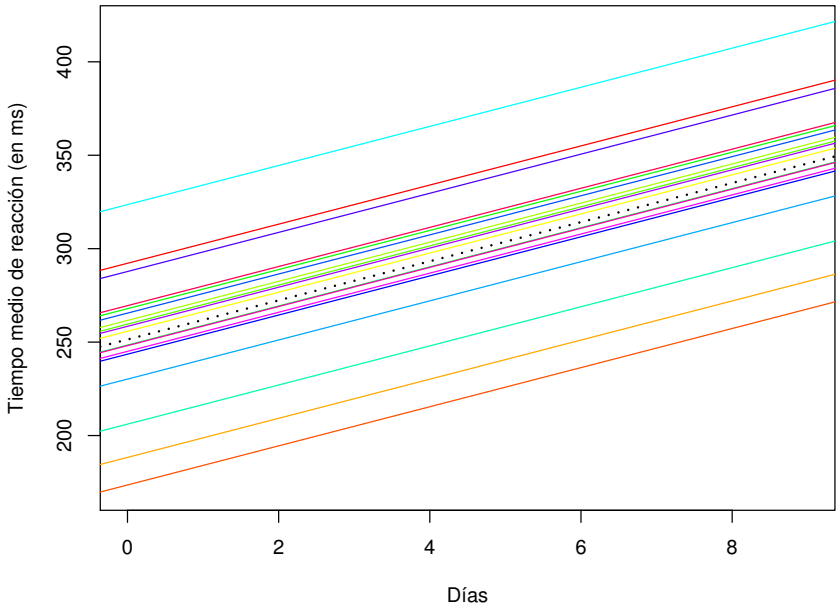


Figura 3.3: Modelo mixto lineal con intercepto aleatorio y pendiente fija para los datos de privación de sueño.

En vista a este último gráfico, y también al obtenido tras el ajuste del modelo RANOVA, parece evidente la utilidad de la consideración de un intercepto aleatorio asociado al individuo, ya que de esta forma se permite que no todos los camioneros tengan el mismo tiempo de reacción cuando duermen su cantidad habitual de horas diarias. También parece adecuado el hecho de introducir una pendiente no nula. Además de esto, si nos fijamos de nuevo en la Figura 3.1 (derecha) podemos ver que quizás podría ser conveniente considerar una pendiente aleatoria asociada a cada individuo, permitiendo así que no todos evolucionen de la misma manera.

3.4.2. Modelo con intercepto y pendiente aleatoria

El modelo lineal mixto con intercepto y pendiente aleatoria se formula del siguiente modo:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_j.$$

En este modelo tanto el intercepto como la pendiente son fijos para cada grupo, pero varían entre grupos. Esto es, para cada uno de los individuos se ajusta una recta con un cierto intercepto y pendiente. Se considera que (β_{0j}, β_{1j}) son variables aleatorias independientes de los errores, con distribución

normal bivariante

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma_{00}\\ \gamma_{10}\end{pmatrix}, \Sigma_u = \begin{pmatrix}\sigma^2_{u0} & \sigma_{u01}\\ \sigma_{u01} & \sigma^2_{u1}\end{pmatrix}\right).$$

Los parámetros a estimar son el intercepto medio γ_{00}, la pendiente media γ_{10}, la varianza del intercepto σ^2_{u0}, la varianza de la pendiente σ^2_{u1}, la covarianza entre intercepto y pendiente σ_{u01} y la varianza de primer nivel σ^2_ϵ. El intercepto del grupo j se puede escribir como $\beta_{0j} = \gamma_{00} + u_{0j}$, con $u_{0j} \sim N(0, \sigma^2_{u0})$, y la pendiente se puede escribir como $\beta_{1j} = \gamma_{10} + u_{1j}$, con $u_{1j} \sim N(0, \sigma^2_{u1})$. De esta forma, el modelo

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_j$$

se puede reescribir como

$$Y_{ij} = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j})X_{ij} + \epsilon_{ij} = \underbrace{\gamma_{00} + \gamma_{10}X_{ij}}_{\text{parte fija}} + \underbrace{u_{1j}X_{ij} + u_{0j} + \epsilon_{ij}}_{\text{parte aleatoria}}.$$

Los efectos aleatorios pueden verse como desviaciones de los parámetros (pendiente e intercepto) en el grupo j-ésimo con respecto al valor medio en la población, es decir, con respecto al intercepto medio y la pendiente media globales. El término $u_{1j}X_{ij}$ puede interpretarse como una interacción entre el grupo y la variable explicativa (interacción entre dos niveles).

En la Figura 3.4 se muestran las 18 rectas ajustadas correspondientes. Las estimaciones obtenidas fueron las siguientes: $\hat{\gamma}_{00} = 251.405$, $\hat{\gamma}_{10} = 10.467$, $\hat{\sigma}^2_{u0} = 611.90$, $\hat{\sigma}^2_{u1} = 35.08$ y $\hat{\sigma}^2_\epsilon = 654.94$. Además, la estimación de la correlación entre intercepto y pendiente es $\hat{\rho}_{u10} = 0.07$. Nótese que las estimaciones del intercepto medio y de la pendiente media se asemejan bastante a las estimaciones del intercepto medio y de la pendiente para el modelo (3.1). No obstante, el modelo mejora con la inclusión de la pendiente aleatoria, ya que se obtiene un AIC de 1755.63 frente a un AIC de 1794.47 para el modelo (3.1), que solo incluye intercepto aleatorio.

Partiendo del modelo con intercepto y pendiente aleatorios podríamos introducir además información sobre alguna característica de segundo nivel, asociada al individuo, como puede ser su edad. Sea W_j la variable de segundo nivel, el modelo resultante al incluirla es el siguiente:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij},$$

con $\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$, $u_{0j} \sim N(0, \sigma^2_{u0})$, y $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$, $u_{1j} \sim N(0, \sigma^2_{u1})$. El modelo se puede expresar como sigue:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij}}_{\text{parte fija}} + \underbrace{u_{0j} + u_{1j}X_{ij} + \epsilon_{ij}}_{\text{parte aleatoria}}.$$

La parte aleatoria de este modelo coincide con la parte aleatoria del modelo que incluía solo la covarible de primer nivel. Se siguen imponiendo las hipótesis de normalidad e independencia sobre los errores, así como las consideraciones ya comentadas sobre los efectos aleatorios. El término W_jX_{ij} puede verse como una interacción entre variables de distintos niveles, siendo γ_{11} el coeficiente de interacción.

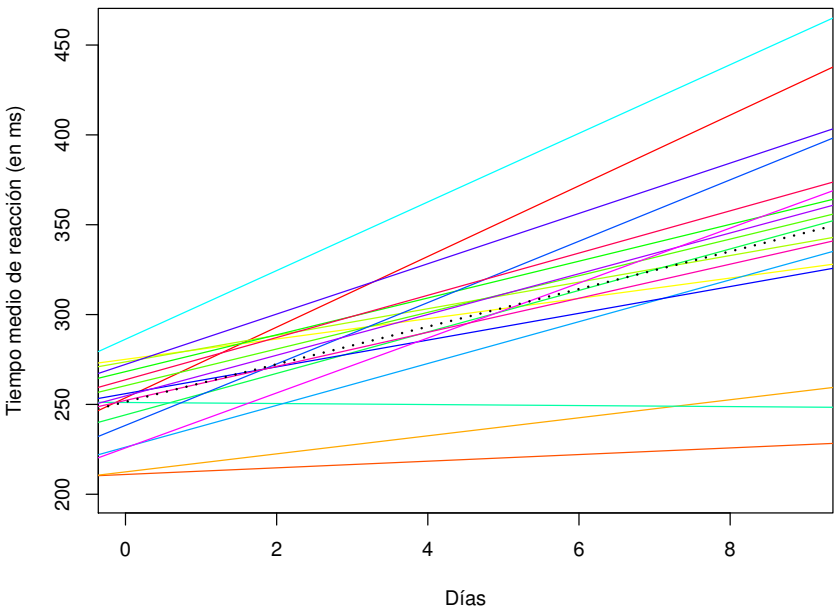


Figura 3.4: Modelo mixto lineal con intercepto y pendiente aleatorios para los datos de privación de sueño.

3.5. Formulación del modelo mixto lineal general

La mejor forma de entender un modelo mixto lineal es recordar el modelo de regresión lineal general. Recordemos que este último se puede expresar como

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

De forma más abreviada,

$$y = X\beta + \epsilon,$$

donde y es el vector de respuestas de tamaño n, X la matriz de diseño de dimensión $n \times p$, β el vector de coeficientes de regresión desconocidos y ϵ el vector (no observable) de errores verificando $\epsilon \sim N_n(0, \sigma^2 I_n)$, siendo I_n la matriz identidad de dimensión n. En este modelo los p coeficientes de regresión se consideran fijos. Sin embargo, ya hemos comentado a lo largo de este capítulo que en algunas ocasiones tiene sentido considerar efectos aleatorios.

La expresión matricial del modelo mixto lineal general es la siguiente:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} z_{11} & \dots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{n1} & \dots & z_{nq} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

De forma más abreviada,

$$y = X\beta + Zu + \epsilon.$$

Nótese que la diferencia entre el modelo lineal general y el modelo mixto lineal general se encuentra en la adición del sumando Zu, el cual corresponde precisamente a la inclusión de los efectos aleatorios en el modelo. La matriz Z, de dimensión $n \times q$, es la matriz de diseño (conocida) asociada a los q efectos aleatorios del vector u. Se asume que $u \sim N_q(0, G)$, $\epsilon \sim N_n(0, R)$ y

$$\text{Var} \begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}.$$

Quizás una de las mayores ventajas que ofrecen los modelos mixtos frente a los modelos lineales generales es la flexibilidad que le permiten a la matriz de varianzas-covarianzas del vector de errores y del vector de efectos aleatorios, y en consecuencia a la matriz de varianzas-covarianzas de la variable respuesta. Una de las hipótesis del modelo lineal general es que el vector de errores verifica $\epsilon \sim N_n(0, \sigma^2 I_n)$, lo cual equivale a que $y \sim N_n(0, \sigma^2 I_n)$. Por lo tanto, para el vector de observaciones se consideran varianzas homogéneas y correlaciones nulas. En presencia de datos en los que existe una estructura de dependencia ya no tiene sentido considerar esta hipótesis, por lo que en principio se debe permitir que $\text{Var}(y)$ puede ser cualquier tipo de matriz. Los modelos mixtos son más flexibles en este sentido, pues permiten varianzas heterogéneas y correlaciones no nulas entre las observaciones. Por ejemplo, en el caso del RANOVA hemos visto que la matriz de varianzas-covarianzas del vector de observaciones tiene una determinada estructura que se conoce como simetría compuesta.

A continuación veremos que los modelos que hemos introducido a lo largo de este tema se pueden reescribir en la notación de modelo mixto lineal general.

RANOVA: $Y_{ij} = \mu + u_j + \epsilon_{ij}$.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_J \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{n_J J} \end{pmatrix}$$

Modelo mixto lineal con intercepto aleatorio y pendiente fija, considerando una variable de primer nivel: $Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \epsilon_{ij}$.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix} = \begin{pmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{n_1 1} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{n_2 2} \\ \vdots & \vdots \\ 1 & X_{1J} \\ \vdots & \vdots \\ 1 & X_{n_J J} \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_J \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{n_J J} \end{pmatrix}$$

Modelo mixto lineal con intercepto aleatorio y pendiente aleatoria, considerando una variable de primer nivel: $Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + u_{1j} X_{ij} + u_{0j} + \epsilon_{ij}$.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix} = \begin{pmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{n_1 1} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{n_2 2} \\ \vdots & \vdots \\ 1 & X_{1J} \\ \vdots & \vdots \\ 1 & X_{n_J J} \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix} + \begin{pmatrix} 1 & X_{11} & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & X_{n_1 1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & X_{12} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 1 & X_{n_2 2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & X_{1J} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & X_{n_J J} \end{pmatrix} \begin{pmatrix} u_{01} \\ u_{11} \\ \vdots \\ u_{0J} \\ u_{1J} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{n_J J} \end{pmatrix}$$

Modelo mixto lineal con intercepto aleatorio y pendiente aleatoria, considerando una variable de primer nivel y una variable de segundo nivel: $Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij} + u_{1j}X_{ij} + u_{0j} + \epsilon_{ij}$.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & W_1 & W_1X_{11} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n_1 1} & W_1 & W_1X_{n_1 1} \\ 1 & X_{12} & W_2 & W_2X_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n_2 2} & W_2 & W_2X_{n_2 2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1J} & W_J & W_JX_{1J} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n_J J} & W_J & W_JX_{n_J J} \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{11} \end{pmatrix} + \begin{pmatrix} 1 & X_{11} & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & X_{n_1 1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & X_{12} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 1 & X_{n_2 2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & X_{1J} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & X_{n_J J} \end{pmatrix} \begin{pmatrix} u_{01} \\ u_{11} \\ \vdots \\ u_{0J} \\ u_{1J} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{n_J J} \end{pmatrix}$$

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

Si separamos la parte fija de la aleatoria, el modelo $Y = X\beta + Zu + \epsilon$ se puede escribir como $Y = X\beta + \epsilon^*$, siendo $\epsilon^* = Zu + \epsilon$. Teniendo en cuenta que $u \sim N_J(0, G), \epsilon \sim N_n(0, R)$ y

$$\text{Var}\begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix},$$

la estructura de covarianzas asociada al modelo viene dada por:

$$\text{Var}(Y) = \text{Cov}(Y, Y) = \text{Cov}(\epsilon^*, \epsilon^*) = \text{Cov}(Zu + \epsilon, Zu + \epsilon) = Z\text{Cov}(u, u)Z' + \text{Cov}(\epsilon, \epsilon) = ZGZ' + R = V.$$

El interés se centra en estimar el vector de parámetros, β, y las componentes de la varianza, G y R. Además, también interesará predecir los efectos aleatorios (u).

Para estimar β, si conociésemos las matrices de varianzas-covarianzas G y R, y por tanto V, podríamos utilizar el método de mínimos cuadrados generalizados. Minimizando la siguiente función

$$G = (Y - X\beta)'V^{-1}(Y - X\beta)$$

se obtiene

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y,$$

que además coincide con el estimador de máxima verosimilitud. El problema es que es muy poco realista suponer que disponemos de G y R, y en tal caso no podemos hallar la matriz V y por tanto necesitaremos estimarla.

Tal y como se indicó, en el caso de los efectos aleatorios u hablamos de predicción y no de estimación, ya que se trata de variables aleatorias en lugar de parámetros. El método de las ecuaciones de modelos

mixtos de Henderson (Henderson et al., 1959) permite obtener el mejor estimador lineal insesgado de β (BLUE, *best linear unbiased estimator*) y el mejor predictor lineal insesgado de u (BLUP, *best linear unbiased predictor*). Dichas ecuaciones se obtienen a partir de la densidad conjunta de y y u:

$$f(y,u) = f(y|u)f(u).$$

Nótese que estamos abusando de la notación, ya que f denota tanto la densidad conjunta, como la condicionada y la marginal en u. Teniendo en cuenta que $y|u \sim N(X\beta + Zu, R)$ y $u \sim N(0, G)$, el logaritmo de la verosimilitud para la densidad conjunta se puede expresar como

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

Derivando con respecto a β y a u se obtiene lo siguiente:

$$\begin{cases} \frac{\partial l}{\partial \beta} = X'R^{-1}(y - X\beta - Zu) \\ \\ \frac{\partial l}{\partial u} = Z'R^{-1}(y - X\beta - Zu) - G^{-1}u \end{cases}$$

Nótese que estamos abusando de la notación, ya que β es un vector. Lo que realmente se hace es la derivada parcial con respecto a cada una de las componentes de β.

Igualando a cero las expresiones anteriores se obtienen las ecuaciones de Henderson de los modelos mixtos:

$$\begin{cases} X'R^{-1}y = X'R^{-1}X\beta + uX'R^{-1}Z \\ Z'R^{-1}y = Z'R^{-1}X\beta + u(Z'R^{-1}Z + G^{-1}) \end{cases}$$

Las soluciones a estas ecuaciones son:

$$\begin{cases} \hat{\beta} = (X'V^{-1}X)^{-1}XV^{-1}y \\ \hat{u} = GZ'V^{-1}(y - X\hat{\beta}) \end{cases}$$

donde $V = ZGZ' + R$.

El problema es que no disponemos ni de G ni de R, y por lo tanto tampoco de V. Para estimar V podemos usar máxima verosimilitud o máxima verosimilitud restringida (Searle et al., 1992).

Con respecto al método de máxima verosimilitud (ML, *maximum likelihood*), hay que tener en cuenta que $y \sim N(X\beta, V)$. Se tiene que el logaritmo de la verosimilitud es

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

Dado que disponemos de un estimador para β podemos construir el perfil de verosimilitud para V:

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

No existe una expresión cerrada al maximizar esta función, y se hace de forma numérica. Se podría obtener también el estimador de V por máxima verosimilitud restringida (REML, *restricted maximum likelihood*). La principal ventaja del REML sobre ML es que REML tiene en cuenta los grados de libertad utilizados para estimar los efectos fijos del modelo. Si el tamaño de la muestra con la que trabajamos es pequeño REML dará mejores estimaciones que ML, mientras que si es grande apenas habrá

diferencias (Durbán, s.f.). En la función de `R` que hemos utilizado a lo largo de este capítulo para ajustar los modelos, que es la función `lmer` de la librería `lme4`, tenemos la opción de seleccionar cualquiera de estos dos métodos. Por defecto considera REML, que es el método que hemos usado en nuestros ajustes.

Además de la flexibilidad que se le permite a la matriz de varianzas-covarianzas, otra de las ventajas de los modelos mixtos es que podemos hacer uso de todos los datos de los que disponemos: los individuos no tienen que ser necesariamente observados en los mismos tiempos ni el mismo número de veces. Por ejemplo, tal y como ya hemos comentado, la formulación de la ecuación (3.1) permite incluir instantes de tiempo diferentes en el seguimiento de los individuos. Supongamos que en el caso del estudio de privación de sueño solamente disponemos de información completa para 13 de los 18 camioneros. Si hacemos un ANOVA, entonces el análisis se basará solo en esos 13 casos completos. Además de una pérdida de poder estadístico, esta pérdida de información puede dar lugar a otros problemas. Supongamos que los camioneros a los que les afecta mucho la privación de sueño, y que por lo tanto adquieren un tiempo de reacción muy alto que puede suponer un peligro vial, son excluidos del estudio, de modo que se les pierde el seguimiento. Para ellos únicamente estarían registradas las primeras medidas, las correspondientes a los primeros días. De este modo habremos eliminado los datos de los camioneros que responden peor a la privación de sueño, y solamente tendremos datos completos de los camioneros que responden mejor. Esto dará como resultado estimaciones sesgadas.

Capítulo 4

Modelos mixtos lineales generalizados

Los modelos lineales generalizados (GLMs, *generalized linear models*) extienden los modelos de regresión ordinarios al caso en que la variable respuesta no sigue una distribución normal. Un caso particular de este tipo de modelos es el modelo de regresión logística, del cual hemos hablado en el Capítulo 2, donde la variable respuesta es binaria. En el contexto de modelos mixtos también se pueden generalizar los modelos estudiados en el Capítulo 3 al caso de respuesta discreta (en particular, nos interesará cuando es binaria), surgiendo así los modelos mixtos lineales generalizados (GLMMs, *generalized linear mixed models*). Comenzaremos planteando el modelo más simple, que es el modelo logístico con efectos aleatorios. Posteriormente introduciremos modelos más complejos mediante la inclusión de covariables, tanto de primer como de segundo nivel.

4.1. Modelo logístico con efectos aleatorios

De la misma forma que hemos hecho en el capítulo anterior, nos centraremos en el caso de medidas repetidas o datos longitudinales, considerando que la variable respuesta binaria Y se observa varias veces en cada individuo. Esto es precisamente lo que ocurre en nuestros datos de TDAH, ya que la presencia o ausencia de cada efecto adverso se mide en cada uno de los 199 individuos en tres instantes temporales. A modo de ejemplo consideraremos el efecto adverso de insomnio. Denotaremos tales observaciones por Y_{ij}, con $j = 1, \ldots, 199$ e $i = 1, 2, 3$. El subíndice i representa el instante temporal de la medida, mientras que j denota al individuo.

Recordemos que en una regresión logística con efectos fijos el modelo lineal se establece sobre la log-Odds. Lo mismo ocurre en el caso de la regresión logística con efectos aleatorios, que se formula de la siguiente forma:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \eta_{ij},$$

donde π_{ij} denota la probabilidad de que Y_{ij} tome el valor 1 condicionada a los efectos aleatorios del grupo y a otras posibles variables explicativas y η_{ij} es el predictor lineal. Para comenzar consideraremos el modelo más sencillo posible, esto es, el que incluye únicamente un posible efecto aleatorio vinculado a la estructura de segundo nivel. Posteriormente podremos formular modelos más complejos mediante la inclusión de covariables de primer o segundo nivel, de forma análoga a lo hecho en el caso de los modelos mixtos lineales.

El modelo más sencillo posible es el que tiene como predictor lineal $\eta_{ij} = \beta_0 + u_{0j}$:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \beta_0 + u_{0j}, \tag{4.1}$$

donde β_0 es un intercepto global y u_{0j} es el efecto aleatorio de individuo, verificando $u_{0j} \sim N(0, \sigma_u^2)$ independientes. Se tiene que $\pi_{ij} = E(Y_{ij}|u_{0j}) = P(Y_{ij} = 1|u_{0j})$ denota la probabilidad de que Y_{ij} tome el valor 1 condicionada al efecto aleatorio de individuo, $Y_{ij}|u_{0j} \sim Ber(\pi_{ij})$. Así,

$$\pi_{ij} = \frac{\exp(\beta_0 + u_{0j})}{1 + \exp(\beta_0 + u_{0j})}.$$

Nótese que en este caso π_{ij} no depende del índice i, por lo que podríamos denotarlo simplemente por π_j. A pesar de ello, mantendremos la notación general que incluye los dos índices.

En este modelo el predictor lineal asociado a cada individuo es una recta con pendiente nula e intercepto $\beta_0 + u_{0j}$, que puede ser mayor o menor que el intercepto global dependiendo de si u_{0j} es positivo o negativo.

En los modelos mixtos lineales generalizados la estimación de los efectos fijos y la predicción de los efectos aleatorios son tareas tediosas y requieren en general de la ayuda de métodos de simulación numérica. Trataremos este tema con más profundidad en la Sección 4.3. Para ajustar el modelo en `R` hemos utilizado la función `glmer` de la librería `lme4`, obteniendo un AIC de 424.7 y las siguientes estimaciones: $\hat{\beta}_0 = -3.1016$ y $\hat{\sigma}_u^2 = 6.514$. Al igual que en el modelo logístico sin efectos aleatorios, el valor de β_0 admite una interpretación en términos de la Odds. En concreto, β_0 puede interpretarse como la log-Odds de que la variable respuesta tome el valor 1 cuando $u_{0j} = 0$, esto es, para un individuo medio. Así, la estimación de la log-Odds de presentar efecto adverso de insomnio en un individuo medio es $\hat{\beta}_0 = -3.1016$. Por tanto, la Odds es $\exp(-3.1016) = 0.045$ y la probabilidad correspondiente de presencia de insomnio es 0.043.

Nótese que bajo el modelo propuesto (4.1) fijado el índice j se tiene que $\pi_{1j} = \pi_{2j} = \pi_{3j}$. En la Figura 4.1 se muestra el histograma y el diagrama de caja de las probabilidades π_{ij}, pues `R` nos permite obtener dichos valores tras la estimación de β_0 y la predicción de los 199 efectos aleatorios u_{0j}. Podemos ver que la probabilidad de insomnio condicionada al efecto aleatorio de individuo es en general muy baja: de los 199 individuos, 131 tienen una probabilidad asociada menor que 0.1. Por otra parte, solamente hay 8 individuos para los cuales la probabilidad asociada es mayor que 0.75.

Bajo el modelo que estamos suponiendo, se puede deducir fácilmente la función de distribución de π_{ij}:

$$F(x) = P(\pi_{ij} \leq x) = P(\text{logit}(\pi_{ij}) \leq \text{logit}(x)) = P\Big(\beta_0 + u_{0j} \leq \log\Big(\frac{x}{1-x}\Big)\Big) = \Phi_{\beta_0,\sigma_u^2}\Big(\log\Big(\frac{x}{1-x}\Big)\Big),$$

donde $x \in [0,1]$ y $\Phi_{\beta_0,\sigma_u^2}$ denota la función de distribución de la normal con media β_0 y varianza σ_u^2. De este modo, la función de densidad es la siguiente:

$$f(x) = F'(x) = \phi_{\beta_0,\sigma_u^2}\Big(\log\Big(\frac{x}{1-x}\Big)\Big)\frac{1}{x(1-x)},$$

donde $\phi_{\beta_0,\sigma_u^2}$ denota la función de densidad de la normal con media β_0 y varianza σ_u^2. Teniendo en cuenta las estimaciones que hemos obtenido para β_0 y para σ_u^2, esto permite estimar fácilmente los cuantiles de la distribución teórica que se supone que siguen los π_{ij} bajo el modelo considerado. Así, podemos ver que se espera que de los 199 individuos alrededor de 127 tengan una probabilidad asociada menor que 0.1 y alrededor de 10 tengan una probabilidad asociada mayor que 0.75. Estas cifras se

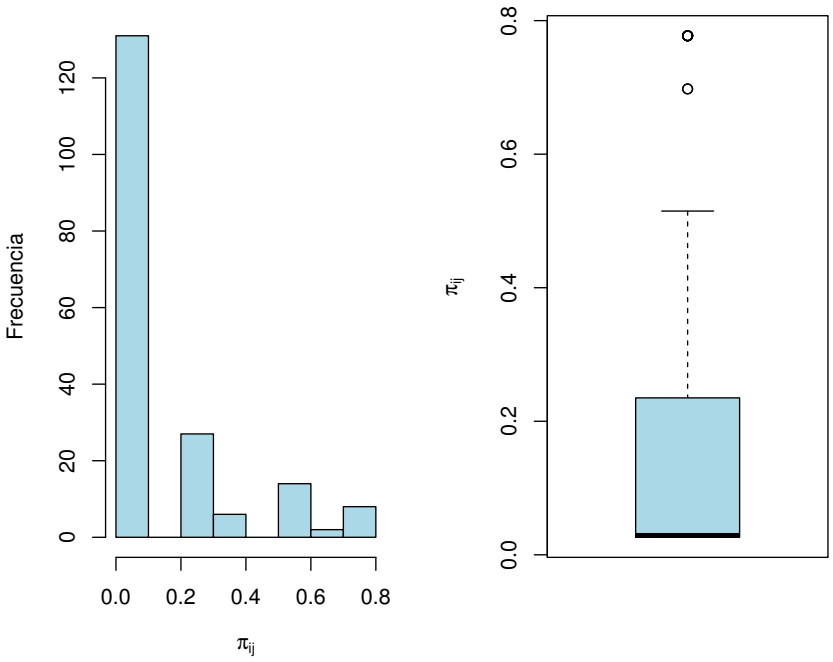


Figura 4.1: Histograma y diagrama de caja de la probabilidad de insomnio condicionada al efecto aleatorio de individuo.

asemejan bastante a las obtenidas tras predecir las probabilidades de cada uno de los 199 individuos de nuestra muestra.

Una vez ajustado este modelo nos podemos preguntar si realmente existe un efecto individuo. Para dar respuesta a esto podríamos aplicar un test de razón de verosimilitudes, contrastando la hipótesis nula de un modelo sin efectos aleatorios frente al modelo ajustado. Lo que obtenemos si hacemos esto es que se rechaza la hipótesis nula a favor de la alternativa, por lo que parece que el efecto niño es importante.

4.2. Introducción de variables explicativas de primer y segundo nivel

Incluiremos ahora, además del efecto aleatorio de individuo, las covariables de primer nivel (fármaco y dosis). El modelo de interés, con intercepto aleatorio y pendiente fija asociada a cada covariable, es el siguiente:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + u_{0j},$$

donde $\pi_{ij} = P(Y_{ij} = 1|X_{ij}, Z_{ij}, u_{0j})$, β_0 es un intercepto global, β_1 es la pendiente asociada a la variable X (fármaco), β_2 es la pendiente asociada a la variable Z (dosis) y u_{0j} es el efecto aleatorio de

individuo (se asume $u_{0j} \sim N(0, \sigma_u^2)$ independientes).

La variable fármaco es categórica (solo hay dos niveles: metilfenidato de liberación inmediata y metilfenidato de liberación prolongada), mientras que la variable dosis es numérica. La variable X_{ij} tomará el valor 0 si el individuo j recibe metilfenidato de liberación inmediata en el momento i, y tomará el valor 1 si el individuo j recibe metilfenidato de liberación prolongada en el momento i.

El modelo ajustado tiene un AIC = 410.3, que es menor que el AIC del modelo que no incluía covariables. Sin embargo, ni el intercepto ni ninguna de las dos pendientes resultan significativamente distintos de cero. Se obtuvieron las siguientes estimaciones: $\hat{\beta}_0 = -1.8$, $\hat{\beta}_1 = -1.36$, $\hat{\beta}_2 = 0.002$ y $\hat{\sigma}_u^2 = 6.136$. El valor de β_0 se puede interpretar como la log-Odds de presentar efecto adverso de insomnio en un individuo con $u_{0j} = 0$ que recibe metilfenidato de liberación inmediata con una dosis de 0 miligramos diarios (esto es, que no recibe fármaco). Como en muchos casos, el intercepto carece de sentido práctico. En nuestro caso la Odds toma un valor de $\exp(-1.8) = 0.17$. Los valores de β_1 y β_2 se pueden interpretar en términos de OR. En concreto, β_1 mide el efecto del paso de metilfenidato de liberación inmediata a metilfenidato de liberación prolongada sobre la log-Odds de $Y = 1$, ajustando por el efecto de grupo (es decir, manteniendo constante el efecto aleatorio) y manteniendo la dosis constante. La OR relativa al fármaco es $\exp(-1.36) = 0.25$, luego parece que el hecho de recibir metilfenidato de liberación prolongada reduce el riesgo se sufrir insomnio. De forma análoga, β_2 mide el efecto del cambio de una unidad en la dosis sobre la log-Odds de $Y = 1$, ajustando por el efecto de grupo y manteniendo constante el fármaco recibido (esto es, sin cambio de fármaco). La OR relativa a la dosis es $\exp(0.002) = 1.002$.

Es posible obtener la probabilidad π_{ij} del siguiente modo:

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + u_{0j})}{1 + \exp(\beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + u_{0j})}.$$

En `R`, además de estimar los efectos fijos, también podemos predecir los efectos aleatorios con la ayuda de la función `ranef`. De ese modo podemos obtener una predicción de las probabilidades π_{ij}. El individuo 1 recibió metilfenidato de liberación prolongada con una dosis de 36 miligramos diarios en los tres instantes temporales, luego $\pi_{11} = \pi_{21} = \pi_{31}$. La probabilidad predicha es $\hat{\pi}_{i1} = 0.027$, con $i = 1, 2, 3$. El individuo 10 recibió siempre metilfenidato de liberación inmediata; en cuanto a la dosis, a los 3 y 6 meses recibió 30 miligramos diarios, mientras que a los 12 meses la dosis bajó a 10. Se tiene que $\hat{\pi}_{1,10} = \hat{\pi}_{2,10} = 0.823$ y $\hat{\pi}_{3,10} = 0.818$. El individuo 83 tuvo cambios tanto de fármaco como de dosis. A los 3 meses recibió metilfenidato de liberación prolongada con una dosis de 54 miligramos diarios. A los 6 y 12 meses el fármaco pasó a ser metilfenidato de liberación inmediata, con una dosis de 40. Se obtuvieron las siguientes probabilidades predichas: $\hat{\pi}_{1,83} = 0.028$ y $\hat{\pi}_{2,83} = \hat{\pi}_{3,83} = 0.097$.

Al igual que en los modelos mixtos lineales con respuesta continua, es posible ajustar modelos mixtos logísticos donde se consideren variables explicativas de segundo nivel. Por ejemplo, podríamos construir un modelo que incluyese además del fármaco y la dosis la edad del niño. En este caso, considerando de nuevo intercepto aleatorio y pendientes fijas, el modelo de interés es el siguiente:

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{01} W_j + \beta_1 X_{ij} + \beta_2 Z_{ij} + u_{0j},$$

donde $\pi_{ij} = P(Y_{ij} = 1 | W_j, X_{ij}, Z_{ij}, u_{0j})$, γ_{00} es un intercepto global, γ_{01} es el coeficiente asociado a la variable W (edad), β_1 es la pendiente asociada a la variable X (fármaco), β_2 es la pendiente asociada a la variable Z (dosis) y u_{0j} es el efecto aleatorio de individuo (se asume $u_{0j} \sim N(0, \sigma_u^2)$ independientes). Para el j-ésimo individuo el intercepto viene dado por $\gamma_{00} + \gamma_{01} W_j + u_{0j}$.

Las estimaciones obtenidas son las siguientes: $\hat{\gamma}_{00} = -2.3$, $\hat{\gamma}_{01} = 0.07$, $\hat{\beta}_1 = -1.29$, $\hat{\beta}_2 = -0.005$ y $\hat{\sigma}_u^2 = 5.933$. No obstante, ninguno de los coeficientes asociados a los efectos fijos resulta significativa-

mente distinto de cero. El AIC del modelo es 411.9, algo mayor que el del modelo que no incluía la covariable edad.

4.3. Estimación de efectos fijos y predicción de efectos aleatorios

Los modelos mixtos lineales generalizados se pueden ver como una extensión o generalización de los modelos mixtos lineales. En el Capítulo 3 hemos visto que en el caso de modelos mixtos lineales con respuesta continua los procedimientos de estimación se basan fundamentalmente en técnicas de máxima verosimilitud (ML y REML). Su extensión al contexto de modelos con respuesta binaria no es inmediata.

Por otra parte, los modelos mixtos lineales generalizados se pueden ver también como una generalización de los modelos lineales generalizados, permitiendo la incorporación de efectos aleatorios. En el Capítulo 2 hemos visto que en los modelos de regresión logística, que son un caso particular de modelos lineales generalizados, la estimación de los parámetros es tediosa. Puesto que las ecuaciones de verosimilitud no tienen en general solución explícita se requiere de métodos numéricos para el cálculo de las estimaciones. Esta dificultad la heredarán los modelos mixtos lineales generalizados. Además, la incorporación de efectos aleatorios en dichos modelos complicará todavía más el tema de la estimación.

Consideremos el modelo de la ecuación (4.1). Para construir la función de verosimilitud, recordemos que $f(y,u) = f(y|u)f(u)$ y que suponemos que los efectos aleatorios u_{0j} siguen una distribución normal con media nula y varianza σ_u^2. Así, la función de verosimilitud se puede escribir en 3 pasos:

1. Verosimilitud condicional para el grupo j:

$$\mathcal{L}_j(\beta_0|u_{0j}) = \prod_{i=1}^{n_j} \pi_{ij}^{Y_{ij}}(1-\pi_{ij})^{1-Y_{ij}}.$$

   Se supone que las observaciones dentro de un grupo son independientes, dados los efectos aleatorios.

2. Verosimilitud marginal en el grupo j:

$$\mathcal{L}_j(\beta_0, \sigma_u^2) = \int \mathcal{L}_j(\beta_0|u_{0j})\phi_{0,\sigma_u^2}(u_{0j})du_{0j},$$

   donde ϕ_{0,σ_u^2} denota la densidad normal de media nula y varianza σ_u^2.

3. Verosimilitud global:

$$\mathcal{L}(\beta_0, \sigma_u^2) = \prod_{j=1}^{J} \int \prod_{i=1}^{n_j} \pi_{ij}^{Y_{ij}}(1-\pi_{ij})^{1-Y_{ij}}\phi_{0,\sigma_u^2}(u_{0j})du_{0j}. \tag{4.2}$$

El problema es que u es no observable, luego la integral de la ecuación (4.2) es difícil de evaluar. En la literatura se han propuesto diversos métodos para solventar esta dificultad. En general, las alternativas que podemos encontrar se basan en modificaciones de la máxima verosimilitud, generalmente apoyadas en métodos de simulación numérica. En concreto, con la función `glmer` de `R`, que es la que hemos utilizado a lo largo de este capítulo para ajustar los modelos, la estimación se obtiene combinando un método QL (*quasi-likelihood*) y la aproximación de Laplace (Rich, 2018).

Con respecto a los efectos aleatorios, la función `ranef` de `R` (perteneciente a la librería `lme4`) nos permite predecirlos a partir de las modas condicionales de dichos efectos, esto es, a partir de las modas de las distribuciones de los efectos aleatorios dados los datos y los parámetros estimados del modelo. Para más información, ver Bates et al. (2015) y Rizopoulos (2012).

Capítulo 5

Aplicación a los datos de TDAH

En este capítulo llevaremos a cabo un estudio de asociación de genoma completo utilizando modelos mixtos lineales generalizados. Como ya sabemos, las variables de efecto adverso de los datos de TDAH son binarias. En el Capítulo 2 hemos visto las limitaciones de los GWAS de modelos logísticos con efectos fijos. Recordemos que para poder ajustar correctamente dichos modelos teníamos que estudiar los tres instantes temporales por separado, con lo cual no podíamos tener en cuenta el progreso a lo largo del tiempo de los individuos. Además, tras hacer una selección de variables, vimos que la inclusión de unas u otras covariables dependía del instante temporal. Los resultados obtenidos no fueron concluyentes. Además, ningún SNP pasó el umbral de significación.

Tras haber introducido los modelos lineales mixtos y los modelos lineales mixtos generalizados se nos abre una nueva puerta para el ajuste de un GWAS. Para empezar, parece claro que nos interesa considerar el *efecto niño* como un efecto aleatorio. Además, con el uso de modelos mixtos podemos incluir toda la información en el mismo modelo, sin la necesidad de tener que hacer un estudio diferente para cada uno de los tres instantes temporales. Para cada uno de los niños se considerará que sus tres medidas pueden estar correlacionadas, mientras que las medidas tomadas sobre dos niños distintos se considerarán independientes.

Recordemos que un estudio de asociación de genoma completo o GWAS consiste en el ajuste de M modelos de regresión, siendo M el número total de SNPs. En nuestro caso, de todos los SNPs de los que tenemos información, seguiremos considerando únicamente los del cromosoma 22, tal y como ya hemos hecho en el Capítulo 2. Además, nos seguiremos centrando en el modo de herencia aditivo.

Para comenzar, realizaremos un GWAS con modelos logísticos mixtos que incluyan el efecto aleatorio del niño, así como la información genotípica de cada SNP. Para cada uno de los M SNPs se ajusta el modelo correspondiente:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{01} SNP_{m,j} + u_{0j},$$

donde $\pi_{ij} = P(Y_{ij} = 1 | SNP_{m,j}, u_{0j})$, γ_{00} es un intercepto global, γ_{01} es el coeficiente asociado al m-ésimo SNP (con $m = 1, \ldots, M$) y u_{0j} es el efecto aleatorio de individuo (se asume $u_{0j} \sim N(0, \sigma_u^2)$ independientes).

En la Fugura 5.1 se muestra el *Manhattan plot* obtenido tras el ajuste. El p-valor más pequeño toma el valor 5.08×10^{-4}, y corresponde al SNP rs17462966. En cuanto a los p-valores asociados al intercepto, en los 17377 modelos se obtiene un valor menor que 4×10^{-3}.

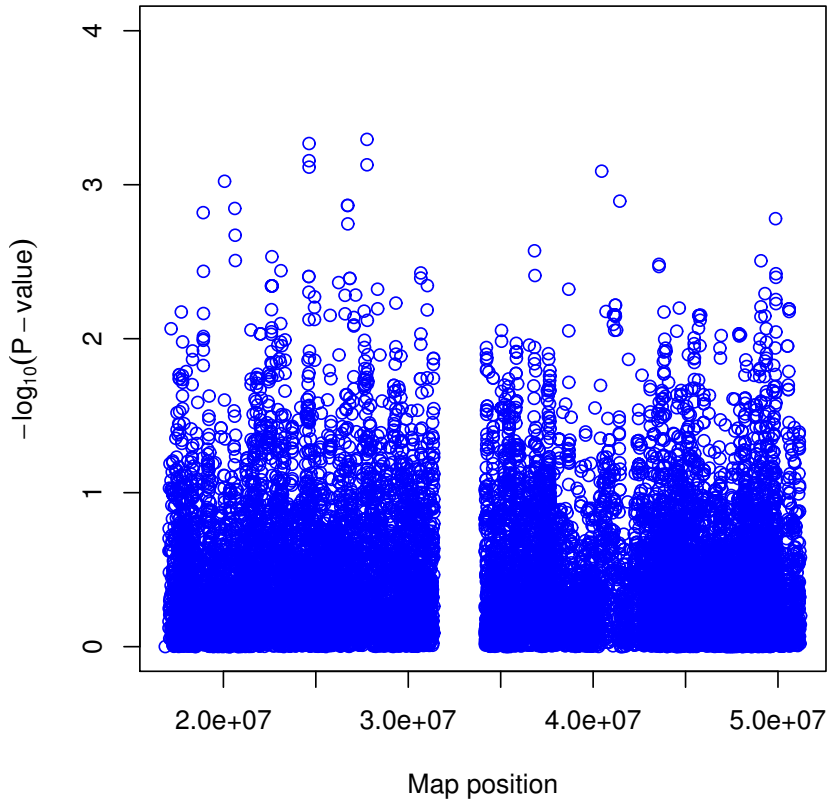


Figura 5.1: *Manhattan plot* correspondiente al GWAS en el que se considera un efecto aelatorio asociado al individuo y la información de los SNPs.

Si en el modelo anterior incluimos la información del fármaco y la dosis (recordemos que se trata de covariables de primer nivel) el modelo pasa a ser el siguiente:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{01} SNP_{m,j} + \gamma_{10}\ \text{fármaco}_{ij} + \gamma_{20}\ \text{dosis}_{ij} + u_{0j},$$

donde $\pi_{ij} = P(Y_{ij} = 1 | SNP_{m,j},\ \text{fármaco}_{ij},\ \text{dosis}_{ij}, u_{0j})$, γ_{00} es un intercepto global, γ_{01} el coeficiente asociado al m-ésimo SNP (con $m = 1, \ldots, M$), γ_{10} el coeficiente asociado al fármaco, γ_{20} el coeficiente asociado a la dosis y u_{0j} el efecto aleatorio de individuo (se asume $u_{0j} \sim N(0, \sigma_u^2)$ independientes).

En este caso, para 44 de los 17377 modelos hubo problemas de convergencia. Recordemos que para la estimación de los parámetros del modelo se recurre a métodos de simulación numérica, que pueden no converger. Por lo tanto, hemos excluido esos 44 modelos para continuar con el estudio. Una vez realizada dicha exclusión, el *Manhattan plot* obtenido es el que se muestra en la Figura 5.2.

El p-valor más pequeño, con un valor de 3.14×10^{-4}, se obtuvo para el SNP kgp6159164. En el caso del intercepto el p-valor más pequeño es 0.16, para el fármaco 0.04 y para la dosis 0.59. Por lo tanto, parece que la inclusión de estas dos covariables no ha producido mejoras.

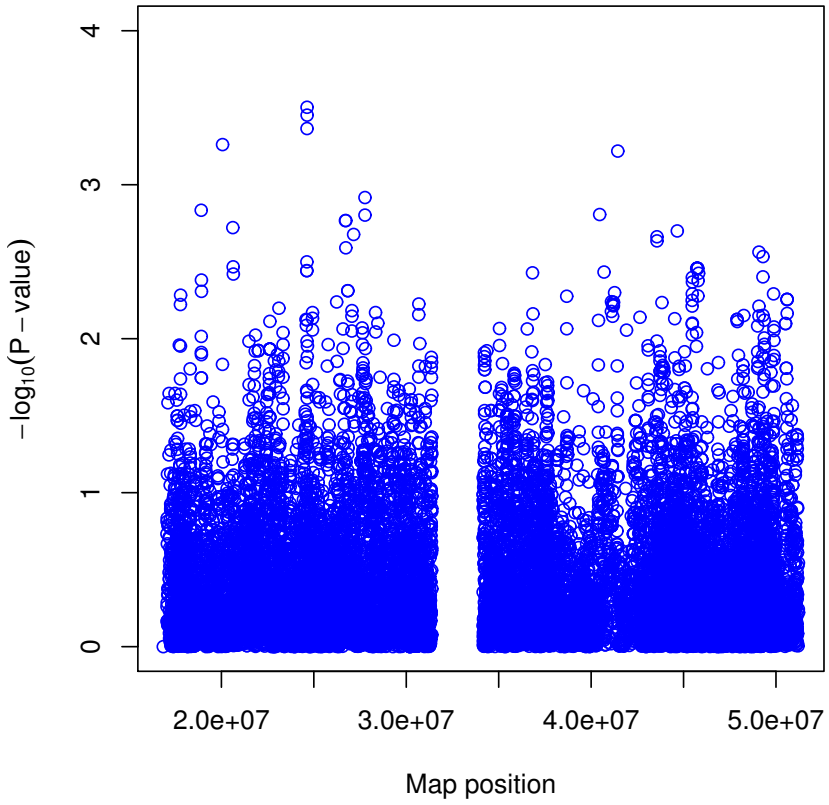


Figura 5.2: *Manhattan plot* correspondiente al GWAS en el que se considera un efecto aelatorio asociado al individuo, la información de los SNPs, el fármaco y la dosis.

Hemos ajustado un tercer GWAS en el que se incluye la covariable tiempo además del SNP, el fármaco, la dosis y el efecto aleatorio asociado al niño. Recordemos que las medidas tomadas sobre cada individuo corresponden a tres instantes temporales distintos, en concreto a los 3, 6 y 12 meses de seguimiento. Consideraremos el tiempo como una covariable de tipo factor, con la finalidad de estudiar el efecto de pasar del instante 3 meses al instante 6 meses, así como el de pasar de 3 a 12. El modelo es el siguiente:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{01} SNP_{m,j} + \gamma_{10}\ \text{fármaco}_{ij} + \gamma_{20}\ \text{dosis}_{ij} + \gamma_{30} I(i=2) + \gamma_{40} I(i=3) + u_{0j},$$

donde $\pi_{ij} = P(Y_{ij} = 1 | SNP_{m,j},\ \text{fármaco}_{ij},\ \text{dosis}_{ij}, i, u_{0j})$, γ_{00} es un intercepto global, γ_{01} el coeficiente asociado al m-ésimo SNP (con $m = 1, \ldots, M$), γ_{10} el coeficiente asociado al fármaco, γ_{20} el coeficiente asociado a la dosis, γ_{30} el asociado al instante temporal 6 meses, γ_{40} el asociado al instante temporal 12 meses y u_{0j} el efecto aleatorio de individuo (se asume $u_{0j} \sim N(0, \sigma_u^2)$ independientes).

En este caso ha habido problemas de convergencia en 698 de los 17377 modelos, por lo que hemos prescindido de ellos. Centrándonos en los modelos restantes, el *Manhattan plot* resultante es el que se muestra en la Figura 5.3.

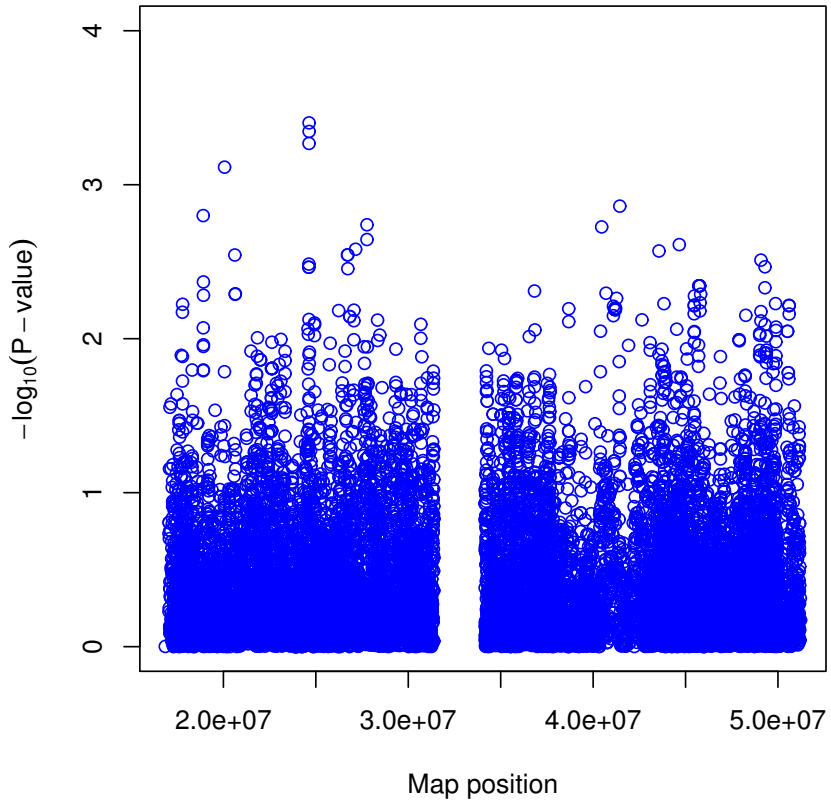


Figura 5.3: *Manhattan plot* correspondiente al GWAS en el que se considera un efecto aelatorio asociado al individuo, la información de los SNPs, el fármaco, la dosis y el tiempo.

El p-valor más pequeño, con un valor de 3.96×10^{-4}, se ha obtenido para el SNP kgp12567948. Para el intercepto el p-valor más pequeño fue 0.16, para el fármaco 0.05, para la dosis 0.32, para el instante 6 meses 0.04 y para el instante 12 meses 0.01. Además, para el instante 12 meses el mayor p-valor obtenido fue 0.06, lo cual corrobora que el impacto del tiempo es importante, al menos cuando nos centramos en el paso de 3 a 12 meses. En tal caso las estimaciones de γ_{40} son todas menores que 0 (con un mínimo de -1.996 y un máximo de -1.173), indicando un efecto protector del paso del tiempo sobre el insomnio.

Por último hemos ajustado un GWAS considerando todas las covariables disponibles, que recordemos que son sexo, edad, subtipo de TDAH, antecedente de tratamiento, fármaco, dosis y tiempo, además de los SNPs y del efecto aleatorio asociado al individuo. En este caso el coste computacional ha sido mayor, y de hecho ha habido problemas de convergencia en todos y cada uno de los 17377 modelos.

Capítulo 6

Conclusiones

En este trabajo hemos llevado a cabo un estudio de asociación de genoma completo con la finalidad de evaluar la respuesta al metilfenidato a largo plazo, considerando un seguimiento de 12 meses. Tal y como indican McCough et al. (2009) y Joensen et al. (2017), la mayoría de los estudios realizados hasta el momento en este contexto evalúan la respuesta al fármaco a corto plazo, y el hecho de seguir a los pacientes a lo largo del tiempo podría darnos información mucho más relevante.

Debido al seguimiento temporal de los pacientes disponemos de medidas repetidas, por lo que hemos utilizado modelos de regresión mixtos para su análisis. Hemos visto que el uso de este tipo de modelos de regresión nos permite analizar toda la información que tenemos de manera conjunta, lo cual supone una gran ventaja frente a los modelos más convencionales utilizados en este contexto, como por ejemplo el modelo de regresión logística. En nuestro estudio las medidas tomadas sobre cada individuo se han registrado en los mismos intantes temporales (a los 3, 6 y 12 meses de seguimiento). No obstante, en el ámbito biomédico es común que no exista homogeneidad en el número de registros ni en los instantes temporales en los que se toman. Los modelos mixtos se pueden utilizar incluso en esas situaciones, por lo que en este sentido son muy flexibles y versátiles.

Desafortunadamente, tras el ajuste del GWAS no hemos encontrado ningún SNP estadísticamente significativo. Es importante recordar que el tamaño muestral de nuestro estudio es $n = 199$. Con la finalidad de obtener mejores resultados, sería conveniente disponer de datos de un mayor número de pacientes.

Además, recordemos que hemos centrado nuestro GWAS únicamente en el cromosoma 22. Sería interesante llevar a cabo un rastreo genómico mucho más completo, considerando todos los SNPs para los cuales tenemos información (que son un total de 2309836, distribuidos por todos los cromosomas). Debido al coste computacional que esto supone, se necesitaría hacer uso de las instalaciones del Centro de Supercomputación de Galicia (CESGA). A causa de la sitación actual, el CESGA prioriza las actividades relacionadas con la lucha contra la COVID-19. No obstante, sería un trabajo interesante de cara al futuro.

Por último, no debemos olvidar que las personas con TDAH y otros trastornos son a menudo objeto de estigmatización y discriminación. La investigación puede ser una buena herramienta para darles visibilidad y para luchar por su inclusión.

Apéndice A

Selección incondicional de covariables

En el Capítulo 2 hemos hecho una selección de covariables condicional al SNP previa al ajuste de los modelos de regresión logística correspondientes a cada instante temporal. Hemos decidido usar ese enfoque teniendo en cuenta que nuestro principal objetivo es estudiar la significación de cada uno de los SNPs. No obstante, podríamos plantearnos también realizar una selección de covariables incondicional al SNP, más general. Tras estudiar la significación de cada covariable siguiendo este enfoque más general, para el efecto adverso insomnio únicamente se han obtenido p-valores menores que 0.1 en los siguientes casos:

- 3 meses. El coeficiente asociado al subtipo, concretamente el correspondiente al paso de presentación predominante de falta de atención a presentación predominante de hiperactividad e impulsividad, tiene un p-valor asociado de 0.09.
- 6 meses. La covariable fármaco tiene un p-valor asociado de 0.07.
- 12 meses. El p-valor asociado a la edad es 0.09; el del coeficiente correspondiente al paso de presentación predominante de falta de atención a presentación predominante de hiperactividad e impulsividad, 0.03; el del fármaco, 0.09.
- Global. Ninguna covariable tiene un p-valor asociado menor que 0.1.

Hemos seguido el mismo procedimiento para el efecto adverso de falta de apetito, obteniendo los siguientes resultados:

- 3 meses. La covariable fármaco tiene un p-valor asociado de 0.08.
- 6 meses. La covariable antecedente de tratamiento tiene un p-valor asociado de 0.02, mientras que la covariable fármaco tiene un p-valor asociado de 0.04.
- 12 meses. El p-valor asociado al fármaco es 0.04.
- Global. Ninguna covariable tiene un p-valor asociado menor que 0.1.

Bibliografía

Bates DM, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1):1–48.

Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, Russo MB, Balkin TJ (2003) Patterns of Performance Degradation and Restoration During Sleep Restriction and Subsequent Recovery: A Sleep Dose-Response Study. *Journal of Sleep Research* 12:1–12.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 57(1):289–300.

Benjamini Y, Heller R (2007) False discovery rates for spatial signals. *Journal of the American Statistical Association* 102:1272–1281.

Biederman J, Mick E, Faraone SV, Braaten E, Doyle A et al. (2002) Influence of gender on attention deficit hyperactivity disorder in children referred to a psychiatric clinic. *American Journal of Psychiatry* 159(1):36–42.

Catalá-López F, Peiró S, Ridao M, Sanfélix-Gimeno G, Gènova-Maleras R, Catalá MA (2012) Prevalence of attention deficit hyperactivity disorder among children and adolescents in Spain: a systematic review and meta-analysis of epidemiological studies. *BMC Psychiatry* 12(1):168.

Chung CC, Magalhaes WC, González-Bosquet J, Chanock SJ (2010) Genome-wide association studies in cancer – current and future directions. *Carcinogenesis* 31(1):111–120.

Dennison CA, Legge SE, Pardiñas AF, Walters J (2020) Genome-wide association studies in schizophrenia: Recent advances, challenges and future perspective. *Schizophrenia research* 217:4–12.

Durbán M (s.f.) *Introducción a los modelos mixtos.* Departamento de Estadística, Universidad Carlos III de Madrid.

Faraone SV, Biederman J, Weber W & Russell RL (1998) Psychiatric, neuropsychological, and psychosocial features of DSM-IV subtypes of attention-deficit/hyperactivity disorder: results from a clinically referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry* 37:185–193.

Faraone SV, Doyle AE (2001) The nature and heritability of attention-deficit/ hyperactivity disorder. *Child and Adolescent Psychiatric Clinics of North America* 10:299–316.

Faraone SV, Perlis RH, Doyle AE, Smoller JW, Goralnick JJ, Holmgren MA et al. (2005) Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57(11):1313-1323.

Gómez-Sánchez CI (2017) *Genética y farmacogenética del trastorno por déficit de atención e hiperactividad en niños de la población española.* Universidad Autónoma de Madrid. Departamento de biología molecular.

Henderson CR, Kempthorne 0, Searle SR, von Krosigk CN (1959) Estimation of environmental and genetic trends from records subject to culling. *Biometrics* 13:192-218.

Hodgkins P, Shaw M, Coghill D, Hechtman L (2012) Amfetamine and methylphenidate medications for attention-deficit/hyperactivity disorder: complementary treatment options. *European Child and Adolescent Psychiatry* 21(9):477-492.

Jannot AS, Ehret G, Perneger T (2015) $P < 5 \times 10^{-8}$ has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology* 68:460-465.

Joensen B, Meyer M, Aagaard L (2017) Specific genes associated with adverse events of methylphenidate use in the pediatric population: a systematic literature review. *Journal of Research in Pharmacy Practice* 6(2):65–72.

Johnston BA, Coghill D, Matthews K, and Steele JD (2015) Predicting methylphenidate response in attention deficit hyperactivity disorder: a preliminary study. *Journal of Psychopharmacology* 29(1):24–30.

Lange KW, Reichl S, Lange KM, Tucha L, Tucha O (2010) The history of attention deficit hyperactivity disorder. *Attention Deficit and Hyperactivity Disorder* 2:241–255.

Larson K, Russ SA, Kahn RS, Halfon N (2011) Patterns of comorbidity, functioning, and service use for US children with ADHD. *Pediatrics* 127:462–470.

McGough JJ, McCracken JT, Loo SK, Manganiello M, Leung MC et al. (2009) A candidate gene analysis of methylphenidate response in attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* 48(12):1155–1164.

Novik TS, Hervas A, Ralston SJ, Dalsgaard S, Rodrigues Pereira R et al (2006) Influence of gender on attention-deficit/hyperactivity disorder in Europe ADORE. *European Child and Adolescent Psychiatry* 15(Suppl 1):15–24.

Rich JL (2018) *Comparison of Generalized Linear Mixed Model Estimation Methods.* Department of Mathematical Sciences Montana State University.

Rizopoulos D (2012) *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R.* Chapman & Hall/CRC Biostatistics Series.

Searle SR, Casella G, McCulloch CE (1992) *Variance Components.* John Wiley & Sons.

The International HapMap Consortium. A haplotype map of the human genome (2005) *Nature* 437:1299-1320.

Thomas R, Sanders S, Doust J, Beller E, Glasziou P (2015) Prevalence of Attention-Deficit/Hyperactivity Disorder: A Systematic Review and Meta-analysis. *Pediatrics* 135:994–1001.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S.* Springer.

Xue A, Wu Y, Zhu Z et al. (2018) Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications* 9(1):2941.