Traballo Fin de Mestrado

Análise comparativa de modelos de regresión en estudos GWAS: aplicación ao estudo da base xenética da severidade de infección por COVID-19

Silvia Diz de almeida

Mestrado en Técnicas Estatísticas

Curso 2020-2021

Proposta de Traballo Fin de Mestrado

Título en galego: Análise comparativa de modelos de regresión en estudos GWAS: aplicación ao estudo da base xenética da severidade de infección por COVID-19

Título en español: Análisis comparativo de modelos de regresión en estudios GWAS: aplicación al estudio de la base genética de severidad de infección por COVID-19

English title: Comparative analysis of regression models in GWAS studies: application to the study of the genetic basis of COVID-19 infection severity

Modalidade: Modalidad B

Autora: Silvia Diz de almeida, Universidade de Santiago de Compostela

Directora: Rosa María Crujeiras Casais, Profesora titular do Departamento de Estatística, Análise Matemática e Optimización na Universidade de Santiago de Compostela

Titora: Raquel Cruz Guerrero, investigadora do CIBERER no Centro de Investigación en Medicina Molecular e Enfermidades Crónicas (CIMUS)

Nos estudos de asociación xenética, a pequena ou gran escala (GWAS, *genome-wide association studies*), a identificacion de que polimorfismos xenéticos son relevantes na determinación do risco dunha patoloxía realízase mediante estudos de tipo caso-control e normalmente se analiza -de forma individual para cada marcador- mediante regresión loxística. O GWAS que se está a realizar no marco do proxecto *Determinantes xenéticos e biomarcadores xenómicos de risco en pacentes con infección por coronavirus* proporcionará información xenética e clínica detallada dunha gran cantidade de individuos, que serán ademais clasificados en función da severidade da súa infección por COVID-19, sendo estes datos idóneos para a aplicación de diferentes modelos de regresión que permitan analizar globalmente o contunxo de casos. Ao incluír no estudo a diferenciación no grupo de casos, categorizados por gravidade da enfermidade ou severidade de síntomas ou consecuencias, faise necesario o uso doutros modelos de regresión que consideren, por un lado, variables resposta con máis de dúas categorías e/ou a inclusión doutras fontes de variabilidade mediante o uso de modelos mixtos.

Dona Rosa María Crujeiras Casais, da Profesora titular do Departamento de Estatística, Análise Matemática e Optimización na Universidade de Santiago de Compostelae dona Raquel Cruz Guerrero, investigadora do CIBERER no Centro de Investigación en Medicina Molecular e Enfermidades Crónicas (CIMUS), informan que o Traballo Fin de Mestrado titulado

Análise comparativa de modelos de regresión en estudos GWAS: aplicación ao estudo da base xenética da severidade de infección por COVID-19

foi realizado baixo a súa dirección por dona Silvia Diz de almeida para o Mestrado en Técnicas Estatísticas. Estimando que o traballo está terminado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 02 de xaneiro de 2021.

A 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: 2021.02.04 09:22:13 +01'00'

Dona Rosa María Crujeiras Casais

A titora:

CRUZ GUERRERO, RAQUEL (FIRMA)

Firmado digitalmente por CRUZ GUERRERO, RAQUEL (FIRMA)
Nombre de reconocimiento (DN): c=ES, serialNumber=12383132R, sn=CRUZ, givenName=RAQUEL, cn=CRUZ GUERRERO, RAQUEL (FIRMA)
Fecha: 2021.02.04 09:59:44 +01'00'

Dona Raquel Cruz Guerrero

A autora:

Firmado digitalmente por Silvia Diz de Almeida (39485530G) el día 04/02/2021.

Dona Silvia Diz de almeida

Agradecementos

En primeiro lugar gustaríame darlle as grazas a Raquel por ter confiado en mi e terme dado esta oportunidade tan bonita, pero sobre todo por permitirme aprender e seguir aprendendo dela.

A Rosa, pola paciencia e pola dedicación en todo momento. Tamén por contaxiarme o seu amor pola estatística.

A Iván, por darme as forzas que necesitaba día a día.

A Valeria e a Nuria, porque foron un apoio fundamental no camiño.

Por último, agradecerlle tamén ao CESGA (Centro de Supercomputación de Galicia) o uso das súas instalacións.

Índice xeral

Resumo

Resumo en galego

Existe unha gran variabilidade ao redor das manifestacións clínicas da enfermidade por SARS-CoV-2 e unha das posibles razóns pode atoparse nos xenes dos hóspedes. Esta base xenética pódese detectar grazas aos estudos de asociación de xenoma completo (GWAS). Os GWAS son habitualmente estudos tipo casos-controis mediante os que se identifican variantes vinculadas á presenza de enfermidade. Así, o estudo clásico consiste no axuste dunha regresión loxística simple que require un control de calidade previo exhaustivo. Ao longo deste traballo explóranse alternativas a estes modelos, por exemplo coa inclusión de efectos aleatorios ou empregando unha variable resposta multinomial. O obxectivo principal será o de determinar os factores xenéticos de risco asociados á gravidade da infección por SARS-CoV-2 a partir dun GWAS que se realizará sobre un conxunto de máis de 9.000 individuos para os que se dispón de datos clínicos e xenéticos.

English abstract

The variability among SARS-CoV-2 clinical manifestations is well-known and one of the reasons behind this might be found in the genes of the hosts. These genetic factors can be detected by conducting genome-wide association studies (GWAS). Usually, GWAS are case-control studies which identify disease-linked polymorphisms by fitting a logistic regression, preceded by an exhaustive quality control. This work aims to explore alternatives to these models, either by including random effects or using a multinomial response variable. The main objective will be to determine risk genetic factors associated with the severity of SARS-CoV-2 infection by carrying out a GWAS on more than 9.000 individuals for which clinical and genetic information is available.

Capítulo 1

Introdución

Os métodos empregados para determinar as bases xenéticas das enfermidades evolucionaron moi rápido nos últimos anos. Nos seus inicios baseábanse na posición física dos xenes, para crear os chamados mapas xenéticos, ou nas análises de pedigrí para establecer os patróns de herdanza. Estas técnicas tiveron moito éxito no descubrimento de marcadores xenéticos[1] para certas patoloxías, mais os chamados caracteres complexos (os cales están determinados por máis dunha variante xenética ou por factores medioambientais) precisaban de investigacións a nivel poboacional (Morris e Cardon, 2019). Como consecuencia cobraron importancia os estudos de asociación, entre os que se atopan os estudos de xenes candidatos e os estudos de asociación de xenoma completo (GWAS, do inglés *Genomic-Wide Association Studies*). O obxectivo dos mesmos é identificar os marcadores que se atopan con maior frecuencia entre a poboación que presenta o carácter de interese e será nos GWAS no que nos centraremos neste traballo.

Os marcadores xenéticos que se empregan na maioría dos estudos son os chamados SNPs (*single nucleotid polymorphism*). Un SNP é un cambio nun único nucleótido (nunha mesma posición) e son o tipo de variación máis común no ADN. Cada variación do SNP chámase alelo e, á sua vez, poden existir varios alelos que presenten frecuencias distintas dependendo da poboación, aínda que soen ser polimorfismos bialélicos. As frecuencias dos SNPs danse habitualmente en función do alelo menos frecuente a través da MAF (*minor allele frequency*). Dependendo de que alelos herde dos pais un individuo para un SNP, o seu xenotipo pode ser homocigoto para o alelo frecuente (AA), heterocigoto (Aa) ou homocigoto para o alelo menos frecuente(aa).

Estes polimorfismos, na súa gran maioría, non provocan cambios substanciais ou funcionais no ADN dos individuos, pero están documentadas e actúan como marcadores. Por tanto, o resultado final dun estudo de asociación será identificar SNPs vinculados co fenotipo[2]. Neste punto é interesante describir o desequilibrio de ligamento, pois é un aspecto chave dos estudos

[1] Os marcadores xenéticos son secuencias de ADN cuxa posición no cromosoma é coñecida.

[2] Fenotipo: Expresión do xenotipo no individuo. Por exemplo, ser albino é un fenotipo, pero portar a mutación do albinismo é o xenotipo.

de asociación.

Nunha mesma poboación existen fraccións do ADN compartidas entre os individuos, chamadas haplotipos. Estas variantes que conforman os haplotipos hérdanse en bloque e, por tanto, que dúas persoas presenten os mesmos alelos para estas variantes non é consecuencia da aleatoriedade. Cando existe asociación non aleatoria entre os alelos fálase de que os loci (plural de locus, posición no cromosoma dunha variante ou xen) se atopan en desequilibrio de ligamento (LD, *linkage desequilibrium*). Se estes son independentes, fálase de equilibrio de ligamento (McVean e Kelleher, 2019). Unha medida do LD é a diferencia entre a frecuencia observada dos n alelos considerados e a frecuencia esperada baixo independencia. En termos de probabilidade é moi sinxelo de entender. Consideremos dous loci 1 e 2, cos alelos (A/a) e (B/b), respectivamente. O alelo A no locus 1 ten unha probabilidade de ocorrer p_A e o mesmo co alelo B do locus 2. Considerando independencia entre ambos, a frecuencia esperada do haplotipo AB sería $p_{AB} = p_A p_B$, sendo p_A a probabiliade do alelo A e p_B a probabilidade do alelo B. O coeficiente de LD, denotado D é $D_{AB} = p_{AB} - p_A p_B$. Así, se a frecuencia co que se atopa o haplotipo AB na mostra é menor ou maior que a frecuencia baixo independencia, os loci atópanse en desequilibrio de ligamento. Este concepto é interesante porque, volvendo aos estudos de asociación, un SNP pode non ser causante dunha patoloxía pero proporcionará pistas sobre a variante causal ao atoparse en desequilibrio de ligamento con ela. A partir dos resultados obtidos realízanse multitude de experimentos ata dar coa variante causal.

Os recentes avances e abaratamento no campo da secuenciación permitiron estender estes experimentos a todo o xenoma. Precisamente, na actualidade os estudos de asociación máis populares son os GWAS, nos que se escanea a totalidade do xenoma a partir de centos de miles de SNPs en busca de variantes asociadas ao fenotipo de análise. O fenotipo investigado pode ou ben ser cuantitativo (por exemplo, o peso) ou binario. Neste último caso realízase un estudo de tipo casos-controis, nos cales se comparan individuos que presentan a enfermidade cos que non a presentan. O enfoque metodolóxico dun GWAS consiste en axustar M modelos de regresión (lineais ou loxísticos), sendo M o número de SNPs considerado e N o número de individuos, baixo un modelo xenético de herdanza[3]. Tras o axuste obtéñense os coeficientes asociados ás variables, entre os que se atopará o asociado ao SNP: será o efecto estimado do marcador. Sobre estes coeficientes lévanse a cabo contrastes de hipóteses, cuxos p-valores nos axudarán a identificar os SNPs significativos.

Aquí o patrón de herdanza que se usará será o aditivo, baixo o cal o SNP pode tomar valores 0, 1 ou 2 segundo o número de copias do alelo menos frecuente (chamémoslle a) que posúa o individuo. Desta forma, o alelo a poderá ter un efecto positivo sobre o caracter de interese ou ter un efecto deletéreo. Un SNP significativo cuxo alelo a se atope máis frecuentemente nos casos que nos controis constituirá un factor de risco xenético. É esencial controlar todos os factores que poidan nesgar o estudo e incrementar a porcentaxe de asociacións falsas. Por último, é importante apuntar que os efectos obtidos para os SNPs son habitualmente moi baixos, pois na expresión dos caracteres complexos non actúa unha única variante. Esta é unha limitación que presentan. Tras a detección das asociacións realízanse estudos de replicación e de detección do xen implicado (Morris e Cardon, 2019).

[3]Hai catro modelos de herdanza: Aditivo, dominante, codominante e recesivo.

Neste traballo levaranse a cabo un GWAS de tres formas diferentes: dúas con variable binaria (casos/controis) e unha terceira cunha variable multinomial, introducindo así un modelo non tan común nestes estudos.

1.1. Motivación do traballo

A hipótese coa que se traballa nos GWAS é que existe unha base xenética de risco compartida entre os casos, polo que ademais de definir os casos e controis adecuadamente, débese controlar calqueira fonte de variabilidade xenética que poida introducir ruído. Ademais, os modelos empregados asumen independencia entre os individuos. Existen varios factores na estructura xenética dunha poboación que alteran esta suposta independencia das mostras e poden actuar como variables confusoras, aparecendo asociacións espúreas entre un fenotipo e certas variantes. Dita cuestión será a motivación principal dos dous primeiros modelos que axustremos: o loxístico simple, con observacións independentes, e o loxístico mixto, asumindo dependencia entre os datos.

Estrutura xenética en estudos GWAS

Os dous factores confusores máis relevantes son a estratificación poboacional e a presencia de relacións de parentesco entre os individuos, como está demostrado en Devlin e Roeder (1999). Estes autores propuxeron un método coñecido como *genomic control*, λ_{GC}[4] que aínda se aplica. Sen embargo, esta aproximación ten limitacións (Price, 2010).

A estratificación poboacional maniféstase nas diferentes frecuencias alélicas entre as múltiples poboacións xenéticas que conforman un estudo. Se estas poboacións non están representadas na mesma magnitude nos casos e controis poden aparecer falsas asociacións. Por exemplificar, supoñamos que como consecuencia das presións selectivas ás que foi sometido un SNP, o alelo menos frecuente en Europa é o alelo común en África. Ademais, no noso estudo hai unha porcentaxe de mostras de orixe africana dentro dos casos pero ningunha nos controis: pode darse un falso positivo entre a variable Y e este alelo por ter maior probabilidade de aparecer entre os casos. Dada a dificultade de equilibrar as poboacións ancestrais entre os grupos dos casos e dos controis, propuxéronse numerosos métodos estatísticos para suavizar o seu efecto como é a análise de compoñentes principais (PCA).

Cavalli-Sforza e Edwards (Cavalli-Sforza et al., 1994) empregaron en 1964 a análise de compoñentes principais sobre frecuencias xénicas poboacionais. Este enfoque permitiulles inferir os eixos de variabilidade xenética a nivel mundial, reducindo a información a unhas poucas compoñentes principais (PCs). Curiosamente, a maioría das veces os mapas de PCs podíanse interpretar xunto con mapas xeográficos, pois a medida que aumenta a distancia xeográfica tende a ocorrer o mesmo coa distancia xenética (Cavalli-Sforza et al., 1994; Novembre et al., 2008). Na actualidade, esta técnica multivariante realízase sobre o xenotipo de miles de SNPs

[4]Defínese como o cociente entre a mediana da distribución dos estatísticos dos tests de asociación e a mediana da distribución do estatístico baixo a hipótese nula (χ^2 con 1 grao de liberdade). Unha vez calculado λ, reescálanse os estatísticos por este valor.

(Price, 2006).

Os autovectores e autovalores calcúlanse a partir dunha matriz formada polos xenotipos dos individuos para un conxunto de miles de SNP, que deben atoparse en equilibrio de ligamento. Dado un conxunto de N individuos e M SNPs, M: vai ser $\{G_{im}\}$ con $m = 1, \dots, M$. Ou sexa, $\{G_{i1}, G_{i2}, \dots, G_{im}\}$ serán os xenotipos para cada individuo e tomarán os valores 0, 1 ou 2 en función do número de copias do alelo menos frecuente. Desta forma temos unha matriz T de dimensión $(N \times M)$. Seguindo o método de Price (2006), T estandarizarase extraendo a media de cada fila

$$\mu_i = \frac{\sum_{i=1}^{N} G_{im}}{N}, \quad i = 1, \dots, N$$

e dividindo por

$$\sqrt{p_i(1 - p_i)};$$

para cada $i = 1, \dots, N$. Á súa vez,

$$p_i = \frac{1 + \sum_{i=1}^{N} G_{im}}{2 + 2N},$$

é a frecuencia alélica do SNP m na mostra. A nova matriz será denotada X e os autovectores e autovalores extraeranse de $\Psi = X^T X$. Unha vez calculadas as compoñentes principais, que se obteñen a partir de $\hat{\Psi}$, engádense como covariables na regresión.

O segundo factor que se mencionou anteriormente é o parentesco críptico (*cryptic relatedness*), relacións de parentesco descoñecidas para os investigadores. Supoñamos agora que temos varios individuos dunha mesma familia no grupo de casos, é dicir, todos manifestan o fenotipo de interese. Estes individuos serán máis cercanos xenéticamente entre eles que cos controis, pois existe correlación alélica (Devlin, 1999). Como consecuencia, os seus xenotipos non serán mostras aleatorias independentes das frecuencias alélicas poboacionais e poden surxir falsas asociacións derivadas de compartir base xenética sen ter esta que ver coa enfermidade. Esta dependencia poderíase correxir en caso de coñecer o pedigrí ou facendo análises intra-familiares, pero en moitas ocasións non é posible ou é moi custoso.

A práctica estándar para solucionar este problema é retirar do estudo todos os individuos que presentan lazos familiares ata certo grao salvo un. Existen moitas medidas coas que determinar a similaridade xenética entre dous individuos, e unha delas é o *identity by descent(IBD)*. Un par de persoas poden compartir secuencias nucleotídicas ou segmentos do xenoma exactamente iguais, e dise que estos son *identical by state(IBS)*. Se estes segmentos son herdados dun ancestro común fálase de que son *identical by descent (IBD)*. Esta medida non só se calcula a partir de rexións do xenoma, senón que se pode calcular a partir de SNPs. Non entraremos en detalles sobre o proceso, pero o software `PLINK` (véxase Apéndice B) emprega o método dos momentos para estimar a probabilidade de compartir 1, 2 ou 0 alelos por IBD para un par de individuos (Purcell et al., 2007). A continuación, calcula a proporción de IBD a partir das probabilidades de compartir 1 ou 2 alelos por descendencia:

$\hat{\pi} = P(Z = 2) + 0{,}5P(Z = 1)$. Os valores esperados de IBD son os mostrados no Cadro 1.1. Anderson et al. (2010) recomendan retirar un individuo de cada par cando o $IBD \geq = 0{,}185$, que estarían a medio camiño entre parentes de segundo e terceiro grao.

IBD	**Parentesco**	**Exemplo**
1	Xemelgos monocigóticos	*Xemelgos. Tamén serve para detectar mostras duplicadas.*
0.5	Parentesco de primeiro grao	*Pais, fillos ou irmáns completos.*
0.25	Parentesco de segundo grao	*Avós, netos, tíos, sobriños e medio-irmáns.*
0.125	Parentesco de terceiro grado	*Bisavós, bisnetos, tíos segundos e primos.*

Cadro 1.1: Táboa cos valores de IBD esperados para cada par de individuos segundo as relacións de parentesco.

Típicamente, no axuste dun modelo loxístico elimínanse as persoas cun alto índice de parentesco entre si a través do cálculo do IBD e engádense como covariables as 10 primeiras compoñentes principais por convenio. Sen embargo, sábese que os modelos mixtos permiten controlar ambos confusores sen a necesidade de reducir o tamaño de mostra, xa que se poden modelizar a estrutura poboacional -en forma de PCs- e os xenotipos para cada SNP como efectos fixos, mentres que a relación de parentesco se axusta como efecto aleatorio (Yu et. al, 2006). Estes modelos están cada vez máis estendidos nos estudos GWAS (Price, 2010) e serán un obxecto de estudo neste traballo.

1.2. Descrición dos datos do estudo

A epidemia do SARS-CoV-2 deulle a volta ao mundo en cuestión de poucos meses. Descuberto en decembro de 2019 (Zhu et al., 2020), este novo tipo de coronavirus (COVID-19) infectou a máis de 100 millóns de persoas en todo o mundo e 2 millóns de mortes (Worldometers, 2021). En España, o número de infectados acumulado en xaneiro de 2021 supera os 2.7 millóns e 55.000 defuncións. Os síntomas varían dende tos a pneumonías severas. Sen embargo, as manifestacións clínicas non están totalmente delimitadas, xa que a maioría dos pacentes ou ben non presentan síntomas ou estes son moderados (The Severe Covid-19 GWAS Group, 2020). Concretamente en España, segundo os datos recollidos por Casas-Rojo et. al (2020), as manifestacións clínicas máis frecuentes nas persoas hospitalizadas son febre, tos, disnea e astenia, ademais de síntomas gastrointestinais como a diarrea. A maioría dos ingresados eran homes de máis de 50 anos. En canto ás comorbilidades que podían presentar, a metade dos enfermos do estudo padecía hipertensión, mentres que tamén era común padecer obesidade, dislipidemia e diabetes mellitus (Casas-Rojo et. al, 2020). Isto demostra que existe unha clara variabilidade en canto á gravidade da enfermidade por COVID-19.

Podería ser unha causa a xenética dos hóspedes? Sabemos de antemán que os nosos xenes evolucionaron á par dos xenomas dos diferentes microorganismos infecciosos grazas a un proceso chamado co-evolución (Kwok et al., 2020). A co-evolución provoca que dúas especies que interaccionan entre si evolucionen paralelamente, xa ben sexa para sobrevivir ou para beneficiarse entre elas. De feito, para enfermidades como a malaria (Malaria Genomic Epidemiology Network, 2015) identificáronse xenes asociados a unha maior suceptibilidade á enfermidade. Tamén se obtiveron resultados en estudos para a hepatitis viral ou a tuberculose (Kwok et al., 2020). Normalmente estas asociacións inclúen o complexo HLA (*human leukocyte antigen*), que codifica unhas moléculas que interveñen na resposta inmunitaria, pero non en todos casos é así. Deste xeito, non sería extraño que existiran xenes nunha parte da poboación que conferiran unha protección -ou susceptibilidade- ao hóspede fronte á infección por coronavirus.

O consorcio *The Severe Covid-19 GWAS Group* (The Severe Covid-19 GWAS Group, 2020) identificou un clúster de xenes nunha rexión concreta do cromosoma 3, entre os que están o SCL6A20 ou LZFTL1, asociados ao fallo respiratorio en hospitalizados por COVID-19. Posteriormente revelouse que as variantes consideradas de risco por este estudo estaban presentes en forma homocigótica (dúas copias) no xenoma Neanderthal (Zeberg e Pääbo, 2020), o que significa que esta rexión de susceptibilidade foi herdada a partir da especie Neanderthal e non se atopa en todos os humanos modernos. Estes autores explican que esta rexión foi sometida a selección positiva en certas poboacións, é dicir, que nun momento da historia recente conferiu protección fronte a algún factor medioambiental, mais agora se atopa en selección negativa debido ao COVID-19. Por outro lado, Pairo-Castineira et al. (2020) atoparon resultados adicionais nos cromosomas 6, 12, 19 e 21. O clúster de xenes OAS do cromosoma 12, por exemplo, xa se identificara previamente como asociado ao SARS-CoV, e intervén na replicación dos coronavirus na célula (Pairo-Castineira et al., 2020). Outros xenes, como o DPP9 no cromosoma 13, están asociados a enfermidades respiratorias como a fibrose pulmonar idiopática. Con todo, este estudo tamén se realizou con pacientes que requeriron apoio cardiorespiratorio continuo, non contemplando máis que o fenotipo severo da enfermidade por SARS-CoV-2.

O presente traballo enmárcase no proxecto *Determinantes xenéticos e biomarcadores xenómicos de risco en pacientes con infección por coronavirus SARS-COV-2*[5], tamén denominado *SCOURGE (Spanish COalition to Unlock Research on host GEnetics on COVID-19)*. Á súa vez é partícipe do consorcio *COVID-19 Host Genetics Initiative* (COVID-19 Host Genetics Initiative, 2020). O obxectivo principal do proxecto é o de analizar a base xenética de pacentes con coronavirus e determinar se existe algunha relación entre esta e as manifestacións clínicas derivadas da infección. A gran cantidade de comorbilidades e antecedentes recollidos no estudo permítenos axustar diferentes modelos de asociación e precisar múltiples factores de risco, analizando o espectro completo da enfermidade e non só a máis grave.

Usaremos só algunhas das variables, a partir das cales se construirá a variable dependente principal: a de gravidade. Na recollida dos datos participaron hospitais de toda España, polo que dispoñemos dunha mostra que cubre a maior parte da poboación. O criterio de inclusión no

[5]Proxecto financiado polo Instituto Carlos III de Madrid. FONDO-COVID19. IP: Dr. Ángel Carracedo e Dr. Pablo Lapunzina.

estudo foi o de ser diagnosticado con COVID-19, xa ben fora mediante diagnóstico molecular (PCR, test de antíxenos) ou clínico. Algúns individuos recrutados no inicio tiveron finalmente un diagnóstico negativo e foron incluídos na poboación control. Xenotipáronse mostras e obtivéronse datos clínicos e demográficos de máis de 10.000 persoas ao longo do ano 2020. Ademais, incluíronse 3358 mostras do Banco Nacional de ADN para comparar os indiviuos enfermos con unha poboación control. Por contextualizar o noso tamaño de mostra, no estudo de The Severe Covid-19 GWAS Group (2020) a mostra estaba formada por 1980 pacentes e 2381 controis; mentres que no de Pairo-Castineira et. al (2020), o número de casos europeos foi de 1676.

Con estes datos levaremos a cabo un estudo de asociación, a modo de proba, sobre o cromosoma 3. A modo de puntualización, este cromosoma ten ao redor de 40.000 SNPs e axustar un número tan alto de modelos require un tempo de computación elevado, polo cal foi necesario usar o CESGA (Centro de Supercomputación de Galicia, ver Apéndice A).

Usaremos dous conxuntos de individuos de ancestralidade únicamente europea. Nun primeiro conxunto de datos mantemos os familiares, con $N = 10190$ individuos. Para o segundo conxunto de datos eliminároнse todos as mostras emparentadas menos 1 por familia, quedando 9896 individuos. É importante destacar que, debido á falta de datos nalgunha das covariables, o N final cambiará. Estas variacións comentaranse nos apartados dos axustes. Ademais, realizouse un control de calidade en `PLINK` (Purcel et al., 2007; Anderson et al., 2010) a nivel de individuo e de SNP (ver Apéndice A). Por simplificación resumiranse, a continuación, as variables no conxunto de individuos con $N = 9896$, xa que o número de parentes eliminados foi de 268 e non variará sustancialmente a análise exploratoria.

- **Variables demográficas**: Empregaranse as variables *sexo* e *idade*. No Cadro 1.2 resúmese a distribución de idade por sexo, tanto para os casos como para o controis do Banco Nacional de ADN. A mediana de idade do total é de 55 anos e a ratio home/muller é de 4881/5015.

Idade		Sexo		Poboación total
		Homes	Mulleres	
Primeiro cuantil	Casos	53	48	50
	Controis	41	41	41
Mediana	Casos	64	61	62
	Controis	46	48	47
Terceiro cuantil	Casos	76	80	77
	Controis	54	56	55

Cadro 1.2: Táboa dos cuantís 25, 50 e 75 da idade na poboación total e estratificados por sexo e por grupo de casos ($N = 6538$) e controis ($N = 3358$).

- **Variables clínicas**: A variable principal será a de *severidade*, unha variable nominal categórica construída a partir doutras variables clínicas. Constará de cinco categorías para os enfermos, que tomarán os valores do 1 ao 5. Ademais, codificados como 0 estarán os controis do Banco Nacional de ADN e as mostras cuxo diagnóstico do COVID-19 foi negativo. A escala seguirá os seguintes criterios:

  - Asintomático: Codificado como 1, o individuos que caen nesta categoría deron positivo en COVID-19 a través dalgún dos diagnósticos pero non presentaron síntomas.

  - Leve: Codificado como 2. O pacente puido estar hospitalizado ou non pero presentou síntomas e non recibiu osíxenoterapia convencional. Ademais, ou ben non se lle realizou radiografía de tórax nin TAC ou ben non se produciu infiltración (non houbo afectación radiolóxica).

  - Moderado: Codificado como 3. O pacente non foi asintomático, e puido ou non estar hospitalizado. Sen embargo, debeu recibir osixenoterapia convencional ou ter infiltración pulmonar ou ter unha afectación máxima menor ao 50 %.
  - Grave: Codificado como 4. O pacente debe cumprir algunha das seguintes condicións: a presión de osíxeno $PaO_2 < 65$ 65mmHg/saturación $SaO_2 < 90\,\%$; SaO_2/FiO_2 mínima < 440; PaO_2/FiO_2 mínima < 300; presentar disnea; ter unha frecuencia respiratoria maior ou igual a 22rpm, ou que a afectación pulmonar máxima en radiografía de tórax fora maior ao 50 %.
  - Crítico: Codificado como 5. O pacente estivo ingresado na UCI, recibiu ventilación mecánica non invasiva, usou cánulas nasais de alto fluxo, ou faleceu por mor do COVID-19.

  Como xa se comentou, unha gran parte dos enfermos por COVID-19 son persoas de entre 50 e 80 anos. Na Figura 1.1 vese con claridade como a enfermidade máis grave se dá en idades avanzadas. Ademais, tamén é interesante comprobar a distribución da severidade por sexos, pois as manifestacións clínicas máis severas danse en homes cuxa idade na maioría supera os 65 anos e que normalmente presentan comorbilidades (Zheng, Z. et al., 2020). Esta estratificación da gravidade en canto ao sexo está confirmada en Takahashi e Iwasaki (2021), que explican o impacto do sexo na inmunidade debido a diversos factores como poden ser os estróxenos ou xenes presentes no cromosoma X.

  Na nosa mostra, máis do 50 % de asintomáticos e pacentes leves son mulleres, mentres que esta tendencia se reverte para as gravidades severa e crítica (Figura 1.3). Dadas estas notables diferencias, engadir estas variables no axuste dos modelos é imprescindible.

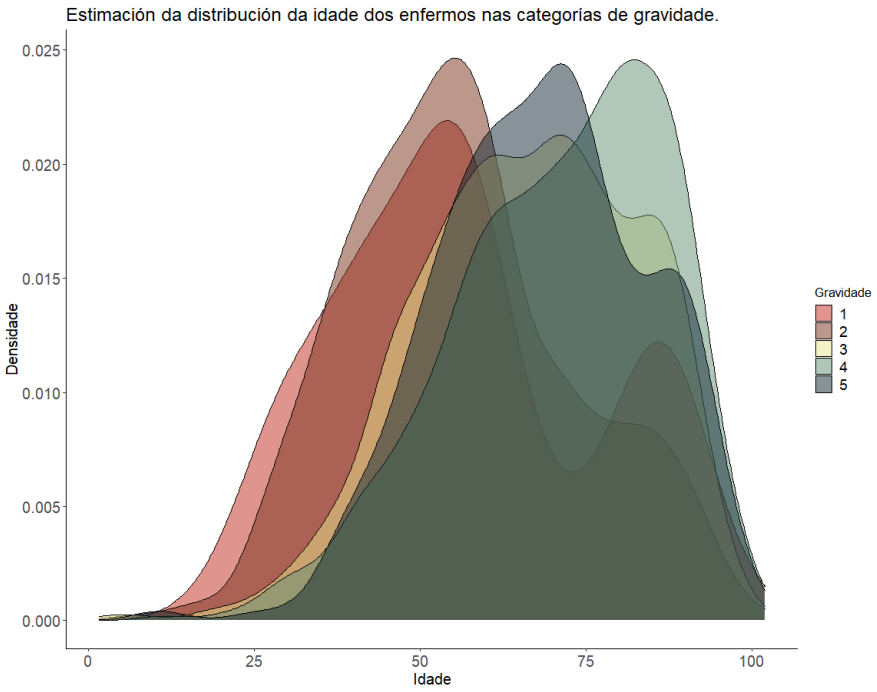

Figura 1.1: Estimación da distribución da idade nas 5 categorías de severidade (dende 1-asintomático a 5-crítico).

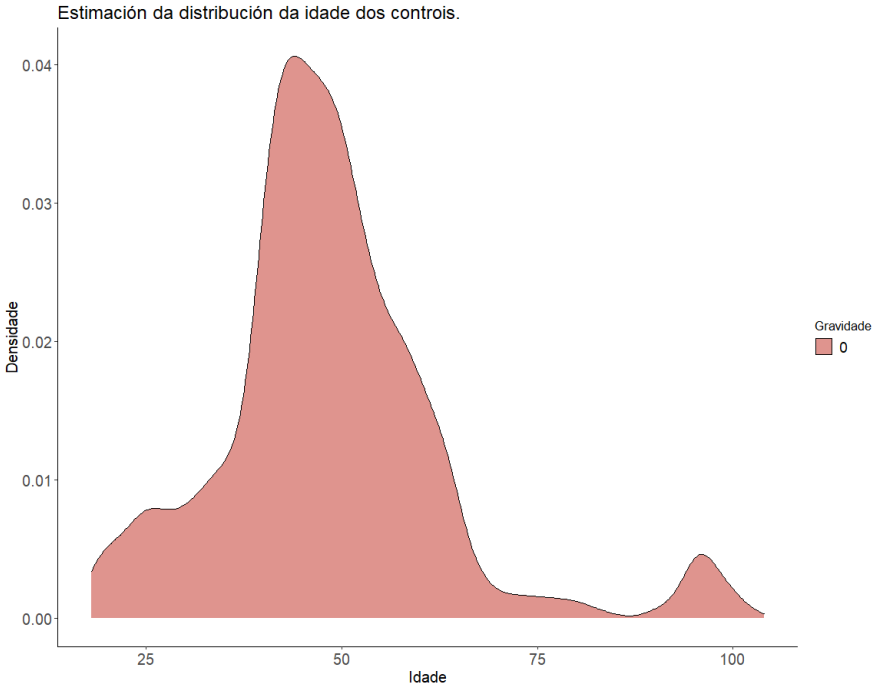

Figura 1.2: Estimación da distribución da idade nos controis.

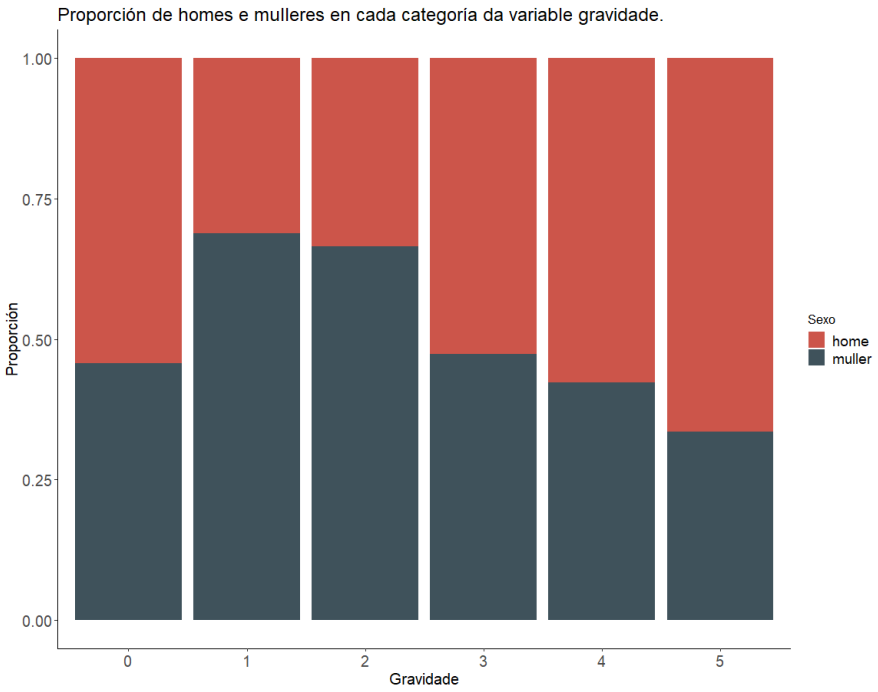

Figura 1.3: Proporcións de homes e mulleres en cada categoría de severidade.

- **Compoñentes principais**: Os enfermos da base de datos completa pertencen a varias poboacións ancestrais, reflictindo a variabilidade étnica dos individuos recrutados. Co fin de controlar a estratificación poboacional realizouse sobre as mostras xenotipadas unha análise de compoñentes principais e seleccionouse únicamente a ancestralidade europea. Este procedemento levouse a cabo co software `PLINK` (véxase Anexos A e B) sobre un conxunto de individuos formado polas mostras do proxecto *1000 Genomes*, para as cales se coñecen as poboacións e superpoboacións ás que pertencen, e as mostras do proxecto *SCOURGE*. Desta forma, os nosos individuos sitúanse no mapa de PCs ao redor das poboacións de 1000G e permítenos inferir as súas poboacións ancestrais. Na figura 1.4 están representadas as dúas primeiras compoñentes principais para a mostra completa (incluíndo o Banco Nacional de ADN e non europeos).

  A distribución dos individuos de 1000G nos dous eixos é moi clara, podendo distinguir 4 clústers moi diferenciados (correspondentes ás poboacións ancestrais africana, europea e as dúas asiáticas). A poboación americana é máis dispersa posto que inclúe mesturas de ancestralidades. En cor gris temos os nosos casos e os nosos controis. Os controis do BNA non se visualizan ao seren puramente europeos. Entre os casos do proxecto *SCOURGE* hai moitos individuos cunhas compoñentes ancestrais americana e asiática fortes, ademáis dalgún individuo que parece de orixe puramente africana. Como a análise principal se realizará sobre a mostra europea, quedámonos cos suxeitos que distan nas súas dúas primeiras PCs menos de tres desviacións típicas dos europeos de 1000G.

  A maiores, no propio continente europeo existe variación xenética e, aínda que a nosa

mostra está recollida en España, pode que queden residuos de variabilidade debido á ancestralidade. Isto solvéntase engadindo as compoñentes principais ao modelo. Segundo Price (2006), as dúas primeiras PCs resumen a variabilidade poboacional, e a terceira as diferencias na xestión das mostras nos diferentes laboratorios. Por convención, engádense as 10 primeiras PCs.

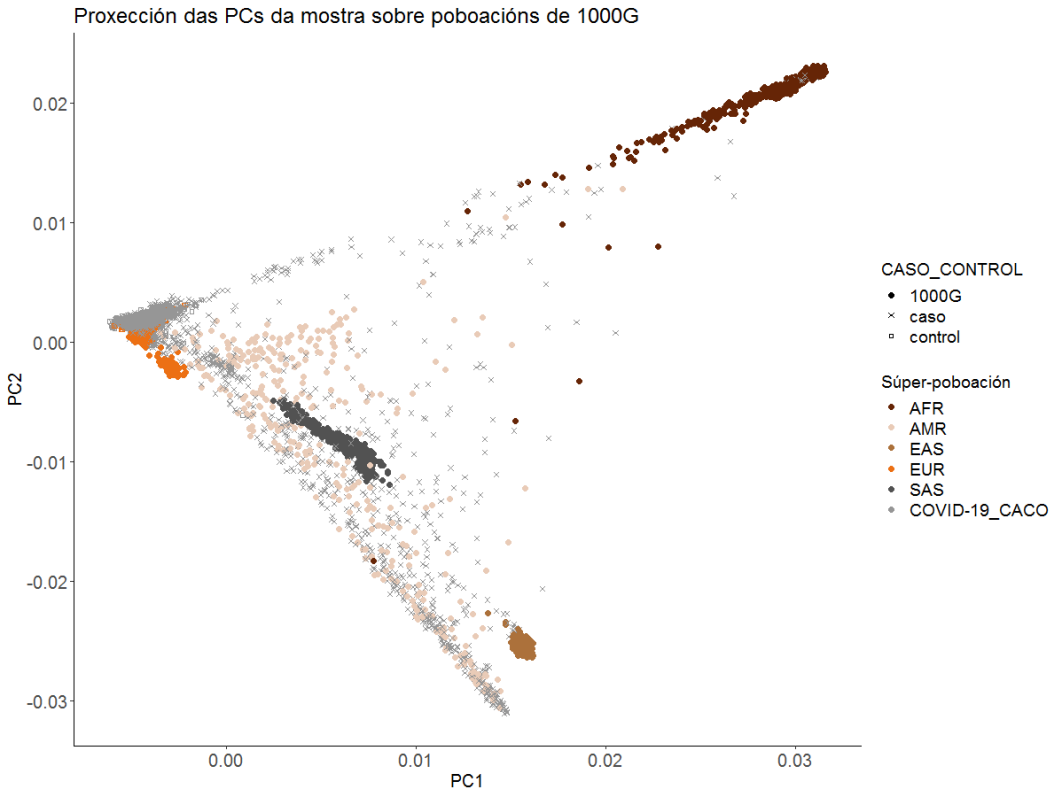

Figura 1.4: Gráfico das primeiras 2 compoñentes principais. Cun círculo están representadas as super-poboacións de 1000 Genomas (AFR-Africanos; AMR-Americanos; EAS-Asiáticos do este; EUR-Europeos; SAS-Asia do sur). Cunha cruz, os casos con COVID-19. Cun cadrado, os controis do BNA.

1.3. Obxectivos e organización do traballo

Ao longo deste traballo realizaranse varios estudos de asociación a partir dos datos xenéticos de dous conxuntos de persoas diagnosticadas por SARS-CoV-2. Algunhas das variables clínicas resumíronse nunha única variable de severidade, con 5 categorías que clasifican a persoas enfermas e unha categoría correspondente ao grupo control. A maioría dos individuos repítense para ambos conxuntos, mais no primeiro poderemos asumir independencia entre os datos e no segundo non. Isto levaranos ao primeiro obxectivo do traballo: comparar un modelo loxístico simple para o que se exclúen individuos (a práctica habitual nos estudos GWAS) e un modelo loxístico con efectos aleatorios que nos permita modelizar esta estrutura

de dependencia.

Nos Capítulos 2 e 3 describiranse a formulación e estimación destes modelos e tamén como se adaptan nos estudos de asociación. Como nun modelo loxístico a natureza da variable resposta é binaria, a avaliación dos mesmos será sobre doentes que presentaron un fenotipo crítico por COVID-19 en contraposición ao resto da poboación. O cromosoma que se usará en todos os estudos do traballo é o 3, pois xa se coñecen as sinais nunha das súas rexións e serviranos para validar os axustes.

Sen embargo, os GWAS sobre o COVID-19 levados a cabo ata o momento só contemplan a enfermidade máis extrema, e non o espectro de gravidade máis leve ou incluso sen síntomas. Non sería raro atopar unha variante asociada a presentar síntomas: a existencia dunha gran proporción de asintomáticos é unha das incógnitas máis resaltadas sobre o virus. O proxecto *SCOURGE* dispón dunha base de datos amplísima e rica en información clínica e bioquímica dos individuos recrutados, polo que se nos presenta unha oportunidade única para estudar a base xenética da enfermidade a todos os niveis. O procedemento típico nos casos nos que o fenotipo é múltiple é o de empregar varios modelos de regresión loxística simple, contrastando os individuos dunha categoría concreta co resto. Isto pode chegar a ter sentido se o número de sub-fenotipos non é demasiado elevado. A nosa variable de gravidade, pola contra, ten 5 categorías: veríamonos obrigados a axustar cinco modelos simples por cada SNP (como recordatorio, nun xenoma completo hai máis de 1 millón). Ademais, perderíamos información relevante relativa ás categorías intermedias. No Capítulo 4 aplicaremos unha regresión multinomial asumindo que cada nivel de gravidade é independente do outro. Compararemos os resultados cos modelos loxísticos, coa esperanza de atopar algunha sinal que se perdera nos axustes anteriores e que puidera estar asociado a unha enfermidade máis leve. Por suposto, este estudo tampouco se realizará sobre todos os cromosomas, e asumir que de existir esta variante estará asociada tamén ao cromosoma 3 non é moi realista.

Nas conclusións comentaranse outras posibles aproximacións e aportacións que se poderán realizar nun futuro tanto no proxecto como nunha ampliación da metodoloxía en GWAS. Finalmente engadíronse dous anexos, un referido a aos aspectos prácticos do traballo e un segundo no que se describe o software e librarías usadas: `PLINK`, `GenABEL`, `SAIGE`, `SAIGEgds`, `mlogit` e `mnlogit`. Destacar, tamén, que foi necesario o uso do ordenador Finis Terrae II do Centro de Supercomputación de Galicia (CESGA).

Capítulo 2

Modelo de regresión loxística

Nos estudos GWAS a variable resposta pode ser binaria ou cuantitativa, dependendo da natureza do carácter fenotípico a estudar. As dúas primeiras aproximacións que se realizarán neste traballo serán cunha variable Y de tipo dicotómica, que tomará o valor 1 cando o paciente presente infección severa por coronavirus e o valor 0 cando esta sexa leve ou asintomática. En función de se existe dependencia entre os datos como consecuencia do parentesco, axustarase un modelo de regresión simple ou un modelo mixto. Nesta sección describiremos o modelo loxístico simple e realizaremos un primeiro GWAS sen individuos emparentados. Desta forma, os individuos serán independentes entre si e non se violará ningunha asunción do modelo.[1]

2.1. Formulación do modelo

A regresión loxística estuda a relación entre a variable resposta Y de tipo binaria, que tomará valores 1 ou 0, e unha ou varias variables explicativas X. Y seguirá unha distribución de Bernouilli tal que

$$E(Y) = P(Y = 1) = p,$$

onde p é a probabilidade de éxito.

Inicialmente poderíamos representar a media de Y condicionada a un vector de variables explicativas X de forma lineal:

$$E(Y = 1|X = x_i) = x_i'\beta, \quad x_i \in \{1, \dots, n\}.$$

Sen embargo, este modelo incumpre as suposicións básicas. En primeiro lugar, a de linealidade, pois pode predecir valores fóra do soporte [0,1]; a de homocedasticidade, posto que a varianza depende da media condicionada e por tanto, da variable explicativa en cada caso, e a de normalidade.

[1] Os aspectos teóricos deste Capítulo están adaptados de Saavedra (2019).

É evidente a necesidade de construír un modelo distinto. A partir de aquí definirase $\pi(x) = P(Y = 1|X = x)$ como a probabilidade de éxito condicionada ao valor da variable explicativa. Para solucionar o problema do soporte de Y, aplícase a π unha transformación lineal mediante unha función link g que transforma o intervalo [0,1] en toda a recta real, de forma

$$g(\pi(x, \beta)) = x'\beta. \tag{2.1}$$

Da expresión en (2.1) deducimos que a probabilidade de éxito π dependerá do vector columna de valores de X e do vector de coeficientes asociado ás variables explicativas. A función link empregada habitualmente para variables dicotómicas, é a *logit* ou loxística. Defínese

$$g(p) = \log\left(\frac{\pi}{1-\pi}\right),$$

onde p será a probabilidade de éxito e $1-p$ a probabilidade de fracaso. Por tanto, é o logaritmo dun cocente coñecido como a Odds:

$$Odds(Y) = \frac{P(Y = 1)}{P(Y = 0)}. \tag{2.2}$$

O concepto formulado na ecuación (2.2) é unha forma diferente de reparametrizar a distribución de Bernouilli: indícanos canto máis probable é que se produza un suceso a que non se produza. A Odds, ao contrario que a probabilidade p, si pode tomar calqueira valor positivo. Se a Odds<1, a probabilidade de fracaso é maior que a de éxito; se a Odds=1, a probabilidade de éxito é de 0.5, e cando a Odds> 1, a probabilidade de éxito é maior que a de fracaso. Como vemos, a Odds pode tomar calqueira valor positivo ao contrario que a probabilidade de éxito, que só pode tomar valores no intervalo $[0, 1]$. Finalmente, aplicando un logaritmo á Odds obtemos unha transformación da variable resposta que tomará valores no intervalo $(-\infty, \infty)$ e poderemos formular o seguinte modelo lineal:

$$\log\left(\frac{\pi(x, \beta)}{1 - \pi(x, \beta)}\right) = \eta_{ij} = x'\beta.$$

Se invertimos a transformación $g(x)$, o modelo quedará expresado en función da probabilidade de éxito en lugar do logaritmo da Odds:

$$\pi(x, \beta) = g^{-1}(x'\beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}.$$

2.2. Interpretación dos parámetros

Unha vez definida a Odds, introducimos o concepto de Odds-ratio. Consideramos agora un modelo de regresión loxística cunha variable resposta binaria Y e unha variable explicativa X, que pode tomar valores 0 ou 1:

$$\pi(0, \beta) = P(Y = 1|X = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}};$$

$$\pi(1,\beta) = P(Y=1|X=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}},$$

onde β_0 é o intercepto e β_j o incremento de pasar do grupo de referencia ao grupo j. A Odds do grupo de referencia (grupo 0) é a exponencial do intercepto:

$$e^{\beta_0} = \frac{\pi(0,\beta)}{1-\pi(0,\beta)} = \frac{P(Y=1|X=0)}{P(Y=0|X=0)}.$$

Pola súa parte, $e^{\beta_0+\beta_1}$ será a Odds do grupo 1 (onde $X=1$). Neste punto cabería preguntarse que interpretación terían os parámetros β_j do modelo. Estes coeficientes son os logaritmos das Odds-Ratio (OR) para cada grupo. A OR defínese como o cociente da Odds nunha poboación e a poboación de referencia:

$$OR = \frac{Odds(Y|X=a)}{Odds(Y|X=b)} = \frac{\frac{P(Y=1|X=a)}{P(Y=0|X=a)}}{\frac{P(Y=1|X=b)}{P(Y=0|X=b)}}. \tag{2.3}$$

Podemos sustituír a Odds de cada grupo en (2.3) polas súas transformacións *logit*:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Temos, por tanto, que a Odds-Ratio de cada grupo, ou a exponencial do coeficiente β_j, é a cantidade pola que hai que multiplicar a Odds do grupo de referencia para obter a Odds do grupo j.

- Para $\beta_j > 0$, a súa exponencial dará unha OR maior que 1, increméntase a Odds do grupo j. Isto é, para o grupo j, a probabilidade de éxito é maior que no grupo de referencia.
- Para $\beta_j < 0$, a súa exponencial dará unha OR menor que 1. A probabilidade de éxito para o grupo j será menor que no grupo de referencia.

Cómpre incidir na importancia de que estas definicións son condicionadas nuns valores das covariables X. No caso dunha variable continua, a exponencial do coeficiente β_j será a OR correspondente a aumentar unha unidade na variable explicativa.

2.3. Estimación dos parámetros e contraste dos coeficientes

Os parámetros do modelo de regresión loxística estimaranse pois mediante o método de máxima verosimilitude. Dada unha mostra aleatoria simple $(X_1,Y_1),\dots,(X_n,Y_n)$, onde Y_i segue unha distribución de Bernouilli condicionada, con parámetros $(\pi(X_i,\beta))$. A función de máxima verosimilitude para o vector de coeficientes $\beta=(\beta_0,\beta_1,\dots,\beta_{p-1})$, supoñendo que os y_i son independentes, toma a forma:

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

Se sobre a función de máxima verosimilitude se aplican logaritmos, obtemos a chamada función de log-verosimilitude:

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

Posteriormente, derívase parcialmente con respecto a cada β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \pi(x_i, \beta)}{\partial \beta} \frac{y_i - \pi(x_i, \beta)}{\pi(x_i, \beta)(1 - \pi(x_i, \beta))}. \tag{2.4}$$

Como sabemos que, ademais, $\pi(x, |\beta) = \pi(x, \beta) = \frac{e^{x'\beta}}{1+e^{x'\beta}}$, derivamos tamén esta expresión con respecto a β e sustituímos o resultado en (2.4), quedando:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i'[y_i - \pi(x, \beta)]. \tag{2.5}$$

O parámetro β que maximizaría a ecuacion de verosimilitude obteríase igualando a ecuación (2.5) a 0. Sen embargo, para resolver esta ecuación hai que usar métodos iterativos como o de Newton-Raphson ou o de IRLS (iterative reweighted least square), usado no software R (R Core Team, 2020).

Unha vez estimados os parámetros β do modelo de regresión pode interesar realizar algún tipo de contraste sobre os mesmos. Os estimadores de máxima verosimilitude permítenos aproximar asintóticamente a distribución dos estimadores a unha distribución normal de forma

$$\hat{\beta} - \beta_0 \sim N_p(0, I(\hat{\beta})^{-1}).$$

$I(\hat{\beta})$ é unha aproximación da matriz de Fisher mediante a sustitución de β polo seu estimador de máxima verosimilitude. Un dos contrastes de hipóteses máis habitual é $H_0 : \beta_p = 0$ vs $H_1 : \beta_p \neq 0$. A continuación presentaranse o test de Wald, empleado na función **glm** de R (R Core Team, 2020) e no paquete `GenABEL` (ver Apéndice B), que será o que usaremos para os nosos datos.

$$Z = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \sim N(0, 1).$$

A inversa da estimación da matriz de Fisher, $I(\hat{\beta})^{-1}$ pode considerarse un estimador da matriz de varianzas-covarianzas de $\hat{\beta}$. A diagonal desta matriz contén as varianzas dos estimadores, polo que a súa raíz cadrada dará as desviacións típicas dos estimadores, que serán o denominador de Z. O valor crítico do contraste virá dado polo cuantil $(1 - \alpha)$ dunha normal estándar.

2.4. Regresión loxística no contexto dos GWAS

Os GWAS tipo caso-control analizan a susceptibilidade xenética dos individuos ao fenotipo Y, avaliando cada SNP por separado. Recordemos ademais que no modelo de herdanza

aditivo asúmese unha tendencia log-lineal do risco da enfermidade en función do aumento do número de alelos de menor frecuencia. É dicir, neste GWAS a variable causal de interese é o xenotipo do individuo i para o SNP_m, G_i, e codificarase en base ao número de copias do alelo raro a: tomará o valor 0 se G_i é homocigoto para o alelo frecuente (AA), 1 se G_i é heterocigoto (Aa) e 2 se G_i é homocigoto para o alelo raro (aa).

Así, sexa $SNP_1, \dots, SNP_M$ un conxunto de M SNPs, $m = 1, \dots, M$; Y unha variable dicotómica que toma valor 1 para os casos e 0 para os controis; $X_1, \dots, X_{p-1}$ un conxunto de $p-1$ covariables e N o número de individuos da mostra, $i = 1, \dots, N$, o modelo sobre Y é:

$$logit(Y_i) = \beta_0 + \beta_1 G_{im} + \beta_2 X_{2i} +, \dots, +\beta_{p-1} X_{p-1}. \tag{2.6}$$

Coa notación establecida en (2.5), $logit(Y_i)$ refírese ao *logit* da probabilidade $P(Y_i = 1|G_{im}, X_i)$, entendendo por $X_i = (X_{2i}, \dots, X_{(p-1)i})$. Ademais, β_1 será o efecto aditivo do SNP_m: o logaritmo da OR de aumentar unha copia do alelo raro no SNP m, é dicir, de aumentar unha unidade en G_m. A OR interpretarase, por tanto, como a Odds do alelo menos frecuente en relación á Odds do alelo máis frecuente. Cando a $OR < 1$, a probabilidade de ter a enfermidade será menor ao aumentar o número de copias do alelo menos frecuente (denominámolo a) para o SNP_m. Nestes casos fálase do alelo a como alelo protector. Se a $OR > 1$, a probabilidade de padecer a enfermidade será maior ao aumentar o número de copias do alelo pouco frecuente: o alelo a é un alelo de risco. Cando a $OR = 1$, non hai efecto do xenotipo sobre o fenotipo. Co contraste de hipóteses para o parámetro β asociado á variable SNP (mediante o test de Wald) determinaremos se dita variante é significativa e, por tanto, atopamos unha sinal no xen ao que pertence.

Recordemos que se axustarán M modelos de regresión e obteranse M p-valores asociados ao contraste de hipóteses $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, polo que nos atopamos ante un problema de tests múltiples. Poñámonos na situación de facer 100 contrastes de hipóteses cuxa hipótese nula se rexeita con probabilidade $\alpha = 0{,}05$: rexeitaranse 5 hipóteses nulas sendo certas. De feito, se consideramos cada test independente dos demais, a probabilidade de atopar polo menos un falso positivo é $P(\text{falsos positivos} \geq 1) = 1(1-\alpha)^m$. A isto coñéceselle como erro de tipo I. Agora ben, se temos en conta que nun estudo GWAS M é da orde de centos de miles, o número de variantes asociadas que poderían ser falsos positivos medra excesivamente. É necesario introducir unha corrección.

A práctica usual nesta área é situar o umbral de significación en 5×10^{-8}, baseado no *Family Wise Error rate (FWER)*. O FWER defínese como a probabilidade de cometer como mínimo un falso positivo. Típicamente o FWER contrólase mediante a corrección de Bonferroni, que consiste en dividir α por m. O umbral 5×10^{-8} vén de considerar 1 millón de SNPs e, por tanto, 1 millón de contrastes. Sen embargo, estudos de simulación demostran que esta aproximación leva a unha perda de potencia estatística salvo para estudos con N elevada, o que ten como consecuencia detectar menos SNPs asociados ou con efectos moi pequenos (Otani, 2018).

Outra posibilidade é a de controlar o *False Discovery Rate (FDR)*, a proporción de falsos positivos con relación aos verdadeiros positivos. Defínese como a esperanza de Q, $Q_e = E(Q) = E(V/V+S) = E(V/R)$, onde V é a proporción de falsos positivos (hipóteses

nulas rexeitadas a pesar de ser verdadeiras) e S a proporción de verdadeiros positivos (hipóteses nulas rexeitadas sendo falsas); o que equivalería a controlar a proporción de falsos positivos (Benjamini e Hochberg, 1993). O FWER sería $P(V \geq 1)$. Na literatura hai propostos numerosos procedementos para controlar o FDR, mais empregaremos o de Benjamini-Hochberg:

Sexan $p_1, \dots, p_m$ os p-valores correspondentes aos M-contrastes de hipóteses e sexan $p_{(1)} \leq , \dots, \leq p_{(M)}$ os p-valores ordenados, k será o i máis alto para o cal

$$p_{(i)} \leq \frac{i}{M} q^*.$$

Rexeitaríamos $H_{(i)}$, sendo $i = 1, \dots, k$. Desta forma, baixo $m = 100$ e fixando q^* a 0.05 (é dicir, asumindo 5 falsos positivos) compararíamos secuencialmente cada p-valor ata chegar a k, e rexeitaríamos as k hipóteses nulas (Benjamini e Hochberg, 1993). Porén, este criterio asume que as m hipóteses son independentes entre si, algo que non é real nos estudos GWAS debido ao LD. Otani et al. (2018) detectaron unha diminuición da potencia estatística baixo simulacións realistas, isto é, tendo en conta o desequilibrio de ligamento, aínda que o FDR continou a ser o procedemento con mellores resultados.

Traballaremos ao longo do traballo co cromosoma 3, que ten ao redor de 40.000 SNPs. Aplicando a corrección de Bonferroni, quedaríamonos cun umbral de $1{,}25 \times 10^{-6}$.

2.5. Aplicación aos datos do COVID-19

Na presente sección analizouse unha única variable *Severidade_4*. Deste xeito, os controis serán todos os individuos que tomen o valor 0 para esta variable; é dicir, que presenten un fenotipo distinto ao crítico. Os casos tomarán o valor 1 e cumprirán os criterios mencionados na Sección 1.3 para gravidade crítica. Estaremos analizando a base xenética do fenotipo máis extremo do coronavirus en contraposición ao resto da poboación. A ratio 0/1 é de 9154/742. Inicialmente hai un total de 9896 individuos que, teóricamente, non están relacionados entre si. Cómpre recordar que no axuste de cada SNP o N non será de 9896, posto que para algunha persoa pode faltar o xenotipo do SNP en concreto. A libería usada é `GenABEL` (ver Apéndice B). Previamente a axustar o modelo de regresión, realizouse un control de calidade sobre todo o xenoma co propio paquete de `GenABEL` (por SNP e por individuo), quedando un total de 589633/714284 marcadores e 9841/9896 individuos. Do total de SNPs usáronse só os do cromosoma 3. Axustáronse dous modelos loxísticos, un sen ningunha covariable e outro con covariables para controlar os efectos ambientais: *sexo*, *idade*, e as 10 primeiras PCs.

O gráfico típico para presentar os resultados dun GWAS é o Manhattan plot: é un diagrama de puntos onde cada un deles representa un SNP, cuxo eixo X mostra a posición de cada marcador no cromosoma e o eixo Y é o $-\log_{10}(p)$ asociado a dito SNP. Este gráfico permítenos visualizar fácilmente a posible asociación dunha rexión do xenoma co fenotipo estudado. Habitualmente engádense dúas rectas horizontais para $-\log_{10}(p = 1e^{-5})$, a partir da cal os SNPs considéranse *suxestivos*, e para $-\log_{10}(p = 5e^{-8})$, o nivel de significación habitual. O Miami-plot é outro gráfico que cada vez se está volvendo máis popular na literatura. Compara

dous Manhattan-plots; por exemplo, de dous estudos diferentes, ou incluso dun mesmo estudo para contrastar SNPs con OR maior e menores que 1.

Consideraríanse significativos os SNPs que obtiveron un p-valor menor a 5×10^{-8} segundo o umbral habitual e menor a $1{,}25 \times 10^{-6}$ pola corrección de Bonferroni. Nos Cadros 2.1 e 2.2 móstranse os resultados dos modelos de regresión para os SNPs cuxo p-valor está por debaixo de 10^{-5}. O único SNP por debaixo de 5×10^{-8} e de $1{,}25 \times 10^{-6}$ en ambos modelos foi o SNP rs7135088, que pertencen ao xen LZTFL1. A OR para este marcador indícanos que, ao aumentar unha copia do alelo menos frecuente, a Odds de padecer un cadro clínico crítico de COVID-19 é dúas veces maior que a dun individuo que non posúa ningunha copia. Típicamente nun GWAS, os efectos dos SNPs son moi limitados, e de aí a dificultade para detectar sinais (sobre todo en variantes moi raras). Porén, que a OR estimada sexa cercana a 2 quere dicir que, en efecto, este SNP constitúe un biomarcador importante na base xenética de enfermidade por COVID-19.

Axuste sen covariables							
SNP	**Crom.**	**Pos.**	**Xen**	**N**	**P-val.**	**OR**	**IC**
rs71325088	3	45862952	LZTFL1	9830	1.15E-10	1.772	1.489-2.109
rs75928798	3	45962603	FYCO1	9829	3.99E-06	1.461	1.244-1.717
rs1994491	3	45960420	FYCO1	9830	5.51E-06	1.443	1.232-1.69
rs1994492	3	45960646	FYCO1	9830	6.59E-06	1.451	1.234-1.706
rs1994493	3	45960700	FYCO1	9827	7.07E-06	1.449	1.233-1.704
rs13079478	3	46007823	FYCO1	9810	7.72E-06	1.448	1.231-1.703
rs13097556	3	46061997	XCR1	9818	9.63E-06	1.429	1.22-1.674
rs13079869	3	46008087	FYCO1	9822	1.02E-05	1.441	1.225-1.694
rs33910087	3	46009487	FYCO1	9825	1.11E-05	1.439	1.223-1.692
rs2230322	3	46063329	XCR1	9785	1.40E-05	1.420	1.212-1.664
rs71327023	3	46131225	XCR1	9833	2.47E-05	1.419	1.206-1.669

Cadro 2.1: Táboa dos SNPs cuxo p-valor$< 1 \times 10^{-5}$ para o axuste sen covariables do modelo loxístico.

A maiores, ilústranse nos Cadros 2.1 e 2.2 outros SNPs que non superan o umbral pero pertencen todos á mesma rexión do cromosoma 3, suxerindo así un sinal forte asociado á mesma. Este sinal, de feito, pódese ver no Miami-plot (Figura 2.1), en verde.

Axuste con covariables							
SNP	**Crom.**	**Pos.**	**Xen**	**N**	**P-val.**	**OR**	**IC**
rs71325088	3	45862952	LZTFL1	8646	1.22E-10	1.830	1.523-2.2
rs1994491	3	45960420	FYCO1	8645	4.88E-06	1.470	1.246-1.734
rs75928798	3	45962603	FYCO1	8644	5.22E-06	1.480	1.25-1.752
rs1994492	3	45960646	FYCO1	8646	7.52E-06	1.473	1.243-1.745
rs1994493	3	45960700	FYCO1	8642	7.76E-06	1.472	1.242-1.743
rs13079478	3	46007823	FYCO1	8624	9.84E-06	1.467	1.238-1.739
rs13079869	3	46008087	FYCO1	8641	1.13E-05	1.463	1.234-1.734
rs13097556	3	46061997	XCR1	8635	1.20E-05	1.448	1.227-1.709
rs33910087	3	46009487	FYCO1	8640	1.34E-05	1.458	1.231-1.729

Cadro 2.2: Táboa dos SNPs cuxo p-valor$< 10^{-5}$ para o axuste con covariables *Sexo*, *Idade* e as 10 primeiras PCs, do modelo loxístico.

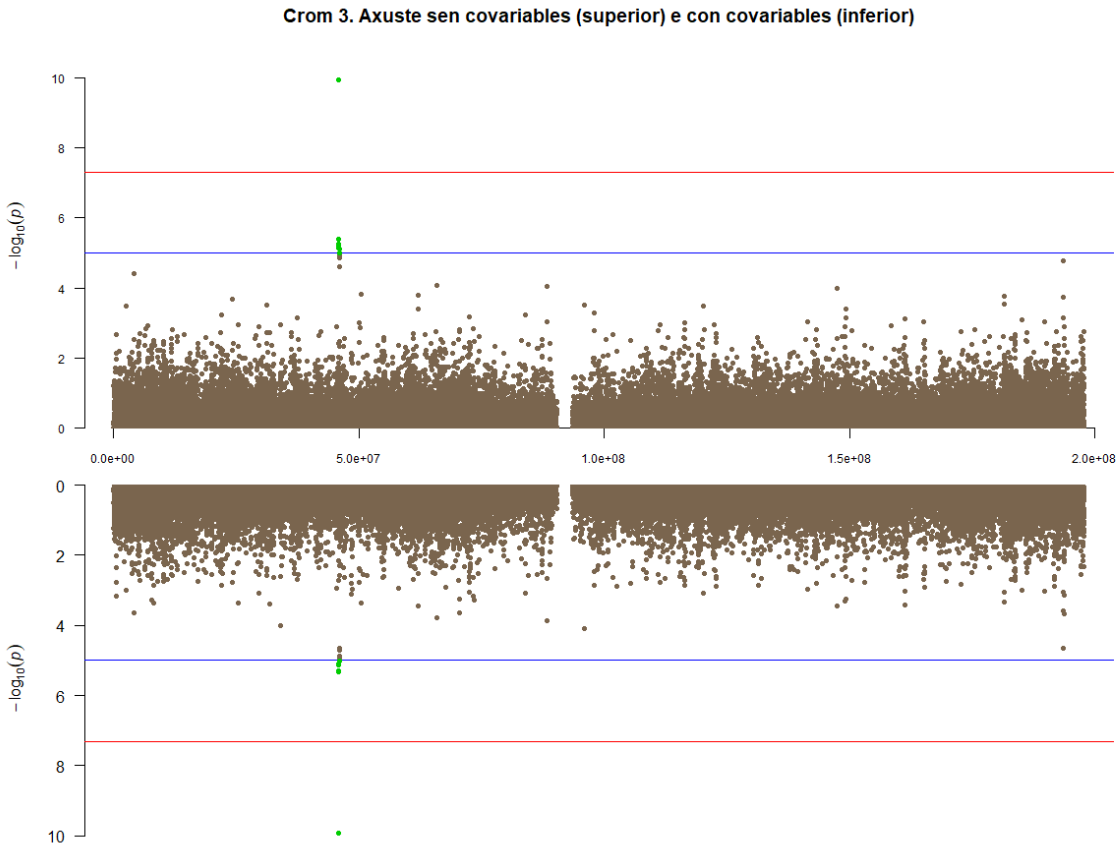

Figura 2.1: Miami plot do cromosoma 3 con e sen covariables co modelo loxístico simple. No eixo Y, $-\log_{10}$(p-valor). No eixo X, posicións no cromosoma.

Aínda non sendo significativos, podemos comprobar que no modelo con covariables os p-valores asociados aos marcadores son algo máis altos que no modelo sen covariables. De feito, os SNPs rs22330322 e rs331910087 superaron o umbral de 1×10^{-5} no primeiro axuste pero non para o segundo. A razón detrás disto é a corrección feita polas covariables *sexo* e *idade*, e as OR asociadas ao SNP deberán interpretarse unha vez fixadas as dúas.

Con todo, a pesar de ter retirado os parentes e engadido as compoñentes principais pode quedar estrutura xenética residual. Elimináronse as relacións de parentesco ata o terceiro grao, pero podería seguir existindo dependencia entre algúns individuos. Ademais, nos estudos GWAS nos que a N é relativamente baixa descartar individuos resulta nunha perda de información importante. É evidente a necesidade de aplicar novos modelos, en concreto os modelos mixtos que trataremos no capítulo seguinte.

Capítulo 3

Modelos de regresión mixtos

3.1. Motivación

É moi frecuente a aparición de xerarquías nos datos en campos como a bioloxía, polo que se precisan modelos que teñan en conta esta estrutura. Nestes modelos, que se denominan modelos mixtos ou modelos multinivel, os individuos están agrupados nun nivel superior, que á súa vez tamén o pode estar. Os individuos dun mesmo grupo terán maior similitude entre si que os individuos de diferentes grupos. Como exemplo deste tipo de datos temos os estudos de medidas repetidas, onde os grupos serían os propios individuos, ou os estudos de reproducción, onde os grupos son as linaxes.

Nun modelo mixto asúmese que o número de grupos é ilimitado e que os empregados no estudo son unha mostra da poboación. Así, non será de interese analizar estes grupos por separado senon estimar o *efecto grupo*. Ten sentido pensar que nun mesmo grupo existirá unha correlación entre os suxeitos que o compoñen, polo que xa non se poderá asumir independencia entre as observacións. Distinguiremos, ademais, uns efectos fixos e uns efectos aleatorios. Os efectos fixos son aqueles nos que recae o interese do experimento e cuxos niveis constitúen toda a poboación. Por outro lado, se os niveis dunha variable se consideran unha mostra da poboación, falamos de efectos aleatorios.

Supoñamos que temos datos da porcentaxe anual de persoas recuperadas na UCI de varios hospitais e queremos comprobar se o número de sanitarios traballando neles inflúe dalgunha forma sobre o número de doentes que reciben a alta. Estes hospitais, ademais, pertencen a diferentes cidades. Xa que o obxectivo é analizar o impacto do número de sanitarios sobre a porcentaxe media de persoas recuperadas, este será o efecto fixo. Por outro lado, as cidades escollidas non constituirán o conxunto de cidades do país, serán so unha mostra do total. Sen embargo, poderíamos supoñer que a calidade de vida nunha cidade (medida a través da renta media, por exemplo) ou incluso a densidade de poboación, afectarán tanto ao número de persoas que ingresan na UCI como ao número de persoas que se recuperan. Neste sentido, que o hospital se sitúe nunha cidade determinada será o efecto aleatorio, xa que só nos importa ter en conta a variabilide que existe na porcentaxe de recuperación media nun hospital en

función da cidade á que pertenza e non estudar que ocorre en cada unha específicamente. Así, a porcentaxe de recuperación media en cada cidade variará con respecto á porcentaxe global nunha cantidade aleatoria (efecto grupo), que haberá que predecir. Cómpre sinalar que definir cal é o efecto fixo e cal é aleatorio depende do experimento en cuestión.

Henderson (1984) propuxo empregar estes modelos para predecir se certo animal presentará un fenotipo ou característica de interese en función da xenética, tendo en conta os efectos aleatorios das proxenies ou relacións endogámicas. Introduciu o concepto de matriz de relacións nos modelos de regresión e foi o primeiro en propoñer un método eficiente para invertir a matriz. Seguindo a súa idea principal, nos estudos GWAS tamén se fai unha predicción da variable dependente -a enfermidade- a partir da información xenética de miles de individuos. Esta información xenética, como sabemos, dase a partir do xenotipo nos SNPs seleccionados. Estas personas poden presentar relacións de parentesco entre si que se manifestarán na súa xenética e por tanto ten sentido pensar en termos de modelos multinivel, onde cada familia constitúe un grupo. Dende un punto de vista biomédico, e sobre todo no contexto da epidemia do COVID-19, poden darse diversas situacións relacionadas co recrutamento de pacientes que acentúen este problema. Por exemplo, é probable que todos os integrantes dun núcleo familiar acudan ao mesmo hospital unha vez infectados e pode que máis de un sexa seleccionado para participar no estudo. Así, teremos datos de persoas emparentadas cuxa relación é difícil de trazar na práctica, polo que é imprescindible incorporar este efecto no modelo matemático. [1]

3.1.1. Modelos lineais mixtos

En primeiro lugar introduciremos brevemente o modelo de regresión lineal mixto (LMM). Dada unha variable Y continua e unha variable X, o modelo de regresión lineal clásico de X sobre Y pódese escribir como $E(Y) = X\beta$. O parámetro β é un efecto fixo. Incorporaremos agora o efecto aleatorio: Zu, onde Z é a matriz dos q efectos aleatorios $n \times q$ e u un vector $q \times 1$. ϵ será un vector $n \times 1$. O LMM é da forma:

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

$$u \sim N(0, G); \epsilon \sim N(0, R).$$

G, a matriz de varianzas-covarianzas de u, é unha matriz simétrica. R é a matriz de varianzas-covarianzas de ϵ, con todos os elementos da diagonal iguais e o resto dos elementos iguais a 0, polo que $Var(\epsilon) = I\sigma^2_\epsilon$, onde I é a matriz identidade. Podemos escribir o modelo en función dos efectos aleatorios e dos fixos como

$$Y = X\beta + \epsilon^*, \tag{3.1}$$

sendo $\epsilon^* = Zu+\epsilon$. Da expresión (3.1) podemos recuperar a estrutura de varianzas-covarianzas do modelo:

$$Cov(\epsilon^*, \epsilon^*) = Cov(Zu + \epsilon, Zu + \epsilon) = Z'Cov(u, u)Z + R.$$

$$Var(Y) = ZGZ' + R, \quad E(Y) = X\beta.$$

[1] Os aspectos teóricos das Seccións 3.1.1 e 3.2. están adaptados de Crujeiras e Conde (2020).

O LMM asume que a distribución da variable resposta segue unha distribución normal. Sen embargo, cando a variable Y é discreta ou segue unha distribución particular, haberá que empregar os modelos lineais xeralizados engadindo un efecto aleatorio. Para o caso dunha variable resposta dicotomizada, como ocorre no noso caso, o modelo será o loxístico.

3.1.2. Modelo dos compoñentes da varianza en xenética cuantitativa

A partir dos modelos polixénicos de Fisher é inmediato visualizar a formulación das compoñentes da varianza e, por tanto, dun modelo mixto. A variabilidade dun fenotipo cuantitativo de interese pódese descompoñer en termos da varianza xenética (V_G) e ambiental (V_E):

$$V_P = V_G + V_E.$$

Tanto os efectos xenéticos como os efectos ambientais son variables aleatorias, que denotaremos como u e ϵ, respectivamente. A V_G está á súa vez descomposta en varianza aditiva, varianza dominante e varianza epistática, pero por simplificación do modelo centrarémonos na varianza aditiva. A varianza aditiva é a única compoñente que se pode estimar directamente a partir das estimacións da mostra e é a causante das similitudes entre parentes. De feito, o cociente V_A/V_P denomínase herdabilidade e interprétase como a porcentaxe de variabilidade do fenotipo que se atribúe á herdanza dos caracteres (Falconer, 1996). A varianza ambiental, V_E, recolle toda a variabilidade do fenotipo que non é xenética. Asúmese independencia entre V_E e V_G. A primeira ecuación poderemos reescribila como (Abney et al., 2000; Kang et al., 2010):

$$Var(Y) = 2\sigma_u^2 K + \sigma_\epsilon^2 I.$$

$Var(Y)$ é un abuso de notación, pois en realidade temos unha matriz de varianzas-covarianzas. A varianza dos efectos ambientais é σ_ϵ^2, e I a matriz identidade. A varianza aditiva xenética, σ_u^2. K será unha matriz de coeficientes de parentesco para cada par de individuos.

Os efectos ambientais capturados polos ϵ_i, ademais, asúmense independentes e distribuídos normalmente, con media cero. Serán o responsables das diferencias entre individuos dentro dunha mesma familia. Cando non existe unha relación de parentesco, as similitudes entre o fenotipo dos individuos non estarán propiciadas por cuestións de herdanza: a varianza aditiva será nula e V_G desaparece da ecuación. Deste xeito é posible aplicar un modelo de regresión simple para Y, no que só se contemplan efectos ambientais:

$$Y = \beta_0 + \beta_1 G_m + \epsilon, \quad Var(Y) = \sigma_\epsilon^2.$$

Sen embargo, isto non é moi realista, pois case sempre vai existir algún tipo de estrutura na mostra. Será necesario estimar V_G. Incluímos así un efecto aleatorio xenético u, cuxa varianza é a proposta anteriormente, $Var(u) = \sigma_u^2 K$. Afortunadamente, coñecida a matriz K, poderase estimar σ_u^2. K será a matriz de varianzas-covarianzas dos xenotipos.

$$Y = \beta_0 + \beta_1 G_m + u + \epsilon, \quad Var(Y) = \sigma_u^2 K + \sigma_\epsilon^2. \tag{3.2}$$

A ecuación en (3.2) é a formulación dun modelo mixto lineal con efectos fixos β_0 e β_1 e parte aleatoria dada por $(u + \epsilon)$.

3.2. Modelos mixtos lineais xeneralizados

O modelo mixto lineal pódese estender a variables respostas que sigan distribucións diferentes á normal, ao igual que ocorre cos modelos de regresión lineais simples. Recordemos que o noso estudo é de tipo caso-control, polo que debemos adaptar o explicado nas Seccións 3.1.1 e 3.1.2 aos modelos lineais xeneralizados (GLMM), en concreto coa transformación *logit*. Variaremos lixeiramente a notación con respecto ao LMM previamente descrito.

No modelo loxístico mixto, os individuos (nivel 1) estarán agrupados nunha estrutura de nivel 2. Por tanto, haberá N individuos e J grupos, e asumiremos un posible efecto aleatorio asociado ao grupo. A transformación que se realiza sobre o preditor lineal é análoga á dos modelos lineais simples e será a través dunha función link, a logit, establecéndose tamén sobre a log-odds:

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

π_{ij} é a probabilidade de $Y = 1$ para o individuo i condicionada aos efectos aleatorios: $\pi_{ij} = E(Y_{ij}|u_{0j})$. Desta forma, o preditor lineal máis sinxelo formularase como

$$\eta_{ij} = \beta_0 + u_{0j}.$$

Os efectos aleatorios asociados ao grupo u_{0j} seguirán unha distribución normal de media 0 e varianza σ_u^2, sendo independentes entre si. β_0 será a log-odds dun grupo medio. Finalmente, o modelo loxístico mixto escribirase como:

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + u_{0j}.$$

$$\pi_{ij} = P(Y_{ij} = 1|u_{0j}), \quad u_{0j} \sim N(0, \sigma_u^2).$$

Neste punto incluiremos variables explicativas. Estas poden ser de primeiro ou de segundo nivel, dependendo de se están asociadas aos individuos ou aos grupos, respectivamente. Se consideramos únicamente variables de nivel 1, podemos axustar ou ben un modelo con intercepto aleatorio ou ben un modelo con intercepto e pendente aleatorias.

- **Modelo con intercepto aleatorio.** Todos os grupos terán a mesma pendente, pois será un parámetro fixo β_1 e constante en todos os grupos. As rectas asociadas a cada j serán paralelas pero terán interceptos distintos, $\beta_0 + u_j$.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 X_{ij} + u_{0j}.$$

$$\pi_{ij} = P(Y_{ij} = 1|X_{ij}, u_{0j}).$$

- **Modelo con intercepto e pendente aleatorias.** A variable explicativa X terá un efecto diferente en cada grupo, permitindo así que cambie a pendente entre eles.

$$\log\left(\frac{\pi_{ij}}{1-\pi ij}\right) = \beta_0 + \beta_1 X_{ij} + u_{1j} X_{ij} + u_{0j}.$$

$$\pi_{ij} = P(Y_{ij} = 1|X_{ij}, u_{0j}, u_{1j}).$$

Introducimos así un efecto aleatorio asociado á pendente, u_{1j}. Por tanto, consideraremos variables aleatorias asociadas tanto a β_0 como a β_1, independentes dos erros, que seguirán unha distribución normal de media 0 e matriz de varianzas-covarianzas dos u.

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

A interpretación dos parámetros é análoga ao modelo loxístico simple, no que falamos de Odds e Odds-ratio. O intercepto global aleatorio β_0 é o logaritmo da Odds para $Y = 1$, $X = 0$ e $u = 0$ na media dos grupos (os valores das covariables son 0 e non existe efecto aleatorio). Se queremos saber cal é o intercepto no grupo J, sumamos ao valor de β_0 o efecto aleatorio asociado u_{0j} e obtemos o valor do logaritmo da Odds para ese grupo. O valor dos β_i asociados ás covariables X continuas indicarán canto cambia o logaritmo da Odds para $Y = 1$ ao aumentar unha unidade de X, fixado sempre o efecto aleatorio do grupo (estaremos medindo o efecto intra-grupo).

Finalmente, comentar que as probabilidades π_{ij} obtéñense a partir de prediccións. Supoñendo que estamos traballando cun modelo con intercepto aleatorio, serían:

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

Hai propostas varias estratexias para calcular estas probabilidades, como por exemplo fixar o efecto aleatorio a 0 (estimar a probabilidade nun grupo medio), ou predicir u_{0j}.

Volvendo aos datos do COVID-19, o modelo que se axustará será un modelo cun efecto aleatorio asociado ao parentesco, e detallarase na sección seguinte. Resumidamente, π_{ij} será a probabilidade para o individuo i de padecer un cadro clínico grave da enfermidade, sempre condicionada aos valores das covariables X e aos efectos aleatorios asociados a pertencer a unha familia. j Como no modelo loxístico, o β asociado a cada SNP medirá o cambio no logaritmo da Odds ao aumentar unha unidade do alelo menos frecuente para individuos dunha mesma familia. Hai que apuntar que é difícil definir o grupo neste caso, pois o concepto de núcleo familiar é moi difuso. Como consecuencia, non coñeceremos o número de grupos e, por tanto, tampouco nos interesará predecir os efectos aleatorios nin engadir variables de segundo nivel.

3.2.1. Inferencia sobre os parámetros

Os estimadores dos parámetros dos efectos fixos e aleatorios, nun modelo lineal mixto, obtéñense a partir de técnicas de máxima verosimilitude como REML ou ML.

Consideremos, por simplificar, un modelo sen covariables e só con intercepto aleatorio.

$$Y_{ij}|u_{0j} \sim Ber(\pi_{ij}), \quad \pi_{ij} = P(Y_{ij} = 1|u_{0j}).$$

A verosimilitude condicional para o grupo j, baixo a suposición de que dados os efectos aleatorios existe independencia entre dúas observacións do mesmo grupo, escríbese:

$$L_j(\beta_0|u_{0j}) = \prod_{i=1}^{n_j} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}.$$

Sen condicionar polo efecto aleatorio podemos calcular a verosimilitude marxinal no grupo j:

$$L_j(\beta_0, \sigma^2_{u0}) = \int L_j(\beta_0|u_{0j})\phi(u_{0j}; 0, \sigma^2_{u0})du_{0j},$$

onde $\phi(u_{0j}; 0, \sigma^2_{u0})$ é a función de densidade dos efectos aleatorios, unha normal con media 0 e varianza $\sigma^2_{u_0}$. Haberá que integrar con respecto a esta densidade, polo que se requiren técnicas de aproximación como a cuasi-verosimilitude.

Finalmente, con ambas ecuacións podemos calcular a verosimilitude global do modelo:

$$L(\beta_0, \sigma^2_{u0}) = \prod_{j=1}^{J} L_j(\beta_0, \sigma^2_{u0}). \tag{3.3}$$

Maximizando (3.3) obtemos os estimadores da varianza dos efectos aleatorios e os estimadores dos efectos fijos. Como non existe solución explícita para esta ecuación empréganse métodos numéricos.

3.3. Formulación no contexto xenético

As expresións das compoñentes da varianza que vimos na Sección 3.2.1 están formuladas para caracteres cuantitativos, mais se poden estender aos caracteres binarios. Definiremos o modelo proposto por Zhou et al. (2018).

O fenotipo Y seguirá unha distribución binomial con $E(Y|u) = \pi_i$ e varianza $Var(Y|u) = \phi v(\pi_i)$, sendo ϕ o parámetro de dispersión; $v(\pi_i) = \pi_i(1-\pi_i)$ e G_m, o xenotipo para o SNP m:

$$\log(\pi_i) = \alpha X_i + \beta G_{mi} + u_i, \quad i = 1, \ldots, N.$$

A probabilidade de presentar o fenotipo de interese, π_i, é $P(y_i = 1|X_i, G_{mi}, u_i)$. X será un vector $1 \times P$, sendo P o número de covariables, cos valores de X asociados ao individuo i. O vector de efectos fixos α será un vector columna $P \times 1$ que incluirá o intercepto e u será un vector columna $n \times 1$ de efectos aleatorios. O efecto fixo asociado ao xenotipo G_{mi} é o parámetro β. Os efectos aleatorios u seguirán unha distribución normal de media 0 e matriz de varianzas-covarianzas

$$Var(u) = \Sigma_{k=1}^{K} \tau_k \psi.$$

τ é a varianza aditiva xenética e ψ, a GRM (*Genetic relationship matrix*). Esta matriz, que será a matriz K, medirá a distancia xenética entre individuos:

$$\psi = \frac{G_C^T G_C}{M1},$$

para a cal G_C é a matriz $M_1 \times N$ dos xenotipos normalizados (a través da media e desviación típica dos xenotipos *crudos*) para M_1 marcadores. Co obxectivo de reducir o tempo computacional, no paquete `SAIGE` (ver Apéndice B) empregan a chamada *low-rank GRM*, calculada a partir de M_1 marcadores independentes (aquí independencia implica ausencia de desequilibrio de ligamento). O parámetro de dispersión fixarase a 1 e non estimaremos o intercepto, pois a súa interpretación carece de valor.

3.3.1. Estimación dos parámetros

A estimación dos parámetros (α, ϕ, τ) obtense a partir da formulación da cuasi-verosimilitude, que se calculará baixo a hipótese nula de $H_0 : \beta = 0$. Porén, fixaremos o parámetro de dispersión $\phi = 1$.

A cuasi-verosimilitude (QL) pódese empregar en situacións onde se coñece unha relación entre a media e a varianza das observacións (Wedderburn, 1974). Supoñamos que temos unhas observacións $z_i, i = 1, \dots, n$; a cuasi-verosimilitude defínese como

$$qL(z_i, \pi_i) = \int_{y_i}^{t} \frac{z_i - t}{V(t)} dt.$$

π_i denotará á media e $Var(\pi_i)$ a varianza; sendo V unha función coñecida. No noso modelo, para o cal $\pi_i = E(Y_i|u)$ e $Var(Y_i|u) = \phi v(\pi_i)$, a cuasi-verosimilitude de α e $\beta = 0$ con respecto a π condicionada aos efectos aleatorios quedaría formulada do xeito:

$$qL(\alpha, \beta = 0|b) = \int_{y_i}^{t} \frac{\alpha_i(y_i - t)}{v(t)} dt,$$

xa que $\phi = 1$. A verosimilitude integrada do modelo, por outro lado, seguirá a expresión:

$$L(\alpha, \phi = 1, \tau) = \int L(\beta_0|u) f(u) du. \tag{3.4}$$

Agora, se asumimos unha distribución normal con media 0 e varianza (τ, ψ) nos efectos aleatorios u e sustituímos $L(\beta_0|u)$ pola súa cuasi-verosimilitude, poderíamos reescribir a ecuación (3.3):

$$\log L(\alpha, \beta = 0, \tau) = \log \int \exp\left\{\sum_{i=1}^{N} qL_i(\alpha, \beta = 0|u)\right\} (2\pi)^{-\frac{N}{2}} |\tau\psi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} u^T (\tau\psi)^{-1} u\right\} du. \tag{3.5}$$

A integral en (3.4) é aproximada polo método de Laplace. O procedemento iterativo para a estimación dos parámetros empregado no paquete `SAIGE` é AI-REML, un algoritmo específico

para estimar as compoñentes da varianza de xeito computacionalmente eficiente (Gilmour, 1995). Este método estima os parámetros mediante máxima verosimilitude restrinxida pero matriz de información esperada é sustituída por unha matriz de información ponderada (*AI, average information matrix*).

Xa sabemos que un dos puntos básicos dos GWAS é o contraste dos coeficientes β_{SNP}. Agora ben, é común que os casos e os controis non estean equilibrados en canto a número nos estudos GWAS. Está demostrado que os tests de Wald, ratio de verosimilitudes e *score test* funcionan ben e son equivalentes ao traballar con $MAF \geq 0{,}05$, mais soen inflar o erro de tipo I cando a frecuencia do alelo menor é máis baixa que 0.05 (Dey et al., 2017). Co obxectivo de corrixir esta inflación, desenvolveuse un *score test* que usa a aproximacción Saddlepoint (SPA, do inglés *Saddlepoint Aproximation*) para estimar a distribución baixo a hipótese nula de $\beta = 0$ (Dey et al., 2017). A a función score U defínese como a primeira derivada de $\log(L)$ con respecto a β. Se a estimación por máxima verosimilitude é igual ao valor da hipótese nula p_0, $U(p_0) = 0$. Por tanto, o score test mide a diferencia entre o valor da función score baixo a hipótese nula con respecto ao 0. Unha vez obtidas as estimacións baixo o modelo nulo dos parámetros $\hat{\alpha}, \hat{\tau}, \hat{\psi}, \hat{u}$, con $\hat{\pi}_i = P(Y = 1|X, \hat{u})$ e $\hat{W}$ unha matriz con elementos da diagonal iguais a $\mu(1-\mu)$, o estatístico do score test escribirase

$$T = G^T(Y - \hat{\pi}_i).$$

O estatístico T segue asintóticamente unha distribución normal, polo que só se usan os dous primeiros momentos (media e varianza). Daniels (citado en Dey et. al, 2017) propuxo empregar a aproximación por Saddlepoint para poder usar todos os cumulantes, resultando nunha maior precisión cando as distribucións son nesgadas. Non se detallará o procedemento pois non entra entre os obxectivos do traballo (para máis información ver Zhou et al., 2018; Dey et al., 2017), mais na práctica usaremos os p-valores con esta corrección.

3.3.2. Aplicación aos datos do COVID-19

O paquete citado na Sección 3.3.1, `SAIGE`, foi o elixido nun primeiro momento para axustar o modelo mixto. En cambio, na práctica foi máis útil unha modificación do mesmo, `SAIGEgds` (ver Apéndice B). Os paquetes `SAIGE` e `SAIGEgds` traballan da seguinte forma: en primeiro lugar axustan un modelo nulo sen a covariable SNP e nun segundo paso levan a cabo a asociación incluíndo cada SNP como variable X. Deste xeito axilízase a computación, pois o cálculo da varianza aditiva xenética e a inversión da matriz ψ só se realiza no primeiro paso. Recordamos que esta matriz se constrúe separadamente a partir dun set moito máis pequeno de marcadores independentes. Isto parece ser suficiente para a obtención dos estimadores, pois cada marcador aporta moi pouco á varianza xenética aditiva (Zhou et al., 2018).

A efectos de reducir ao máximo posible o erro no cálculo dos estimadores, o modelo nulo axustouse co total dos cromosomas e marcadores. Filtráronse por LD (escolléronse os SNPs cuxo $LD < 0{,}2$, ver Anexo 3) para estimar a GRM: dun total de 656.883 marcadores, seleccionáronse 153.955. O número inicial de individuos é de 10.190, dos cales 9429 toman o valor "0"para a variable *Severidade_4* e 761 toman o valor "1". Fixouse o parámetro de dispersión a 1.

Axuste sen covariables							
SNP	**Crom.**	**Pos.**	**Xen**	**N**	**P-val.**	**OR**	**IC**
rs71325088	3	45862952	LZTFL1	10179	1.24E-10	2.040	1.642-2.536
rs75928798	3	45962603	FYCO1	10178	2.40E-06	1.556	1.295-1.871
rs1994491	3	45960420	FYCO1	10179	2.53E-06	1.537	1.285-1.839
rs13097556	3	46061997	XCR1	10166	2.85E-06	1.532	1.281-1.831
rs1994492	3	45960646	FYCO1	10179	3.85E-06	1.543	1.283-1.854
rs2230322	3	46063329	XCR1	10129	3.90E-06	1.522	1.273-1.819
rs1994493	3	45960700	FYCO1	10176	4.10E-06	1.540	1.282-1.851
rs13079478	3	46007823	FYCO1	10158	4.17E-06	1.541	1.282-1.852
rs13079869	3	46008087	FYCO1	10171	5.38E-06	1.531	1.274-1.839
rs33910087	3	46009487	FYCO1	10173	5.87E-06	1.528	1.272-1.835
rs71327023	3	46131225	XCR1	10182	9.10E-06	1.512	1.26-1.815

Cadro 3.1: Táboa dos SNPs cuxo p-valor$< 10^{-5}$ para o axuste sen covariables do modelo mixto loxístico.

A saída dos modelos con esta libraría tamén é limitada, pero devolve os estimadores da varianza aditiva xenética: $\hat{\sigma}^2_u$ no modelo sen covariables é de 0.0629; No modelo con covariables, $\hat{\sigma}^2_u = 0{,}0623$. A asociación e o cálculo dos β_{SNP} e dos p-valores só se levaron a cabo co cromosoma 3.

De novo, o SNP rs7132088 é significativo. Os OR asociados ás variantes e os p-valores dos contrastes para o modelo mixto incluíndo parentes non difiren notablemente dos resutados do modelo loxístico simple excluíndo parentes. En canto ao Miami plot, a Figura 3.1 é moi parecida ao Miami plot do modelo loxístico simple (Figura 2.1) tanto coa inclusión de covariables como sen elas. En realidade, esta situación era a esperable, pois estamos corrixindo o parentesco críptico de dúas maneiras. A vantaxe do modelo mixto é que nos permite saltarnos o paso da identificación e retirada dos individuos, o que ademais nos evita reducir a N final. Concretamente neste estudo o número de casos (761) é limitado e non compensa perder mostras. Con respecto a isto último, o desequilibrio entre casos e controis é moi acusado e a aproximación por SPA é moi útil (nos cadros 3.1 e 3.1 únicamente se mostran os p-valores coa corrección aplicada).

Axuste con covariables							
SNP	**Crom.**	**Pos.**	**Xen**	**N**	**P-val.**	**OR**	**IC**
rs71325088	3	45862952	LZTFL1	8899	7.91E-10	1.99282521	1.599-2.483
rs1994491	3	45960420	FYCO1	8898	3.22E-06	1.54883845	1.288-1.862
rs75928798	3	45962603	FYCO1	8897	4.64E-06	1.5531627	1.286-1.875
rs13097556	3	46061997	XCR1	8887	6.21E-06	1.52651676	1.271-1.834
rs1994492	3	45960646	FYCO1	8899	6.40E-06	1.5442651	1.279-1.865
rs1994493	3	45960700	FYCO1	8896	6.59E-06	1.54302796	1.278-1.863
rs13079478	3	46007823	FYCO1	8877	7.28E-06	1.54124329	1.276-1.862
rs13079869	3	46008087	FYCO1	8895	8.25E-06	1.53621722	1.272-1.855
rs33910087	3	46009487	FYCO1	8893	9.12E-06	1.5326119	1.269-1.851

Cadro 3.2: Táboa dos SNPs cuxo p-valor $< 10^{-5}$ para o axuste coas covariables *Sexo*, *Idade* e as dez primeiras PCs do modelo mixto loxístico.

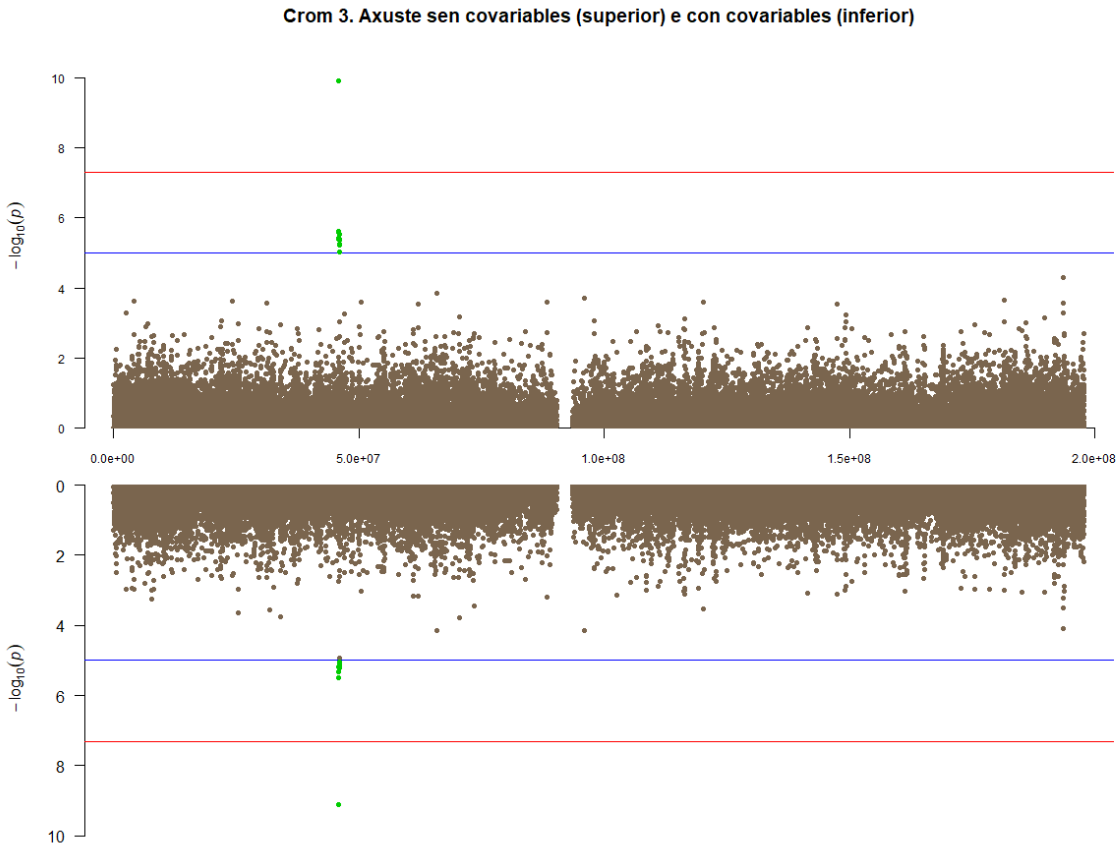

Figura 3.1: Miami plot do cromosoma 3 con covariables co modelo mixto. No eixo Y, $-\log_{10}$(p-valor). No eixo X, posicións no cromosoma.

Capítulo 4

Modelo multinomial

Os resultados obtidos ata o momento únicamente explican a base xenética do COVID-19 para o seu fenotipo máis crítico. Porén, dispoñemos dunha variable resposta *severidade* que clasifica aos enfermos en función das súas condicións clínicas e sintomatolóxicas, e non interesa perder a información relativa ás categorías de gravidade intermedias. Habitualmente, nos estudos GWAS con sub-fenotipos trabállase coa categoría de maior importancia, mantendo a estrutura de casos-controis. Neste Capítulo introduciremos unha variable Y que seguirá unha distribución multinomial e, por tanto, deberemos explorar alternativas ao modelo loxístico. Contemplaremos só efectos fixos, polo que o conxunto de datos sobre o que se axustará o modelo ten un tamaño de mostra $N = 9896$. Na literatura dos estudos GWAS o modelo multinomial non é moi empregado, quizais polo elevado número de parámetros e o aumento da dificultade na interpretación. [1]

4.1. Formulación do modelo

A variable Y, no modelo multinomial en comparación co loxístico, ten máis de dous niveis. O obxectivo será o de modelizar a probabilidade de pertencer á categoría k dados uns valores das covariables X, $P(Y = k|X)$. Por un lado, ignórase a orde das categorías, pero por outro lado non se asume proporcionalidade nas Odds entre elas. De feito, unha das hipóteses básicas do modelo multinomial é a independencia das alternativas irrelevantes (IIA, *independence of irrelevant alternatives*): incluír ou excluír categorías non cambiará a ratio de dúas probabilidades.

O modelo máis simple é o basal, que toma como categoría de referencia k_0, e será coa que se contrasta cada unha das demais. Denotemos a probabilidade de pertencer á categoría K como $\pi_k = Pr(Y = k)$, $k = 1, \dots, K$. Ao igual que no modelo loxístico, a transformación

[1] Os aspectos teóricos deste Capítulo están adaptados de Hosmer et al., (2013).

$g(x)$) será o *logit*. O preditor lineal basal η_k podemos escribilo como:

$$\eta_k = \log\left(\frac{\pi_k}{\pi_0}\right), \quad k = 1, \dots, K-1. \tag{4.1}$$

Na fórmula (4.1), π_0 denota a probabilidade de $Y = 0$, sendo 0 a categoría basal. Neste modelo engadimos unha restricción: $\eta_0 = 0$. Cando $K = 2$ temos un modelo loxístico habitual e só necesitamos un único preditor lineal η. Para unha variable resposta que tome os valores 0, 1 e 2 ($K = 3$), teremos dous preditores lineais:

$$\eta_1 = \log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \beta_{10} + \beta_{11}X_1 = x'\beta_1.$$

$$\eta_2 = \log\left(\frac{P(Y=2|X=x)}{P(Y=0|X=x)}\right) = \beta_{20} + \beta_{21}X_1 = x'\beta_2.$$

Invertindo a transformación $g(x)$ e sendo β_{kp} o coeficiente asociado á variable X_p (con $p = 0$ para o intercepto) no nivel k predeciremos as probabilidades de pertencer á categoría k:

$$P(Y=0|X=x) = \frac{1}{1 + e^{\beta_{10}+\beta_{11}X_1} + e^{\beta_{20}+\beta_{21}X_1}};$$

$$P(Y=1|X=x) = \frac{e^{x'\beta_1}}{1 + e^{\beta_{10}+\beta_{11}X_1} + e^{\beta_{20}+\beta_{21}X_1}};$$

$$P(Y=2|X=x) = \frac{e^{x'\beta_2}}{1 + e^{\beta_{10}+\beta_{11}X_1} + e^{\beta_{20}+\beta_{21}X_1}}.$$

Podemos xeralizar a formulación a K categorías da variable Y da forma

$$P(Y=k|X=x) = \frac{e^{\eta_k(x)}}{\sum_{k=0}^{K} e^{\eta_k(x)}},$$

onde $\eta_0 = 0$. Finalmente, teremos $K-1$ vectores de parámetros β, porque o vector asociado ao grupo de referencia ou basal β_0 será un vector de ceros.

4.1.1. Interpretación dos parámetros

A interpretación dos parámetros será un pouco máis complicada que no modelo binomial. No caso de que os coeficientes asociados á variable X teñan o mesmo signo, si se pode sacar unha conclusión sobre as probabilidades de pertencer a un grupo distinto do de referencia en relación a aumentar unha unidade en X. Por exemplo, supoñendo coeficientes positivos para todas as categorías, aumentar unha unidade na variable explicativa si que se pode interpretar como unha diminuición da probabilidade de pertencer ao grupo de referencia. Sen embargo, para calcular estas probabilidades é necesario traballar cos efectos marxinais. β_{kp} será o efecto marxinal de X_p na log-OR da alternativa k fronte á de referencia.

As OR, en cambio, terán a interpretación habitual, pero cada categoría compararase coa de referencia que se escollera na definición do modelo. O máis lóxico é situar o nivel máis baixo

como o de referencia. De novo, consideremos o modelo máis básico cos nosos datos, cunha covariable X que toma os valores 0, 1 ou 2 segundo o número de alelos menos frecuentes para un SNP dado. A categoría de referencia para a variable *gravidade* será 0, controis, e os demais grupos serán 1 (asintomático), 2 (enfermidade leve), 3 (enfermidade moderada), 4 (enfermidade grave) e 5 (enfermidade crítica).

$$OR_k = \frac{Odds(Y = k|X = 1)}{Odds(Y = 0|X = 0)} = \frac{\frac{P(Y=k|X=1)}{P(Y=0|X=1)}}{\frac{P(Y=k|X=0)}{P(Y=0|X=0)}}.$$

Así, e^{β_k} é o efecto multiplicativo da Odds do grupo k en relación ao nivel basal por cada unidade que aumenta X. Por exemplo, una $OR_3 = 2$ significará que un individuo medio, ao aumentar unha copia do alelo menos frecuente, terá unha Odds dúas veces maior de presentar un fenotipo moderado con relación ao grupo de controis.

Porén, estas Odds son relativas sobre o grupo de referencia. É dicir, supoñamos que ademais, a OR para a categoría 4 é $OR_4 = 2{,}5$. Incrementar unha copia do alelo menos frecuente irá asociado a unha maior probabilidade relativa de padecer enfermidade moderada que de estar no grupo control fixadas as covariables, pero é aínda maior a probabilidade relativa de padecer enfermidade grave que de estar no grupo control. Pola contra, non sabemos cal é probabilidade de presentar enfermidade grave en función dun aumento do número de alelos menos frecuentes na variable SNP: para sabelo hai que calcular os efectos marxinais. De feito, o signo desta relación non ten que ser o mesmo que o do coeficiente β_k.

A interpretación pode parecer directa no noso caso porque, no fondo, estamos tratando unha variable Y naturamente ordeada onde a categoría basal é non padecer a enfermidade. Poñámonos agora na situación oposta, na que a variable resposta é categórica pura e non existe unha orde natural dos niveis: queremos analizar o efecto da renta sobre a elección de someterse a varios chequeos médicos. A categoría basal será *ir o dentista*. Os resultados da regresión multinomial indican que a xente con máis ingresos prefire invertir no dentista que no psicólogo. Porén, prefiren invertir no pediatra que no dentista. Por tanto, coñecemos as preferencias relativas en función da renta, mais non coñecemos o efecto da renta sobre cada unha das categorías: obter maiores ingresos implica unha maior probabilidade de invertir no psicólogo? Ao contrario que no modelo loxístico, as estimacións dos coeficientes do modelo multinomial non nos indican cal é a categoría máis probable ao aumentar unha unidade da variable X. Compre ter coidado na interpretación dos coeficientes.

4.1.2. Inferencia sobre os parámetros

A estimación dos parámetros do modelo multinomial tamén será a partir da súa función de verosimilitude. Para construíla crearemos tres variables *dummy* Y_k, que tomarán valor 1 se a observación i pertence ao grupo K e 0 no caso contrario. A suma das variables Y_k será igual a 1.

$$L(\beta) = \prod_{i=1}^{n} [\pi_0(x_i)^{y_{0i}} \pi_1(x_i)^{y_{1i}} \pi_2(x_i)^{y_{2i}}].$$

Tendo en conta que a suma de $Y_k = 1$, aplicamos agora o logaritmo e derivamos parcialmente con respecto a β:

$$logL(\beta) = \sum_{y=1}^{n} y_{1i}\eta_1(x_i) + y_{2i}\eta_2(x_i) - \log(1 + e^{\eta_1(x_i)} + e^{\eta_2(x_i)}).$$

$$\frac{\partial L(\beta)}{\partial \beta_{kp}} = \sum_{i=1}^{n} x_{pi}y_{ki} - \pi_k(x_i);$$

para $k = 3$ e $P = 1, \dots, p$ covariables X.

O problema de optimización para obter as estimacións dos coeficientes resólvese con algoritmos iterativos. Nos paquetes `mlogit` e `mnlogit` empregan o algoritmo de Newton-Raphson, que require o cálculo da matriz Hessiana e do gradiente da función de verosimilitude (Croissant, 2020; Hasan, Zhiyu e Mahani, 2019).

Determinar a significancia dos coeficientes no modelo multinomial varía un pouco con respecto ao modelo binomial, pois agora teremos que contrastar, por un lado, se existen diferencias entre o grupo K e o de referencia e, por outro, se a variable SNP é significativa no modelo. O primeiro contraste consiste en aplicar un test de Wald a cada coeficiente β_{kSNP} baixo a hipótese nula de que é igual a cero, ou equivalentemente, que a súa OR é igual a 1. Así, se a hipótese nula é certa para unha categoría esta poderíase combinar coa de referencia. Finalmente, emprégase un test de ratio de verosimilitudes para o contraste de hipóteses sobre a covariable SNP:

$$LR = -2\log\left(\frac{L_0}{L_a}\right). \tag{4.2}$$

Na expresión (4.2), L_0 denota o valor da verosimilitude no modelo baixo a hipótese nula (é dicir, sen a covariable en cuestión), e L_a a verosimilitude no modelo menos restrinxido. O estatístico LR seguirá unha distribución X^2 con t graos de liberdade, sendo t a diferencia de parámetros entre ambos modelos.

4.2. Aplicación aos datos do COVID-19

A librería escollida para o axuste do modelo multinomial é a `mnlogit`, a cal é unha adaptación de `mlogit` pero reducindo considerablemente o tempo e velocidade de computación. De novo, a elección dos paquetes de R optimizados é esencial ao traballar con conxuntos tan grandes de datos. O preditor lineal xeral do modelo multinomial para a variable *gravidade* é:

$$\eta_k = log\left(\frac{P(Y = k|Idade, Sexo)}{P(Y = 0|Idade, sexo)}\right), \quad k = 1, \dots, 5.$$

O grupo control será o grupo de referencia.

Como non introducimos un efecto aleatorio asociado ao parentesco, o conxunto de individuos estudado é o mesmo que no modelo loxístico. Entre eles temos 3357 persoas no grupo

control; 407 asintomáticos; 1606 con fenotipo leve; 1042 con fenotipo moderado; 1229 con fenotipo grave, e 742 con fenotipo crítico. A función usada para o test de ratio de verosimilitudes pertence á librería `lmtest`.

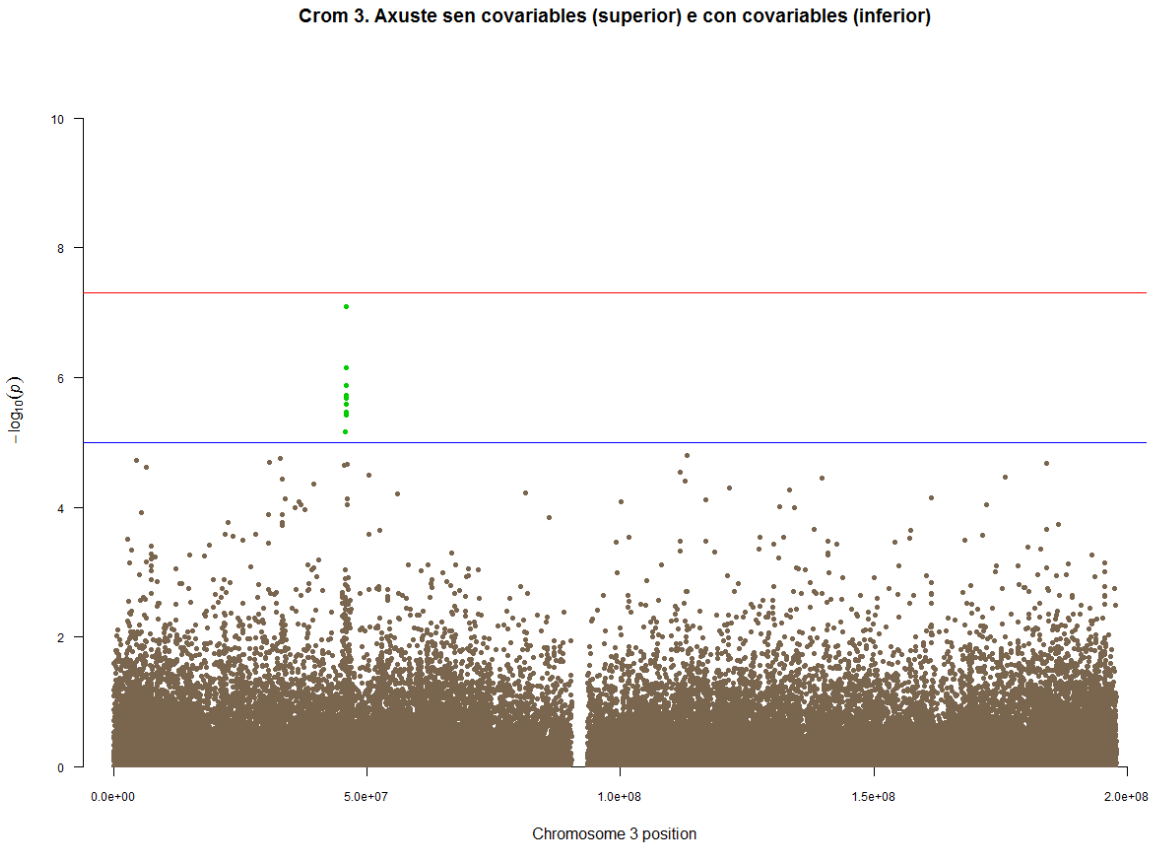

Figura 4.1: Manhattan plot do cromosoma 3 con covariables co modelo multinomial. No eixo Y, $-\log_{10}$(p-valor). No eixo X, posicións no cromosoma.

No Cadro 4.1 están resumidos os resultados para a regresion multinomial sempre que o p-valor asociado ao SNP para o test de razón de verosimilitudes estivera por debaixo de $10e^{-5}$. Para 8 dos 9 SNPs a OR entre a categoría 1 (asintomáticos) e a categoría 0 (controis) non é significativamente distinta de 1, ou o que é o mesmo, o seu β_k non é significativamente distinto de cero. A interpretación disto é que se poderían fusionar ambas categorías sen perder información, pois a variable SNP non ten efecto sobre os asintomáticos máis alá do que ten sobre o grupo control. Por outro lado, as OR para as categorías 4 e 5 (grave e crítico, respectivamente) tamén son moi similares entre si para todos os SNPs. Os p-valores asociados, sen embargo, son cercanos a cero, pero recordemos que estamos comparando sempre coa categoría de referencia.

En xeral, os resultados para esta regresión non se afastan demasiado dos resultados para a regresión loxística: a rexión do cromosoma 3 é a mesma e só o SNP rs71325088 é significativo no umbral habitual. Sen embargo, coa corrección de Bonferroni ($< 1{,}25e^{-6}$) obtivemos outro

SNP significativo: o rs12079478, do xen FYCO1.

Aínda que non o pareza a simple vista, o resultado é satisfactorio. A pesar de que finalmente non identificamos ningún sinal novo, si que nos permitiu observar de forma global cal é a contribución xenética ás categorías de gravidade na enfermidade por COVID-19: polo menos para o cromosoma 3, sabemos que os asintomáticos non presentan diferencias xenéticas detectables con este tamaño de mostra, e que estas se empezan a manifestar a partir dos fenotipos leve-moderado. Para futuras análises, a categoría 0 colapsarase coa categoría control.

Axuste sen covariables									
SNP	**Crom.**	**Pos.**	**Xen**	**P-val.**	**OR 1:0**	**OR 2:0**	**OR 3:0**	**OR 4:0**	**OR 5:0**
rs71325088	3	45862952	LZTFL1	8.02E-08	1.058 (0.755-1.483)	1.33** (1.109-1.595)	1.611*** (1.316-1.973)	1.693*** (1.387-2.066)	1.73 (1.38-2.169)
rs13079478	3	46007823	FYCO1	7.03E-07	0.982 (0.738-1.308)	1.172* (1.002-1.371)	1.557*** (1.307-1.854)	1.454*** (1.219-1.736)	1.531*** (1.252-1.871)
rs1994492	3	45960646	FYCO1	1.34E-06	1.032 (0.778-1.369)	1.198* (1.025-1.401)	1.590*** (1.337-1.892)	1.424*** (1.193-1.701)	1.485*** (1.214-1.817)
rs1994493	3	45960700	FYCO1	1.91E-06	1.03 (0.777-1.366)	1.194* (1.022-1.396)	1.580*** (1.328-1.88)	1.421*** (1.19-1.696)	1.482*** (1.211-1.813)
rs1994491	3	45960420	FYCO1	2.07E-06	0.998 (0.756-1.317)	1.168* (1.002-1.361)	1.519*** (1.28-1.801)	1.416*** (1.191-1.683)	1.515*** (1.246-1.841)
rs75928798	3	45962603	FYCO1	2.53E-06	1.041 (0.787-1.378)	1.182* (1.011-1.382)	1.577*** (1.326-1.876)	1.41*** (1.181-1.683)	1.474*** (1.205-1.803)
rs13079869	3	46008087	FYCO1	3.37E-06	0.965 (0.725-1.285)	1.147 (0.981-1.341)	1.522*** (1.279-1.812)	1.414*** (1.185-1.687)	1.495*** (1.224-1.827)
rs33910087	3	46009487	FYCO1	3.83E-06	0.959 (0.72-1.276)	1.14 (0.975-1.333)	1.518*** (1.275-1.806)	1.412*** (1.183-1.684)	1.487*** (1.217-1.817)

Cadro 4.1: Táboa dos SNPs cuxo p-valor$< 1\times10^{-5}$ para o axuste do modelo multinomial. Danse as OR para cada nivel con respecto ao grupo control xunto cos seus intervalos de confianza. As significacións para as OR son: p-val $< 0{,}05$ (*),p-val $< 0{,}01$ (**) e p-val $< 0{,}001$ (***).

Capítulo 5

Conclusións

O obxectivo práctico deste traballo era o de analizar a base xenética de infección por COVID-19. Para isto seleccionáronse tres aproximacións teóricas diferentes sobre o cromosoma 3.

A presenza de estratificación poboacional conleva unha dependencia nos datos que viola unha das suposicións básicas do modelo lineal xeral. Nunha primeira parte comparouse a estratexia tradicional de exclusión de individuos que presentaban dependencia xenética cunha segunda estratexia na que se modelizaba esta estrutura mediante un efecto aleatorio. Estes dous modelos axustáronse sobre unha variable binaria que consideraba como casos só aos enfermos críticos, polo que se precisou a transformación logit sobre os datos. Así, o que se fixo na práctica e dende unha perspectiva máis biolóxica foi estudar a base xenética do fenotipo máis extremo.

Ambos procedementos deron resultados similares, concluíndo que o modelo mixto é un bo enfoque para os estudos GWAS, xa que permite manter todos os individuos e á súa vez controlar os distintos factores confusores sen violar ningunha asunción teórica do modelo. Por outro lado, a flexibilidade dos modelos mixtos facilitarían a inclusión doutros efectos aleatorios, incluso medioambientais, que puideran afectar á probabilidade de padecer un cadro clínico grave de COVID-19. En canto aos resultados xenéticos, confirmamos o sinal na rexión do cromosoma 3 que xa fora reportada con anterioridade.

Con todo, os datos recollidos no proxecto *SCOURGE* son moi variados e moi ricos en información. Isto proporciónanos unha oportunidade única para explorar outros modelos de regresión aplicables aos GWAS, puidendo ampliar a análise ao espectro completo de gravidade. Na segunda parte do traballo axustouse un modelo multinomial basal, nada común en estudos de asociación, que contrastaba a poboación control con doentes distribuídos en cinco categorías de severidade. O modelo multinomial evidenciou que a base xenética asociada aos distintos graos de severidade é detectable a partir do fenotipo leve-moderado. Ademais, non se atoparon diferencias significativas entre as categorías control e asintomático. Porén, unha limitación dos nosos datos é que o número de asintomáticos en comparación co N do resto das categorías é moi baixo, e este desequilibrio pode repercutir na significación dos coeficientes asociados.

De novo, confirmouse o sinal do cromosoma 3.

O SNP significativo en todos os modelos foi o rs71325088, do xen LZTFL1, expresado nas células do epitelio respiratorio (The Severe Covid-19 GWAS Group, 2020). A maiores, no modelo multinomial superou a corrección de Bonferroni para 40.000 SNPs o rs12079478 do xen FYCO1. Estes resultados son satisfactorios, agora ben, estas análises deben realizarse sobre os 22 cromosomas restantes. De feito, o modelo multinomial non se axustou ata o momento sobre os datos do COVID-19 en ningún estudo, e farase no proxecto *SCOURGE*.

Non obstante, non se introduciu un efecto aleatorio no modelo multinomial e será interesante investigar esta opción. As limitacións computacionais das librerías existentes en R suporán un obstáculo, sobre todo á hora de computar a matriz de varianzas-covarianzas dos efectos aleatorios a partir do conxunto dos xenotipos. Por outro lado, comentamos tamén que no modelo multinomial asumimos que as categorías da variable resposta non están ordeadas. A vantaxe é que tampouco se asume unha proporcionalidade entre cada nivel, pero na realidade sabemos que a categoría "grave"é superior á categoría "leve"e pode que perdamos información. Unha alternativa plausible sería a de axustar un modelo multinomial ordinal.

Finalmente, é importante aclarar que este traballo é unha introdución a un estudo que aínda non está rematado e que, por suposto, queda moito traballo por facer ata esclarecer todas as dúbidas surxidas ao redor da interacción virus-hóspede para o SARS-CoV-2. De feito, neste traballo non se exploraron outras covariables como as comorbilidades e pode que estas, xunto coa aproximación multinomial, poidan aportar novos datos á investigación.

Apéndice A

Aspectos prácticos do estudo

Este primeiro apéndice resume con brevedade os procedementos realizados no estudo e as dificultades que se atoparon durante o seu curso.

Previamente ao inicio do estudo foi necesaria a limpeza e organización da ampla base de datos do proxecto para posteriormente escoller as variables pertinentes para o GWAS. Esta parte foi realizada mediante scripts de R.

A maiores, os datos de xenotipado requeriron un control de calidade exhaustivo co obxectivo de reducir ao máximo o ruído introducido tanto por erros no laboratorio como por erros no algoritmo de xenotipado. O control de calidade realizado en `PLINK` (ver Apéndice B) sobre os mesmos é unha adaptación de Anderson et al. (2010) e tomáronse os seguintes pasos:

- Selección dos cromosomas 1-22 (autosomas, excluímos os cromosomas sexuais).
- Selección de marcadores cuxa MAF$>$ 0,05.
- Cálculo de heterozigosidade e LD.
- Eliminación de marcadores e individuos cuxa porcentaxe de *missings* superaba o 2 %.
- Eliminación de marcadores que se desviaban significativamente do equilibrio Hardy-Weinberg.
- Selección de marcadores que se atopan en equilibrio de ligamento (LD$<$ 0,2). Con este conxunto de marcadores procedeuse ao seguinte:

  • Fusión dos individuos do proxecto *SCOURGE* cos individuos de 1000G e análise de Compoñentes Principais (10 PCs).

  • Cálculo de IBD para identificar parentes cercanos.

  • En R: Detección de outliers de IBD e ancestralidade non europea (individuos cuxas PCs se desvían máis de dúas desviacións típicas dos europeos de 1000G).
- Eliminación de outliers: parentes + heterocigosidade excesiva + non europeos (conxunto de individuos 1); únicamente non europeos + heterocigosidade excesiva (conxunto de

individuos 2).

- Selección dos cromosomas 1-22 e X.
- Repetición do control de calidade básico: Selección dos marcadores cuxa MAF$>$ 0,05, porcentaxe de *missings* $<$ 0,02 e que non se desvíen do equilibrio Hardy-Weinberg; selección de individuos cuxo porcentaxe de *missings* $<$ 0,02.

Por outro lado, o gran número de modelos a axustar (ao redor de 40.000 para o cromosoma 3) require que o software que se usa estea orientado a cantidades masivas de datos, aínda mellor se é específico para datos de alta dimensión como son os xenéticos. Para os axustes dos Capítulos 2 e 3 non houbo problemas, pois as librarías escollidas (ver Apéndice B) están dirixidas para estudos GWAS. Sen embargo, para axustar o modelo multinomial do Capítulo 4 si atopamos dificultades. Concretamente, o paquete `mlogit` escollido nun inicio probouse sobre un conxunto de 20 SNPs e resultou ser excesivamente lento. Hai que ter en conta que o test de ratio de verosimilitudes que devolve a función do axuste contrasta o modelo con todas as covariables sobre o modelo con intercepto, polo que foi necesario axustar un modelo nulo sen a covariable SNP. Noutras palabras, por cada marcador axustáronse dous modelos multinomiais (un total de 80.000 aproximadamente). Finalmente atopamos unha versión computacionalmente máis eficiente, `mnlogit`.

Porén, fíxose necesario a utilización das instalacións do CESGA en todos os casos. Os traballos foron enviados a diferentes nodos do ordenador Finis Terrae II. Procurouse minimizar os recursos como a RAM pero si usar un número elevado de núcleos. Esta paralelización permitiu reducir considerablemente o tempo de execución dos códigos de R para `SAIGEgds` e `mnlogit`, resultando en 20 minutos e 3 horas respectivamente para o cromosoma 3. Non foi posible paralelizar co paquete `GenABEL` e neste caso si houbo que aumentar a RAM. Comentar tamén que non foi posible a inclusión da matriz de parentesco en `mnlogit`, polo que non se puido incluír un efecto aleatorio no modelo multinomial.

Apéndice B

Descrición do software e librarías empregadas

A continuación describirase o software e os paquetes de R mencionados ao longo do traballo.

B.1. PLINK

`PLINK` (Purcell et al., 2007) é un software libre para a ánalise de xenotipos e fenotipos. As súas funcións varían dende a lectura e conversión de arquivos, cálculo de estatísticos de resumo, control de calidade, detección de estrutura poboacional e estudos de asociación e metaanálise. Durante o curso deste traballo usáronse dúas versións, `PLINK` 1.9 e `PLINK` 2.0. Actualmente está mantido por Cristopher Chang, Carson Chow, Shashaank Vattikuti, Laurent Tellier e James Lee. O control de calidade dos datos de xenotipado realizouse con este programa.

B.2. GenABEL (R)

`GenABEL` (Aulchenko et al., 2007) é unha librería de R especializada en realizar GWAS a través de modelos loxísticos simples. Ademais, tamén contén funcións para unha análise estatística de datos xenéticos estendida, incluíndo control de calidade propio e análise descritiva.

Neste traballo empregouse para o axuste do modelo loxístico simple coa función **mlreg**. Tamén se realizou un segundo control de calidade previo ao axuste mediante a súa función **check.marker**.

B.3. SAIGE e SAIGEgds (R)

A libraría `SAIGE` (Zhou et al., 2018) emprega modelos mixtos para levar a cabo estudos de asociación tanto en caracteres cuantitativos como binarios. Actualmente só funciona en entornos de UNIX e non é posible a súa instalación en Windows. As funcións principais son **step1_fitNULLGLMM.R**, que axusta o modelo nulo, e **step2_SPAtests.R**, para as asociacións e cálculo dos p-valores.

Sen embargo, o paquete usado no traballo foi `SAIGEgds` (Zheng, X. et al., 2020). É unha adaptación do paquete orixinal `SAIGE` a Windows, mantendo a metodoloxía. A principal diferencia entre ambos paquetes é que `SAIGEgds` emprega un obxecto GDS (*genomic data structure*) descrito en Zheng et al. (2012) para almacenar os xenotipos e outros metadatos en formato *array*. Ademais, implementa outros procedementos numéricos. Ambas cousas aceleran considerablemente a computación, se ben os cálculos das compoñentes dos modelos e a estratexia en dous pasos se conserva.

A selección de marcadores en equilibrio de ligamento foi feita coa función **snpgdsLDpruning** da libraría `SNPRelate` (Zheng et al., 2012).

B.4. mlogit e mnlogit (R)

O axuste do modelo multinomial para o conxunto de SNPs realizouse coa libraría `mnlogit` (Hasan et al., 2019). É unha versión da libraría `mlogit` (Croissant, 2020) adaptada a conxuntos de datos máis complexos. Ambas librarías son usadas habitualmente no campo da econometría e psicoloxía. Nótese que `mlogit` permite incluír un efecto aleatorio pero non ocorre o mesmo con `mnlogit`. Pola contra, ambas permiten a inclusión de covariables asociadas á variable Y.

Bibliografía

[1] Abney, M., Mcpeek, M. S., Ober, C. (2000). Estimation of Variance Components of Quantitative Traits in Inbred Populations. *The American Journal of Human Genetics*, 66(2), 629-650. doi:10.1086/302759

[2] Anderson, C., Pettersson, F., Clarke, G. et al. Data quality control in genetic case-control association studies. *Nature Protocols*, 5, 1564-1573 (2010). doi: 10.1038/nprot.2010.116

[3] Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.* Series B (Methodological), 57(1), 289-300. Retrieved from http://www.jstor.org/stable/2346101

[4] Aulchenko, Y., Ripke, S., Isaacs, A. , Van Duijn, Cornelia. (2007) GenABEL: an R library for genome-wide association analysis, *Bioinformatics*, 23(10), 1294-1296. doi: 10.1093/bioinformatics/btm108

[5] Casas-Rojo JM, Antón-Santos JM, Millán-Núñez-Cortés J, Lumbreras-Bermejo C, Ramos-Rincón JM, Roy-Vallejo E, Artero-Mora A, Arnalich-Fernández F, García-Bruñén JM, Vargas-Núñez JA, Freire-Castro SJ, Manzano-Espinosa L, Perales-Fraile I, Crestelo-Viéitez A, Puchades-Gimeno F, Rodilla-Sala E, Solís-Marquínez MN, Bonet-Tur D, Fidalgo-Moreno MP, Fonseca-Aizpuru EM, Carrasco-Sánchez FJ, Rabadán-Pejenaute E, Rubio-Rivas M, Torres-Peña JD, Gómez-Huelgas R; en nombre del Grupo SEMI-COVID-19 Network. (2020). Clinical characteristics of patients hospitalized with COVID-19 in Spain: Results from the SEMI-COVID-19 Registry. *Revista Clínica Española*, 2, 220(8), 480-494. doi: 10.1016/j.rce.2020.07.003

[6] Cavalli-Sforza, L., Menozzi, L., Piazza, A. (1994) Introduction to Concepts, Data, and Methods. En Cavalli-Sforza, L., Menozzi, L., Piazza, A. *The History and Geography of Human Genes* (p. 3-54). New Jersey: Princeton University Press.

[7] COVID-19 Host Genetics Initiative. (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, 28, 715-718. doi: 10.1038/s41431-020-0636-6.

[8] Croissant, Y. (2020). Estimation of Random Utility Models in R: The mlogit Package. *Journal of Statistical Software*, 95(11), 1-41. doi: 10.18637/jss.v095.i11

[9] Crujeiras, R., Conde, M. (2020). *Regresión xeralizada e modelos mixtos*. [PDF]. Material non publicado.

[10] Devlin, B., Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, 55(4), 997-1004. doi:10.1111/j.0006-341x.1999.00997.x

[11] Dey, R., Schmidt, E. M., Abecasis, G. R., Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *American Journal of Human Genetics*, 101, 37-49. doi:10.1016/j.ajhg.2017.05.014

[12] Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics* (4a ed). Harlow: Addison Wesley Longman.

[13] Gilmour, A. R., Thompson, R., Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, 51(4), 1440-14450.

[14] Hasan, A., Wang, Z., Mahani, A. (2016). *Fast Estimation of Multinomial Logit Models: R Package mnlogit*. Journal of Statistical Software, 75(3), 1-24. doi:http://dx.doi.org/10.18637/jss.v075.i03

[15] Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Guelph: University of Guelph.

[16] Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied logistic regression* (2da ed). Hoboken, NJ: Wiley.

[17] Kang H. M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C., Eskin E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4), 348-54. doi: 10.1038/ng.548

[18] Kwok, A. J., Mentzer, A., Knight, J. C. (2020). Host genetics and infectious disease: new tools, insights and translational opportunities. Nature reviews. Genetics, 1-17. Advance online publication. https://doi.org/10.1038/s41576-020-00297-6

[19] Malaria Genomic Epidemiology Network, Band G., Rockett K.A., Spencer C.C., Kwiatkowski D.P. (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, 8, 526(7572), 253-257. doi: 10.1038/nature15390

[20] McVean, K., Kelleher, J. (2019). Linkage Desequilibrium, Recombination and Haplotype Structure. En Balding, D., Moltke, I., Marioni, J. (Ed.) *Handbook of Statistical Genomics* (4ª ed., vol. 2, p. 51-76). UK: Wiley.

[21] Morris, P., Cardon, L. (2019). Genome-Wide Association Studies. En Balding, D., Moltke, I., Marioni, J. (Ed.) *Handbook of Statistical Genomics* (4ª ed., vol. 2, p. 597-623). UK: Wiley.

[22] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218), 98-101. https://doi.org/10.1038/nature07331

[23] Otani, T., Noma, H., Nishino, J., Matsui, S. (2018). Re-assessment of multiple testing strategies for more efficient genome-wide association studies. European journal of human genetics : EJHG, 26(7), 1038-1048. https://doi.org/10.1038/s41431-018-0125-3

[24] Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, Walker S, Parkinson N, Fourman MH, Russell CD, Furniss J, Richmond A, Gountouna E, Wrobel N, Harrison D, Wang B, Wu Y, Meynert A, Griffiths F, Oosthuyzen W, Kousathanas A, Moutsianas L, Yang Z, Zhai R, Zheng C, Grimes G, Beale R, Millar J, Shih B, Keating S, Zechner M, Haley C, Porteous DJ, Hayward C, Yang J, Knight J, Summers C, Shankar-Hari M, Klenerman P, Turtle L, Ho A, Moore SC, Hinds C, Horby P, Nichol A, Maslove D, Ling L, McAuley D, Montgomery H, Walsh T, Pereira A, Renieri A; GenOMICC Investigators; ISARICC Investigators; COVID-19 Human Genetics Initiative; 23andMe Investigators; BRACOVID Investigators; Gen-COVID Investigators, Shen X, Ponting CP, Fawkes A, Tenesa A, Caulfield M, Scott R, Rowan K, Murphy L, Openshaw PJM, Semple MG, Law A, Vitart V, Wilson JF, Baillie JK. (2020). Genetic mechanisms of critical illness in Covid-19. *Nature*, doi: 10.1038/s41586-020-03065-y

[25] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904-909. doi:10.1038/ng1847

[26] Price, A. L., Zaitlen, N. A., Reich, D., Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7), 459-463. doi:10.1038/nrg2813

[27] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559-575. doi:10.1086/519795

[28] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[29] Saavedra, P. (2019). *Modelos de regresión. Tema 8: Regresión loxística.* [PDF]. Material non publicado.

[30] Takahashi, T., Iwasaki, A. (2021). Sex differences in immune responses. *Science*, 371(6527), 347-348. doi:10.1126/science.abe7199

[31] Wedderburn, R. W. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61(3), 439-477. doi:https://doi.org/10.2307/2334725

[32] Worldometers, 2021. Recuperado de: https://www.worldometers.info/coronavirus/.

[33] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208. https://doi.org/10.1038/ng1702

[34] Zeberg, H., Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, 587, 610-612. https://doi.org/10.1038/s41586-020-2818-3

[35] Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24), 3326-3328. doi:10.1093/bioinformatics/bts606

[36] Zheng, X., Davis, J. W. (2020). SAIGEgds - an efficient statistical tool for large-scale PheWAS with mixed models. Bioinformatics (Oxford, England), btaa731. Advance online publication. https://doi.org/10.1093/bioinformatics/btaa731

[37] Zheng Z., Peng F., Xu B., Zhao J., Liu H., Peng J., Li Q., Jiang C., Zhou Y., Liu S., Ye C., Zhang P., Xing Y., Guo H., Tang W. (2020). Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysisJournal of Infection, 81(2). doi:10.1016/j.jinf.2020.04.021

[38] Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics, 50(9), 1335-1341. https://doi.org/10.1038/s41588-018-0184-y

[39] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., Tan, W., & China Novel Coronavirus Investigating and Research Team (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England journal of medicine*, 382(8), 727-733. https://doi.org/10.1056/NEJMoa2001017