eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Análisis de modelos estadísticos para estudios de asociación del genoma completo

Trabajo Fin de Grado
Grado en Matemáticas

Ania Gorostiza López de Subijana

Trabajo dirigido por
Dae-Jin Lee
Irantzu Barrio

Leioa, 28 de junio de 2016

Índice general

Introducción

En este trabajo se introducen técnicas estadísticas que permiten abordar un tipo de problema genético: los estudios de asociación del genoma completo o GWAS (del inglés *Genome Wide Association Studies*). Estos estudios pretenden identificar las variables genéticas que están relacionadas o influyen en el riesgo de padecer una determinada enfermedad.

La variable respuesta en este tipo de estudios es la ausencia o presencia de enfermedad. Sin embargo, aunque nos encontremos ante una variable dicotómica, los métodos de regresión estudiados en la asignatura de "Análisis Multivariante" no serán adecuados para estos problemas por distintos motivos:

- Al tratarse de estudios del genoma completo, el número de variables predictoras es muy elevado, en ocasiones mucho mayor que el número de observaciones, lo que dará lugar a un sobreajuste del modelo y a coeficientes con alta varianza.

- Es necesario estudiar la interacción entre variables, ya que muchas enfermedades son producto de interacciones entre genes. Esta necesidad se traduce en un aumento en el número de variables a introducir en el estudio.

- Muchas de las variables genéticas están muy correladas entre sí, por lo que no siempre contamos con variables predictoras independientes.

Es por ello que extenderemos los métodos estudiados en el grado para amoldarlos a la clase de problemas que nos ocupan.

El trabajo está dividido en tres partes: en la primera introducimos los conceptos genéticos básicos a tener en cuenta para realizar estudios de asociación genética. En la segunda parte hacemos un breve repaso del modelo de regresión logística y lo extendemos de manera que sea adecuado para abordar problemas de asociación del genoma completo. Las extensiones que estudiaremos son las regresiones "*Ridge*", "*Lasso*" y "*Elastic Net*", que mejorarán el modelo logístico reemplazando el ajuste por máxima verosimilitud por

otros ajustes alternativos. En tercer lugar, compararemos los métodos estadísticos introducidos en la segunda parte aplicándolos a una base de datos genéticos. Este estudio nos permitirá observar las propiedades generales y las diferencias entre las distintas técnicas estadísticas, así como comprobar su habilidad para detectar variables genéticas relacionadas con enfermedades. Al final del trabajo se muestran las conclusiones obtenidas tras la realización del mismo y se adjunta un apéndice en el que se muestra un extracto del código utilizado para realizar el estudio en R.

Capítulo 1

Estudios de asociación genética

Los estudios de asociación genética tienen como objetivo identificar variables genéticas que influyen o están relacionadas con padecer una determinada enfermedad. En este capítulo explicaremos los conceptos y propiedades necesarios para poder realizar dicho tipo de estudios.

1.1. Conceptos genéticos

El ácido desoxirribonucleico (ADN) es la molécula que contiene la información genética de todos los seres vivos. Esta molécula consiste en dos cadenas que se enrollan entre ellas formando una estructura de doble hélice. Cada cadena está formada por azúcares, grupos de fosfato y bases o nucleótidos. Las dos cadenas se mantienen unidas por enlaces entre las bases.

El ADN está a su vez formado por cromosomas, que en la especie humana son 23 pares: 22 autosomas o pares de cromosomas homólogos y un par de cromosomas que determinan el sexo del individuo. Para cada par de cromosomas, se hereda uno de cada progenitor. En organismos diploides como los nuestros, cada uno de los pares de cromosomas contiene para cada carácter o rasgo una pareja de genes (o alelos) en posiciones (*locus*, en plural *loci*) análogas. Llamamos gen a un fragmento de ADN que contiene la información necesaria para construir una determinada proteína, que a su vez controla la manifestación de un determinado rasgo. Los alelos, por su parte, son formas alternativas de un mismo gen, informan sobre un mismo rasgo y ocupan la misma posición en los cromosomas homólogos. La Figura 1.1 puede ayudarnos a entender estos conceptos con mayor claridad.

El genotipo es la combinación de alelos en el mismo *locus* para dos cromosomas homólogos. Si denominamos a los dos posibles alelos indicadores de

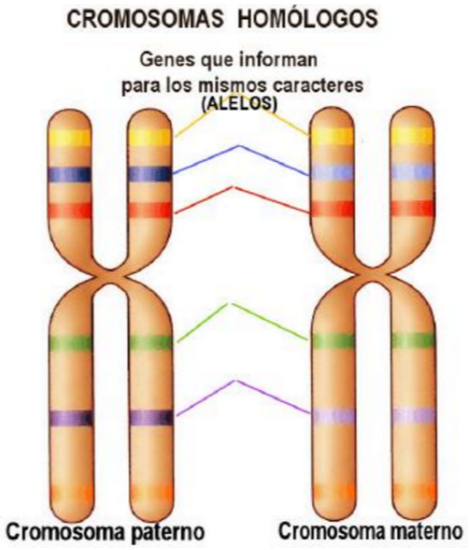


Figura 1.1: Esquema representativo de los conceptos de cromosomas homólogos y alelos. Fuente: http://cienaturales8.blogspot.com.es

un rasgo como "A" y "a", el genotipo se clasifica en homocigoto, si ambos alelos son iguales ("AA" o "aa") o heterocigoto cuando son distintos ("Aa"). Llamamos fenotipo a la manifestación externa del genotipo, es decir, a los rasgos observables en un individuo. Es el resultado de la interacción entre el genotipo y el ambiente. La manifestación de algunos rasgos depende solo de un par de alelos, pero otros dependen de varios genes. En el contexto de estudios de asociación genética que nos ocupa, el fenotipo suele referirse a si un individuo está o no afectado por una determinada enfermedad. Lo denominaremos Y, siendo $Y = 1$ enfermos, $Y = 0$ no enfermos.

Un polimorfismo genético son los múltiples alelos de un gen entre una población, normalmente expresados como diferentes fenotipos. Es la variación en la secuencia de un *locus* determinado del ADN entre los individuos de una población. Esta variación debe estar presente en al menos un 1 % de la población para poder considerarse polimorfismo. Aunque la mayor parte de ellos no tienen efecto sobre el fenotipo, algunos son responsables de las enfermedades genéticas. Los polimorfismos objeto de estudio más habituales en los estudios de asociación son los polimorfismos de nucleótido único o SNPs (*Single Nucleotide Polymorphysms*). Los SNPs son variaciones en el genoma en las que un único nucleótido es sustituido por otro, sin afectar a los nucleótidos colindantes.

Por otro lado, un haplotipo es un conjunto de polimorfismos que tienden

a heredarse juntos, se puede referir a una combinación de alelos o a un conjunto de SNPs que se encuentran en el mismo cromosoma.

Modelos genéticos

Los modelos genéticos describen la relación entre el genotipo de un individuo y un rasgo específico. Un parámetro comúnmente utilizado para describir los modelos genéticos para un rasgo binario es la penetrancia: la probabilidad de un fenotipo particular Y (p.e. la presencia de enfermedad) condicionada al genotipo del individuo.

Los modelos genéticos más comunes son:

- **Modelo codominante.** Cada genotipo aporta un riesgo de enfermedad diferente. $P(Y = 1|AA) \neq P(Y = 1|Aa) \neq P(Y = 1|aa)$
- **Modelo dominante.** En estos casos es suficiente con que haya una copia del alelo de riesgo para modificar el riesgo de enfermedad. Suponiendo que el alelo de riesgo sea A, $P(Y = 1|AA) = P(Y = 1|Aa)$.
- **Modelo recesivo.** Hacen falta dos copias del alelo de riesgo para modificar el riesgo de enfermedad. $P(Y = 1|Aa) = P(Y = 1|aa)$.
- **Modelo aditivo.** El riesgo asociado al heterocigoto es intermedio entre el asociado a cada homocigoto.

Desequilibrio de ligamento

El desequilibrio de ligamento (LD, del inglés *Linkage Disequilibrium*) se refiere a la asociación entre alelos de diferentes *loci* en un cromosoma. Hace referencia a la disposición no causal de alelos en dos *loci*, de manera que estos dos alelos serán heredados conjuntamente. Este concepto será muy útil en los estudios de asociación del genoma completo, como explicaremos más adelante. No será necesario genotipar dos SNPs que estén en alto LD, ya que el genotipo de los alelos de un SNP se predecirá del otro.

Enfermedades mendelianas y enfermedades complejas

Las enfermedades monogenéticas mendelianas son las causadas por la mutación de un único gen. Las enfermedades complejas son las causadas por el efecto de varios genes y de la interacción entre ellos (epístasis) y con el ambiente. En estos casos, la presencia de un gen defectuoso no causa necesariamente la enfermedad, pero modifica el riesgo de padecerla. Los modelos genéticos definidos anteriormente son determinísticos en el caso de las enfermedades mendelianas y probabilísticos para las complejas. Es decir, para las enfermedades complejas el genotipo influye en la probabilidad de desarrollar

una enfermedad.

Los estudios de asociación pretenden localizar los genes y sus polimorfismos que contribuyen a padecer una enfermedad compleja en diferentes ambientes. Para ello, se asume la hipótesis de "enfermedades comunes-variantes comunes", que indica que las enfermedades complejas comunes son la consecuencia del efecto combinado de muchos polimorfismos frecuentes en la población.

Los resultados obtenidos hasta el momento en esta materia solamente han permitido identificar una pequeña parte del riesgo de enfermedad, lo que lleva a los expertos a pensar que tal vez la hipótesis de "enfermedades comunes–variantes comunes" no sea correcta y que las enfermedades complejas sean en realidad fruto de variantes infrecuentes [1].

1.2. Tipos de estudios genéticos

Los métodos más comunes a la hora de identificar y localizar los genes causantes de enfermedades son los análisis de ligamento genético y los estudios de asociación. Los análisis de ligamento genético han demostrado ser exitosos en lo que se refiere a analizar enfermedades mendelianas, pero no han sido de gran ayuda para detectar genes asociados a enfermedades complejas [2]. Es por ello que en este trabajo nos centraremos en los estudios de asociación.

1.2.1. Estudios de asociación

Estudian la relación entre el genotipo y el fenotipo, siendo este último enfermo o no enfermo. Lo hacen buscando los *loci* para los cuales los alelos varían entre los individuos sanos y los enfermos. Diremos que un *locus* está asociado con una cierta enfermedad si distintos genotipos tienen distribuciones estadísticamente diferentes en los individuos enfermos y en los que no lo están.

Las mayores ventajas de estos estudios frente a los anteriores son la capacidad de detectar los pequeños efectos de las variantes comunes y la posibilidad de utilizar muestras de individuos no relacionados en lugar de familias, lo cual permite que las muestras sean de mayor tamaño, aumentando el poder de detección. Dentro de los estudios de asociación también hay diferentes modelos dependiendo del número de SNPs estudiados y de las hipótesis previas:

Estudios de genes candidatos: consisten en aplicar conocimientos biológicos o estudios anteriores para escoger aquellos genes sobre los que exista cierta evidencia de que puedan estar relacionados con una enfermedad. Se

estudiarán entre 5 y 50 SNPs pertenecientes a cada gen.

Genome Wide Association Studies (GWAS): son estudios de asociación de todo el genoma que pretenden determinar qué SNPs se asocian con la enfermedad, sin la necesidad de establecer ninguna hipótesis previa. Estos estudios son posibles gracias a los avances hechos en técnicas de secuenciación, que han permitido que secuenciar el genoma sea un proceso menos costoso tanto en términos temporales como económicos. La ausencia de hipótesis previas permite examinar todas las diferencias existentes en el genoma, lo cual aumenta la probabilidad de encontrar los genes relacionados con la enfermedad.

Una vez encontramos un SNP asociado con la enfermedad, pueden ocurrir diferentes cosas: que el SNP esté realmente asociado con la enfermedad, que los resultados sean erróneos por problemas de la estratificación poblacional o que el SNP encontrado esté en desequilibrio de ligamento con algún SNP realmente relacionado con la enfermedad.

Recalcamos la importancia del concepto de LD en los estudios de asociación, ya que permiten que no sea necesario genotipar la variante causal para detectar la asociación. La Figura 1.2 ilustra lo anterior: aunque no seamos capaces de detectar las dos asociaciones directas, si ambos *loci* están en alto LD, probablemente podamos detectar la asociación indirecta entre el SNP marcador y el fenotipo de enfermedad.

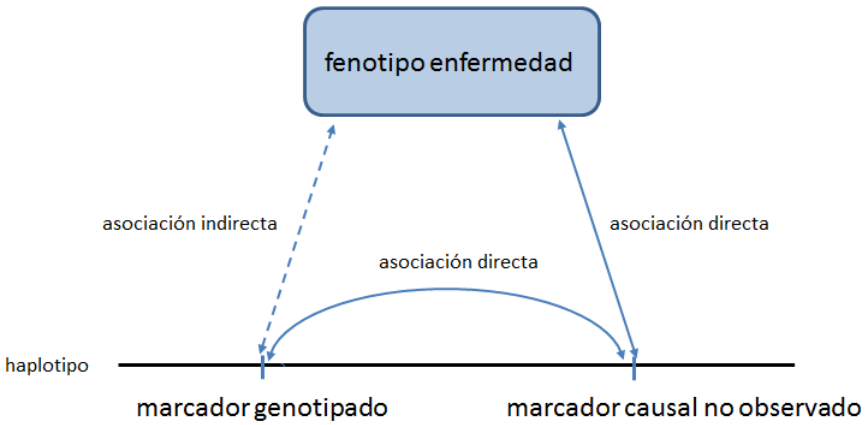


Figura 1.2: Detección de asociaciones indirectas.

Los GWAS presentan ciertas limitaciones, y hasta el momento no han da-

do los resultados que de ellos se esperaban. Algunos de los problemas que presentan son los siguientes:

- Hay un elevado número de variables predictoras en relación con el número de observaciones. Tengamos en cuenta que la cantidad de SNPs introducidos en el estudio será mayor que 300.000.
- Habrá casos en los que la variación genética tenga que interactuar con algún factor ambiental para dar lugar a la enfermedad, y otros en los que el riesgo de padecer una enfermedad dependa de la interacción gen–gen. Los GWAS permiten introducir este tipo de variables en el estudio, pero el número de interacciones a estudiar será muy elevado.
- Al explorar tantos SNPs, hay un alto riesgo de considerar influyentes variables que en realidad no lo son.
- La hipótesis de "enfermedades comunes-variantes comunes" en la que se apoyan estos análisis deja fuera del estudio variantes que aparecen en menos de un 1 % de la población. No podremos medir el efecto de estas variantes raras sobre la enfermedad.

Capítulo 2

Métodos estadísticos para datos genéticos

En este capítulo introduciremos los métodos estadísticos utilizados para detectar relaciones entre un SNP y la presencia o ausencia de enfermedad en el contexto de GWAS. Como ya hemos adelantado, uno de los mayores inconvenientes en los estudios del genoma completo es el alto número de variables a estudiar: cientos de miles de SNPs genotipados en muestras de miles de individuos.

Dado que la variable respuesta de nuestro estudio genético será dicotómica, en concreto:

$$Y = \begin{cases} 0 & \text{para individuos sanos} \\ 1 & \text{para individuos enfermos} \end{cases}$$

haremos uso del modelo logístico para estudiar la asociación entre variables predictoras y la variable respuesta Y.

2.1. Regresión logística

En primer lugar, vamos a introducir la regresión logística múltiple. Dadas una variable respuesta dicotómica Y y las variables predictoras $X_1, X_2, ..., X_p$, buscamos un modelo capaz de explicar la probabilidad $p(\mathbf{X}) = Pr(Y = 1|\mathbf{X})$, con $\mathbf{X} = (X_1, X_2, ..., X_p)$.

Aunque no podemos explicar la probabilidad con un modelo de regresión lineal $p(\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$, ya que toma valores en toda la recta real, sí que podemos modificarlo para obtener resultados en el intervalo $(0, 1)$. Usaremos el siguiente modelo al que denominamos "*logit*":

$$logit(p(\mathbf{X})) = ln\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p. \qquad (2.1)$$

Dada una muestra de tamaño n, podemos escribir el modelo como:

$$logit(p(\mathbf{X} = \mathbf{x}_i)) = ln\left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$

Para estimar los parámetros β_j con $j \in \{1, 2, ..., p\}$ maximizaremos la función de verosimilitud, que en el caso que nos ocupa es:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{1-y_i}, \tag{2.2}$$

cuyo resultado es el mismo que el obtenido al maximizar el logaritmo neperiano de la función de verosimilitud (función de log–verosimilitud) o minimizar la log–verosimilitud negativa, función a la que a partir de ahora denotaremos como L,

$$L(\boldsymbol{\beta}) = -\ln(l(\boldsymbol{\beta})). \tag{2.3}$$

Significatividad de las variables

Para estudiar si las variables predictoras son significativas se realiza un test de razón de verosimilitudes. Se quieren contrastar las hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 & \text{(modelo } M_0\text{)} \\ H_1 : \exists j \in \{1, 2, ..., p\}/\beta_j \neq 0 & \text{(modelo } M_1\text{)} \end{cases}$$

El estadístico que utilizamos para hacer el contraste es:

$$G = D(M_0) - D(M_1) \approx \chi^2_p,$$

donde D es la *deviance* del modelo: un estadístico que nos permite comparar el modelo estimado con el modelo completo (aquel que ajusta perfectamente los datos, con un parámetro por observación) y que se define como:

$$D = -2ln\left(\frac{\text{verosimilitud(m. estimado)}}{\text{verosimilitud(m. completo)}}\right).$$

Interpretación de los parámetros

La interpretación de estos parámetros se hace mediante el "*Odds Ratio*"(OR), definido como $OR = e^{\beta_j}$, y es la siguiente:

- X_j **variable dicotómica:** el OR representa el aumento de la probabilidad de éxito en individuos con $X_j = 1$ frente a aquellos con $X_j = 0$ cuando el resto de variables se mantienen constantes.
- X_j **variable continua:** OR indica cuánto aumenta el riesgo por unidad de aumento de X_j, cuando las demás variables se mantienen constantes.

Observamos que un valor del OR mayor que uno significa que el correspondiente factor aumenta el éxito, menor que uno lo disminuye y uno no afecta. Así, cuando el intervalo de confianza de un determinado OR contenga a 1, la variable correspondiente no será significativa.

Bondad de ajuste del modelo

Recordemos que el modelo logístico predice la probabilidad de que un individuo esté enfermo dadas unas ciertas características. Será necesario definir un punto de corte o umbral c a partir del cual consideremos que el individuo está enfermo. Es decir, habrá que decidir cuál es el valor de c tal que:

$$\begin{cases} p(\mathbf{x}_i) < c & \Rightarrow \hat{Y}_i = 0 \\ \\ p(\mathbf{x}_i) \geq c & \Rightarrow \hat{Y}_i = 1 \end{cases}$$

Dada una base de datos, dividiremos las observaciones en dos subconjuntos: el "conjunto de entrenamiento", con n_e observaciones y el "conjunto de test", con n_t observaciones, de forma que $n = n_e + n_t$. En general, la proporción de datos en el conjunto de entrenamiento será mayor que en el conjunto de test. Si utilizamos únicamente el primer conjunto para ajustar el modelo, podremos utilizar el conjunto de test para comprobar la capacidad de discriminación del modelo para observaciones fuera de muestra. Tras dividir la base de datos en dos subconjuntos, ajustaremos el modelo con el conjunto de entrenamiento. Aplicaremos el modelo ya ajustado al conjunto de test y compararemos los resultados obtenidos con los observados. Así obtendremos los valores de la Tabla 2.1, y a partir de ella calcularemos los valores de la sensibilidad y la especificidad.

	Estimados		
Observados	0	1	
0	Verdadero Negativo	Falso Positivo	VN+FP
1	Falso Negativo	Verdadero Positivo	FN+VP
	VN+FN	FP+VP	n_t

Tabla 2.1: Tabla de clasificación en la muestra de test

- Sensibilidad: $Se = \dfrac{VP}{VP + FN}$, la proporción de éxito clasificados correctamente.
- Especificidad: $Es = \dfrac{VN}{FP + VN}$, la proporción de fracasos clasificados correctamente.

Observamos que los valores de la tabla cambiarán con el punto de corte c elegido.

A continuación, definiremos una medida basada en la sensibilidad y la especificidad denominada curva ROC (*Receiver Operating Characteristics*). Considerando "Se" como eje de ordenadas y "1 - Es" como eje de abscisas, podemos marcar la sensibilidad y especificidad correspondientes a cada posible valor del punto de corte c. Si unimos todos esos puntos, obtendremos la denominada "curva ROC".

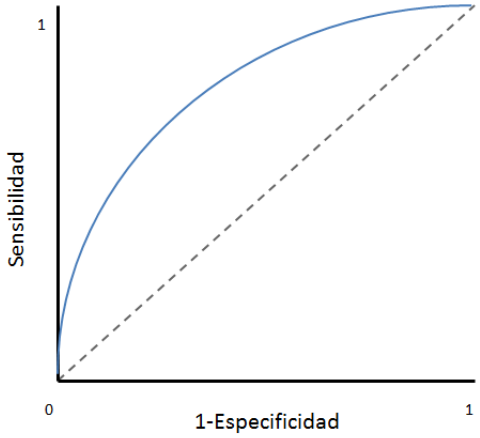


Figura 2.1: Ejemplo de curva ROC

Cuanto más arriba esté un valor en el espacio de la curva ROC, mayor será la sensibilidad, y por lo tanto, el valor del punto de corte c correspondiente dará lugar a menos falsos negativos. Cuanto más a la izquierda esté, mayor será la especificidad y menor el número de falsos positivos.

Podemos también definir el área bajo la curva ROC o AUC (*Area Under de Curve*), que mide la capacidad de discriminación del modelo. Observamos que este valor está entre 0 y 1. Un valor de AUC= 0,5 indica clasificación al azar, mientras que AUC= 1 significa que el clasificador es perfecto.
Es importante recordar que hemos construido la curva ROC en base a una muestra, por lo que no es si no una estimación de la curva real.

La Figura 2.2 ilustra el proceso para medir la capacidad de discriminación de un modelo logístico.

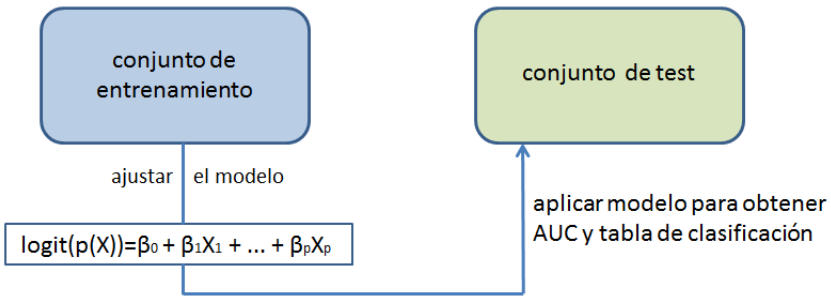


Figura 2.2: Proceso para medir capacidad de discriminación de un modelo logístico

2.2. Análisis simple: asociación entre un SNP y la enfermedad

Podemos analizar la asociación de cada SNP con la enfermedad de manera independiente, aunque estaremos perdiendo información interesante proveniente de la interacción entre SNPs.

Para cada SNP, queremos contrastar las siguientes hipótesis:

$$\begin{cases} H_0 : \text{no está relacionado con la variable respuesta} \\ \\ H_1 : \text{está relacionado con la variable respuesta} \end{cases}$$

Podemos presentar la información genética de cada SNP mediante una tabla de contingencia 2×3 como la siguiente:

	aa	Aa	AA
Enfermos	r_0	r_1	r_2
No enfermos	s_0	s_1	s_2

El modelo logístico correspondiente será:

Dada Y la variable respuesta definida como

$$Y = \begin{cases} 0 & \text{para individuos sanos} \\ 1 & \text{para individuos enfermos} \end{cases}$$

y X una variable discreta que puede tomar las modalidades AA, Aa y aa, podemos definir las variables indicadoras:

$$X_{AA} = \begin{cases} 1 \ si \ X = AA \\ 0 \ si \ X \neq AA \end{cases} \qquad \text{y} \qquad X_{Aa} = \begin{cases} 1 \ si \ X = Aa \\ 0 \ si \ X \neq Aa \end{cases}$$

que indican la pertenencia a la modalidad señalada, tomando como referencia el genotipo "aa".

El modelo de regresión logística es:

$$logit(p(X)) = \beta_0 + \beta_{AA} X_{AA} + \beta_{Aa} X_{Aa}. \tag{2.4}$$

Tras estimar los valores de los diferentes β por el método de máxima verosimilitud, podemos realizar un test de razón de verosimilitudes para estudiar si la covariable X es significativa.

2.3. Análisis de múltiples SNPs

Aunque el estudio anterior puede ofrecernos información a cerca de las variables relacionadas con una enfermedad, es conveniente considerar conjuntamente la asociación entre varios SNPs y la misma, así como introducir covariables e interacciones. Sin embargo, este tipo de consideraciones acarrean un gran problema al que tendremos que dar solución: hay un elevado número de variables, muy superior al de observaciones ($n << p$). Es decir, estamos ante un problema de alta dimensión.

Los métodos de regresión habituales no son válidos para analizar este tipo de problemas, ya que generarían unos coeficientes que ajustarían muy bien los datos, dando lugar a un sobreajuste y una gran variabilidad: pequeños cambios en los datos darían lugar a grandes cambios en el modelo.
Vamos a estudiar procesos alternativos para ajustar el modelo logístico evitando el sobreajuste y mejorando la interpretabilidad del modelo. Con la interpretabilidad del modelo nos referimos a que en un modelo con tantas variables predictoras, habrá muchas que estén poco relacionadas con la variable respuesta y será más fácil interpretar un modelo en que solo se tengan en cuenta las que influyen en la respuesta.

Una manera de afrontar el problema es aplicar técnicas de selección de variables como "*Best Subset Selection*" o "*Stepwise Selection*", que consisten en identificar un subconjunto de las variables predictoras relacionadas con la respuesta y ajustar el modelo para ese subconjunto [4, capítulo 6.1]. Sin embargo, estas técnicas sufren a menudo de alta variabilidad. Otros métodos disponibles son las técnicas de regularización o *shrinkage*, que ajustan el modelo con todos los predictores, pero fuerzan los coeficientes predichos hacia cero. Además, veremos que hay métodos de regularización que ayudan a la selección de variables.

2.3.1. Métodos de regularización

En nuestro contexto de GWAS, estamos interesados en ajustar un modelo de regresión logística como el descrito anteriormente en la Ecuación (2.1), pero evitando el sobreajuste proveniente de la gran cantidad de variables predictoras. Para ello, además de minimizar la log-verosimilitud negativa de la Ecuación (2.3), aplicaremos una penalización sobre los coeficientes β_j, controlada a través de un parámetro de penalización. Explicaremos a continuación las técnicas de regresión *ridge*, *lasso* y *elastic net* en el contexto de regresión logística.

Regresión Ridge

Fue introducida en 1970 por Hoerl y Kennard [5]. Dado el Modelo (2.1), los coeficientes $\hat{\boldsymbol{\beta}}^{ridge} = (\hat{\beta}_0^{ridge}, ..., \hat{\beta}_p^{ridge})$ estimados se obtienen de minimizar la función:

$$L_{ridge} = L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2.5}$$

donde $\lambda \geq 0$ es el parámetro de penalización y determina la fuerza de penalización ($\lambda = 0$, no hay penalización, $\lambda \to \infty$, todos los parámetros se contraen hacia cero).
La regresión *ridge* busca coeficientes que se ajusten bien a los datos, minimizando la log–verosimilitud negativa. Pero el segundo término, $\lambda \sum_j \beta_j^2$, llamado término de penalización, tiene el efecto de contraer los coeficientes β_1,...,β_p hacia cero, ya que la penalización es menor cuanto más cercanos a cero sean los β_j. Observamos que no imponemos penalización al coeficiente β_0, ya que no muestra la relación entre una variable concreta y la respuesta. Cada valor de λ dará lugar a un conjunto de coeficientes estimados $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$, y elegir el λ adecuado será crucial.

Regresión Lasso

La regresión *lasso* (*least absolute shrinkage and selection operator*) fue introducida por Tibshirani en 1996 [6] y es una manera alternativa de ajustar el modelo logístico (2.1). En este caso estimamos los coeficientes $\hat{\boldsymbol{\beta}}^{lasso} = (\hat{\beta}_0^{lasso}, ..., \hat{\beta}_p^{lasso})$ minimizando la función

$$L_{lasso} = L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{2.6}$$

El término de penalización en este caso es $\lambda \sum_j |\beta_j|$, es decir, el método *lasso* usa una restricción de norma l_1 mientras el método *ridge* usa una l_2.

Esta técnica no solo contrae los coeficientes hacia cero, si no que fija algunos en cero si el valor de λ es adecuado. Así, *lasso* hace también selección de variables, dando lugar a modelos más fáciles de interpretar y mostrando qué variables tienen mayor relación con la variable respuesta. Dependiendo del valor de λ tendremos modelos con distinto número de variables.

Regresión Elastic Net

Propuesto en 2005 por Zou y Hastie [7], es un método que combina la penalización l_1 de *lasso* y la l_2 de *ridge*. Como ya hemos mencionado anteriormente, en los GWAS podemos encontrar SNPs que aunque no estén fuertemente relacionados con la enfermedad, estén en desequilibrio de ligamento con SNPs influyentes. En este contexto, es interesante aplicar una técnica de regresión capaz de representar a todos los SNPs asociados con el SNPs causante, para poder estudiarlos más adelante meticulosamente.

La penalización *lasso* escoge una sola de las variables altamente correladas y descarta el resto. La penalización *ridge* tiene una propiedad de agrupamiento que tiende a contraer los coeficientes de las variables correladas conjuntamente, generando para ellas coeficientes estimados parecidos. El método que presentamos a continuación tiene un término de penalización que combina la norma l_1, que dará lugar a selección de variables con la norma l_2, con la que mantendremos la propiedad de agrupamiento del método *ridge* ([8], [14]).

Los coeficientes $\hat{\boldsymbol{\beta}}^{en} = (\hat{\beta}_0^{en}, ..., \hat{\beta}_n^{en})$ estimados se obtienen minimizando la función:

$$L_{en} = L(\beta) + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha \left|\beta_j\right| \right]. \tag{2.7}$$

Recordando que $||\beta||_1 = \sum_j |\beta_j|$ y $||\beta||_2^2 = \sum_j \beta_j^2$, podemos escribir la ecuación anterior como:

$$L_{en} = L(\beta) + \lambda \left(\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right).$$

La penalización *elastic net* es equivalente a la penalización *lasso* cuando $\alpha = 1$ y a la penalización *ridge* cuando $\alpha = 0$.

2.3.2. Elección del parámetro de penalización

La elección del parámetro de penalización λ es crucial cuando aplicamos esta clase de métodos, ya que determina la magnitud de los coeficientes por un lado, y el número de variables que formarán parte del modelo por el otro.

La manera habitual de proceder es mediante validación cruzada: seleccionamos primero unos posibles valores para λ, computamos el criterio de validación cruzada para cada uno de ellos y seleccionamos el mejor en base al criterio. Por último, reajustamos el modelo utilizando todas las observaciones y el λ seleccionado.

Validación cruzada

La validación cruzada es una técnica que nos permite validar parámetros estadísticos. Podemos utilizarla para comparar distintos métodos o un mismo método con diferentes niveles de flexibilidad, y seleccionar el más adecuado. Consiste esencialmente en dividir el conjunto de observaciones en dos subconjuntos: el conjunto de entrenamiento, con el que ajustamos el modelo y el conjunto de validación, con el que validamos el parámetro estadístico considerado, computando un criterio previamente establecido. Un criterio común en problemas de clasificación con respuesta binaria es el de minimizar la *deviance* de los residuos.

Uno de los procedimientos de validación cruzada más usuales es el "*k-fold cross-validation*", que se aplica como sigue: en primer lugar, dividimos las observaciones en k grupos de igual tamaño de forma aleatoria. Ajustamos el modelo con los datos de todos los grupos excepto el primero y aplicamos el modelo a este primer grupo, obteniendo el valor Dev_1, que es el valor de la *deviance* de los residuos para los datos del primer grupo. Volvemos a ajustar el modelo, pero tomando ahora como conjunto de entrenamiento las observaciones de todos los grupos excepto el segundo. Utilizamos el segundo grupo como conjunto de validación para obtener el valor Dev_2. Repetimos este proceso k veces, usando cada vez un grupo como conjunto de validación. En la Figura 2.3.2 se ilustra el proceso anterior. El valor de k más utilizado es $k = 10$.

Por último, computamos la *deviance* de validación cruzada:

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} Dev_i \tag{2.8}$$

En el momento de decantarnos por un método o un nivel de flexibilidad, elegiremos aquel para el que el valor CV_k sea menor.

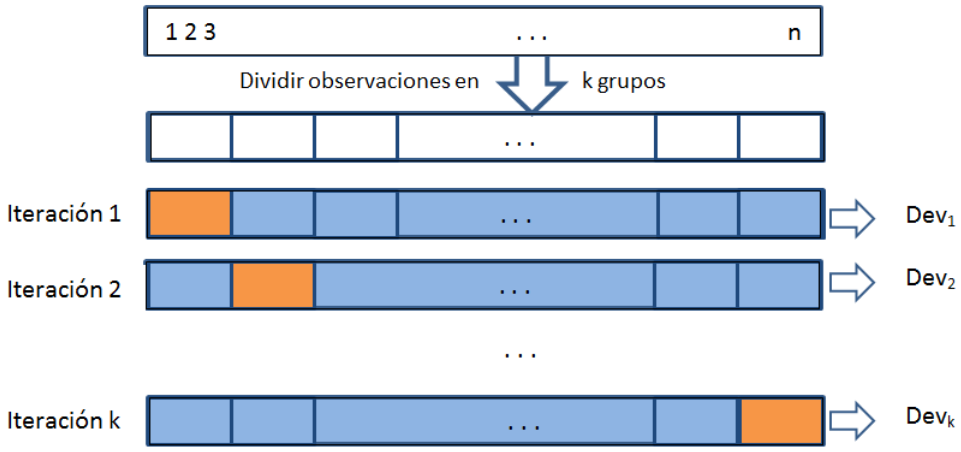


Figura 2.3: Esquema del proceso *k-fold cross-validation*. En azul, los conjuntos de entrenamiento de cada iteración y en naranja los de validación.

Capítulo 3

Aplicación de métodos de regularización a datos genéticos

En este capítulo presentaremos un estudio realizado sobre una base de datos genéticos con el objetivo de trabajar con los tres métodos de regresión penalizada propuestos en el capítulo anterior y comparar sus propiedades. Realizaremos un estudio de asociación en una región genética relacionada con la enfermedad de estudio. Nuestro objetivo será identificar las variables genéticas que influyen o están relacionadas con la enfermedad, usando para ello técnicas de regresión penalizada.

3.1. Descripción de los datos

En este estudio hemos usado los datos utilizados por Ayers y Cordell [8]. Éstos consisten en observaciones de 2000 individuos, la mitad de los cuales están afectados por una determinada enfermedad. Para cada individuo, se dispone del genotipo correspondiente a 228 SNPs de la región genética CTLA4 [*]. Cabe destacar que como Ayers y Cordell simularon los datos, conocemos la posición de los verdaderos SNPs causales. A continuación mostramos un extracto de la base de datos, que cuenta con 2000 observaciones de 129 variables. Cada una de las variables representa lo siguiente:

- La variable respuesta "Fenotipos" es dicotómica y toma el valor 0 para individuos sanos y el valor 1 para los afectados por la enfermedad.

[*]Región genética codificadora de proteinas. Las mutaciones en esta región han sido asociadas a la diabetes mellitus insulina-dependiente, la enfermedad de Graves, la tiroiditis de Hasmimoto, la enfermedad celiaca, el lupus eritematoso sistémico y otras enfermedades autoinmunes. http://www.iqb.es/reumatologia/fichas/ctla.html

- Las demás 128 variables predictoras corresponden a los *loci* de cada uno de los SNPs considerados. Están codificadas como 0, 1, ó 2 según el número de alelos de referencia.

Fenotipos	V1	V2	V3	V4	V5	V6	V7	V8	V9	...
1	0	0	0	0	0	0	1	0	0	...
0	0	0	0	0	1	0	2	0	0	...
0	0	0	0	0	0	0	2	0	0	...
1	2	1	2	2	0	1	2	2	2	...
1	2	2	2	2	0	0	2	2	2	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Tabla 3.1: Extracto de la base de datos

Los SNPs causales se corresponden con las variables $V14$, $V46$, $V98$, $V164$ y $V176$, que influirán conjuntamente en el riesgo de padecer la enfermedad.

Además de los datos de los genotipos y los fenotipos, contamos con un archivo que almacena los p-valores correspondientes a un test simple para estudiar la asociación individual de cada SNP con la enfermedad. Hemos representado gráficamente en la Figura 3.1 el valor negativo del logaritmo neperiano de cada uno de los p-valores, así como las posiciones de los verdaderos SNPs causales (lineas verticales).

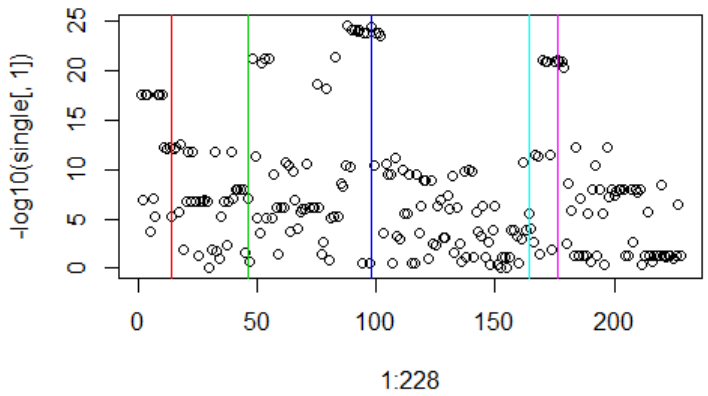


Figura 3.1: Representación de los p-valores correspondientes a cada SNP.

Los elementos más relacionados con la enfermedad de manera individual, es decir, aquellos con menor p-valor, serán aquellos situados en la parte

más alta del eje ordenado. Pero hay muchos valores muy significativos y es difícil distinguir que hay 5 *loci* causales y cuáles son. Esta observación nos lleva a plantearnos el uso de los métodos de regularización para localizar las posiciones o *loci* de los SNPs relacionados con la enfermedad.

3.2. Estudio del problema mediante métodos de regularización

Para aplicar los métodos de regularización estudiados en el capítulo anterior, haremos uso del paquete de R *glmnet*, que permite ajustar modelos logísticos por máxima verosimilitud penalizada. Bastará con indicar el valor de α deseado, siendo α el parámetro de la penalización elastic net:

$$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1$$

para aplicar cada una de las técnicas de penalización. En el Apéndice A mostramos algunos de los comandos que hemos utilizado.

Vamos a realizar el estudio en tres partes:

- En primer lugar, aprovecharemos que conocemos las posiciones de los verdaderos SNPs causales para comparar los coeficientes estimados por cada una de las técnicas. En este caso podremos escoger el parámetro de penalización manualmente, basándonos en nuestro conocimiento de los *loci* causales.
- Como dichas posiciones serán en general desconocidas, tendremos que usar validación cruzada para escoger un valor adecuado para el parámetro λ. En la segunda parte escogeremos el parámetro de penalización por validación cruzada y compararemos la capacidad discriminatoria de los tres métodos de regularización.
- Por último, haremos una comparativa de resultados en función del parámetro α.

3.2.1. Comparación de coeficientes para los tres métodos

En esta sección trabajamos con la función *glmnet*, que ajustará nuestro modelo 100 veces, para 100 valores de penalización λ diferentes. La secuencia de valores de λ es dependiente del método.

Regresión lasso

Fijamos $\alpha = 1$ y obtenemos los resultados que se muestran a continuación. En la Figura 3.2 se muestra la evolución del vector de coeficientes respecto

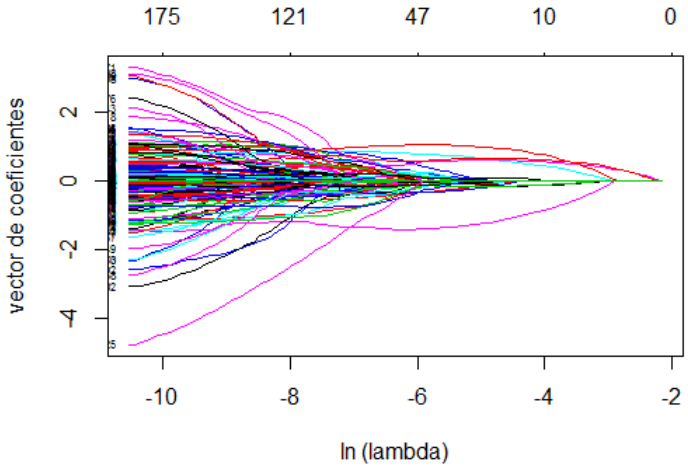


Figura 3.2: Coeficientes del método en función del nivel de penalización

a los valores de $\ln(\lambda)$. El gráfico nos permite observar distintos aspectos interesantes como por ejemplo:

- Los coeficientes de qué variables son los primeros en tomar el valor cero.
- De qué manera disminuye el número de variables consideradas por el modelo a medida que aumenta la penalización.
- Cómo disminuye también el tamaño de los coeficientes.

Escogemos seis de los cien valores de λ considerados por *glmnet* y mostramos en la Figura 3.3 los coeficientes estimados por el método para cada uno de ellos. Además de eso, indicamos con líneas horizontales la posición de los verdaderos marcadores causales. Observamos que a medida que λ decrece (la penalización sobre los coeficientes disminuye), hay más variables que entran en el modelo y además los coeficientes de las mismas crecen. Según apreciamos en la Figura 3.3, un valor de λ en torno a $0{,}02 - 0{,}05$ sería adecuado para estos datos, ya que el modelo consideraría un marcador para cada verdadera variable causal.

Eligiendo los parámetros de penalización $\lambda = 0{,}05034$ y $\lambda = 0{,}01986$, el modelo contará con 6 ó 10 variables respectivamente, como podemos observar en la Figura 3.4. Concretamente, las variables seleccionadas para los valores de λ considerados han sido las siguientes (marcamos con un asterisco los SNPs verdaderamente causales):

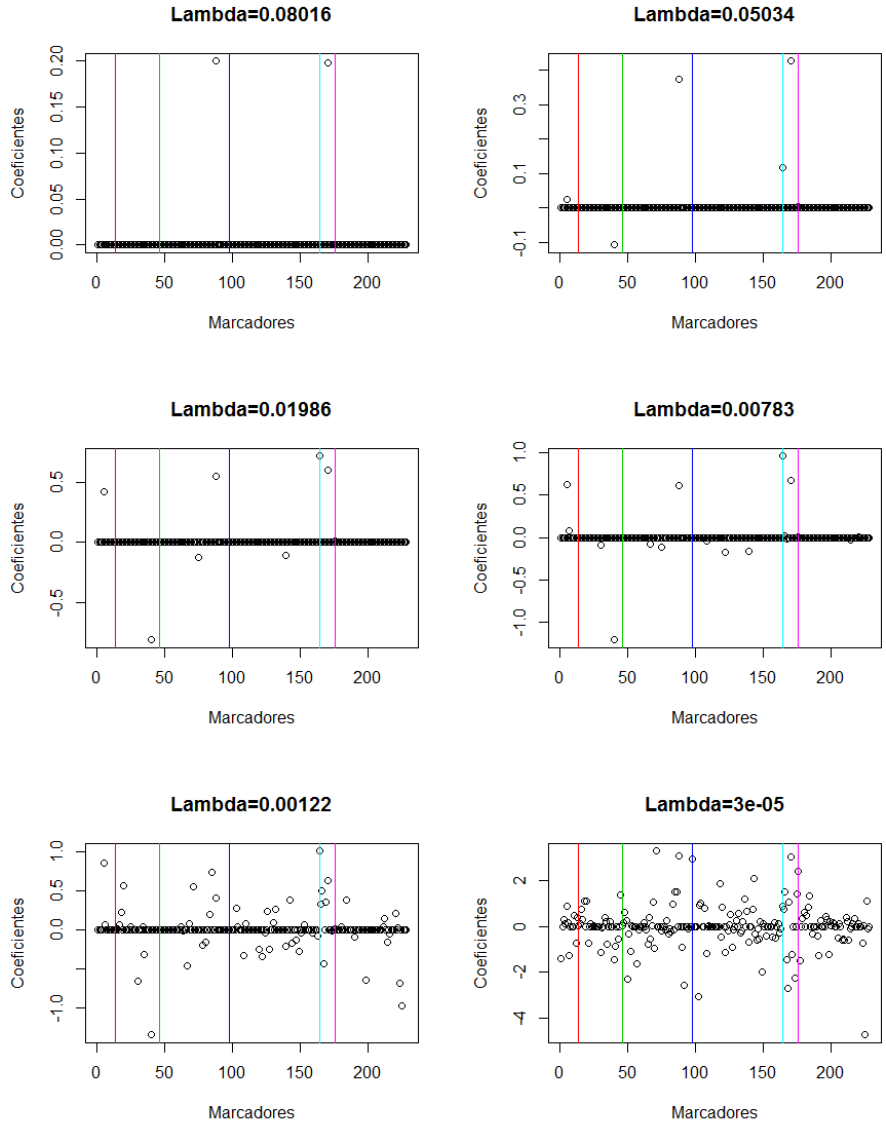


Figura 3.3: Coeficientes del modelo *lasso* para distintos valores de λ

- Para $\lambda = 0{,}05034$: $V5$, $V40$, $V88$, $*V164$, $V170$, $*V176$.
- Para $\lambda = 0{,}01986$: $V5$, $V7$, $*V14$, $V40$, $V75$, $V88$, $V139$, $*V164$, $V170$, $*V176$.

El método ha localizado 2 y 3 de las 5 variables causales respectivamente.

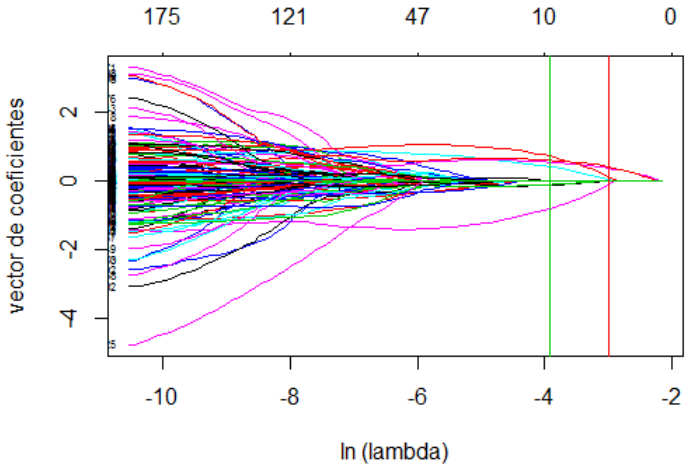


Figura 3.4: Número de coeficientes para regresión *lasso* con $\lambda = 0{,}05034$ y $\lambda = 0{,}01986$

Regresión ridge

En este caso, aunque nuestro parámetro de penalización haga que los coeficientes se contraigan a cero, no conseguiremos modelos dispersos (llamamos así a los modelos con muchos coeficientes iguales a cero), si no modelos con coeficientes muy pequeños.

La Figura 3.5 muestra los coeficientes correspondientes a seis valores distintos de λ. Se observa cómo para valores grandes de λ, algunos de los coeficientes son prácticamente nulos (observar la escala) y a medida que imponemos una menor penalización, los coeficientes crecen. Aunque obtengamos coeficientes muy pequeños no podemos considerar que el modelo sea disperso, por lo que será más difícil alcanzar nuestro objetivo de localizar las variables relacionadas con la enfermedad que utilizando el método *lasso*, que nos indicará directamente cuáles son las variables más convenientes a tener en cuenta.

Regresión elastic net

Elastic net admite cualquier valor de α entre 0 y 1, en este apartado escogeremos el valor $\alpha = 0{,}5$, dando la misma fuerza a la norma l_1 y a la l_2. Ajustamos el modelo con la función *glmnet* y considerando seis valores diferentes para la penalización λ.

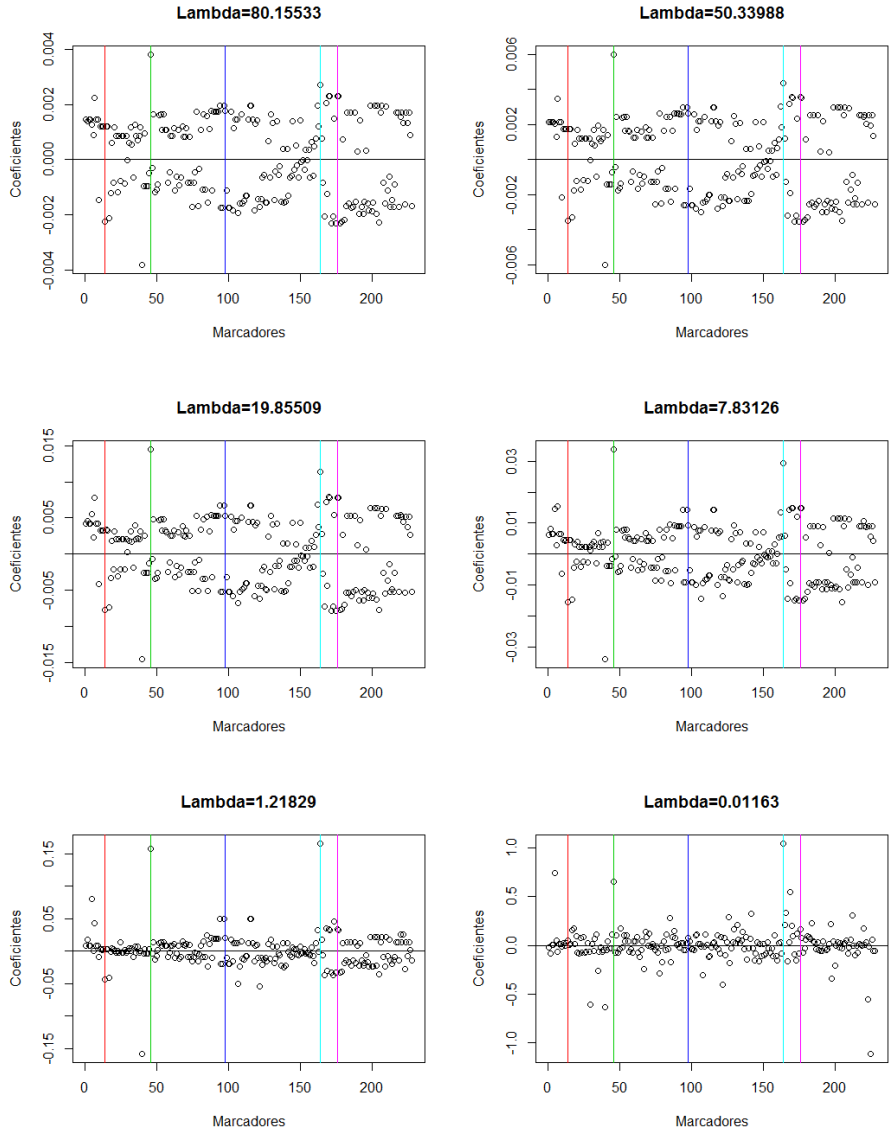


Figura 3.5: Coeficientes del modelo *ridge* para distintos valores de λ

Este método hace selección de variables, pero de una manera más moderada que *lasso*. Esta tendencia se hace visible comparando el número de coeficientes distintos de cero estimados para cada uno de los 100 valores λ por el método *lasso* y el método *elastic net* con $\alpha = 0{,}5$. En la Figura 3.6 mostramos el número de coeficientes no nulos estimados por ambos métodos en función de sus correspondientes secuencias de 100 valores de λ ordenados de mayor a menor penalización. El método *lasso* tiende a estimar más valores como nulos.

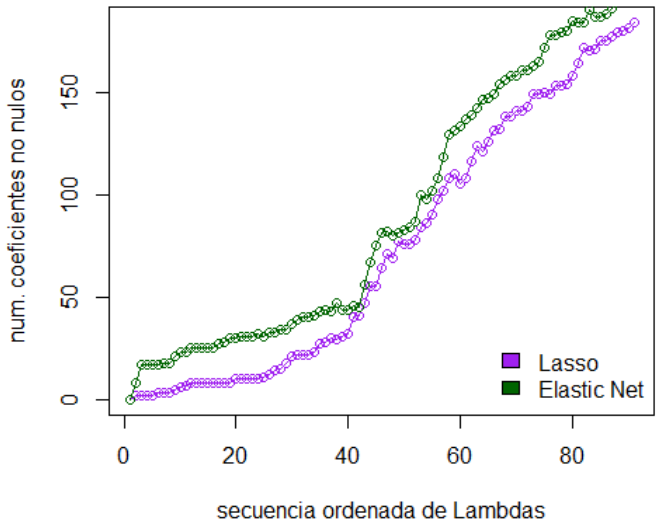


Figura 3.6: Número de coeficientes no nulos en función de las secuencias de los 100 valores de λ ordenados de mayor a menor penalización.

En la Figura 3.7 se muestran los coeficientes estimados por el método elastic net para seis valores diferentes de λ. De entre ellos, los que mejor parecen reconocer los *loci* causales son el segundo y el tercero. El modelo ajustado para $\lambda = 0{,}10068$ considera 23 de las 228 variables y es capaz de localizar 4 de los 5 SNPs causales. El modelo con $\lambda = 0{,}05249$ por su parte, considera 30 variables, localizando además todos los SNPs causales.

Comparación de los resultados

Tras analizar los coeficientes obtenidos con cada una de las tres técnicas de regularización, podemos concluir lo siguiente:

- Los métodos *lasso* y *elastic net*, al contrario que el método *ridge*, han realizado selección de variables, lo cual es muy beneficioso para nuestro objetivo de encontrar los SNPs relacionados con la respuesta.
- El método *elastic net* selecciona más variables que *lasso*. De esta manera, será más fácil seleccionar las variables influyentes en el fenotipo, pero a cambio estaremos incluyendo en el modelo más variables que,

aunque tengan relación con la enfermedad, no sean las variables causales buscadas.

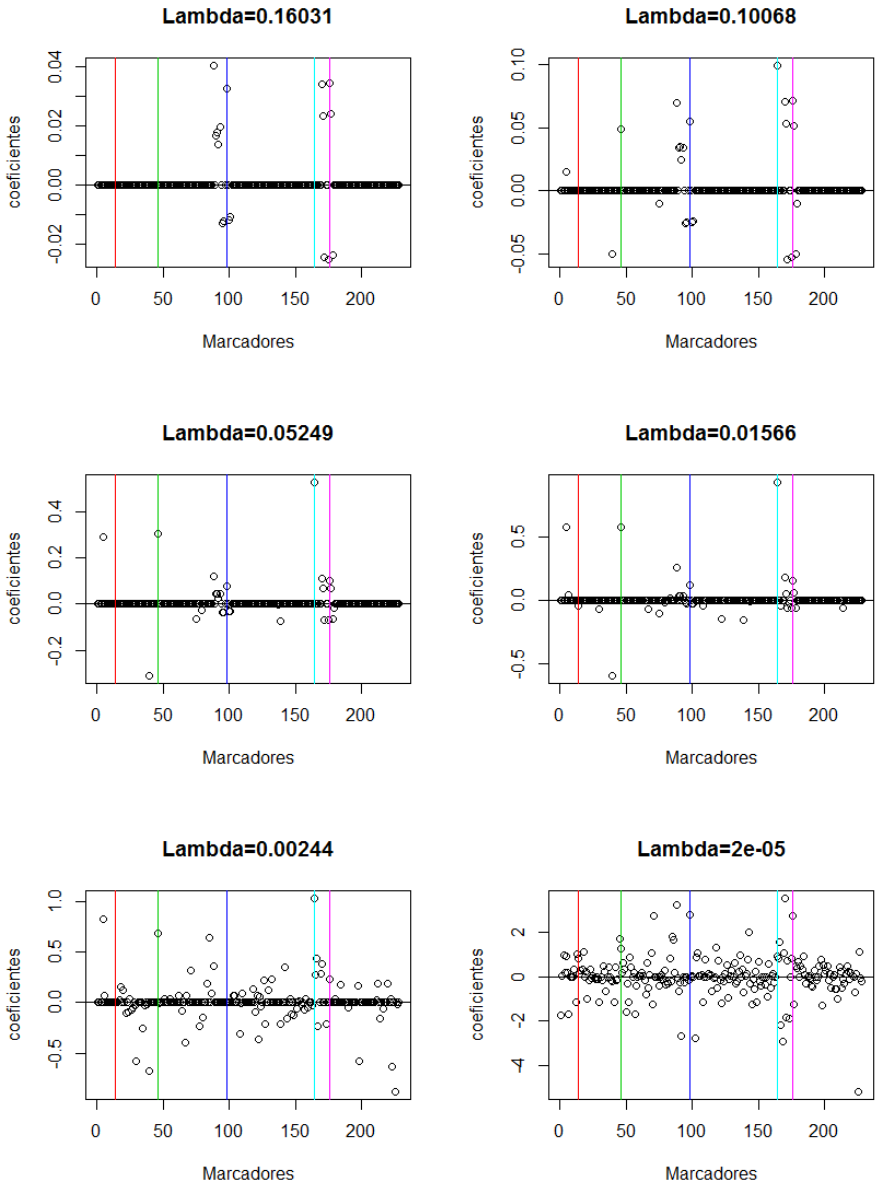


Figura 3.7: Coeficientes del modelo *elastic net* para distintos valores de λ

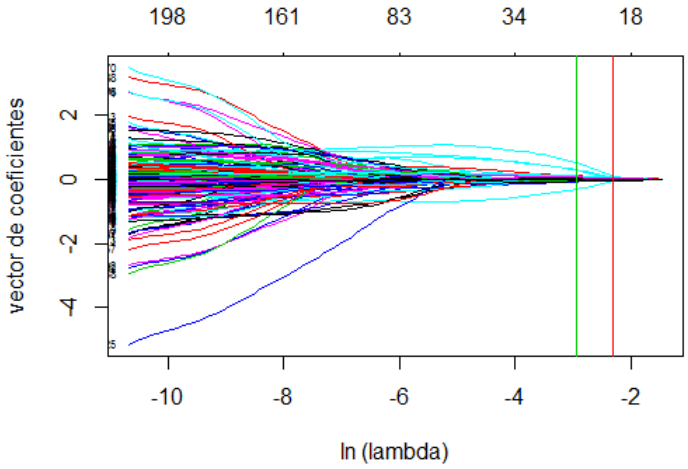


Figura 3.8: Número de coeficientes para regresión *elastic net* con $\lambda = 0{,}10068$ y $\lambda = 0{,}05249$

3.2.2. Selección de λ y comparación de la capacidad discriminatoria

En el apartado anterior hemos aprovechado que conocemos las posiciones de los SNPs causales para elegir manualmente un valor de λ que se adecue a nuestro problema. Sin embargo, éste no será el procedimiento habitual, ya que dichos *loci* serán desconocidos. En caso de estar buscando las posiciones de los SNPs relacionados con la enfermedad, elegiremos el parámetro de penalización mediante el método de validación cruzada. El paquete *glmnet* incluye una función que combina función *glmnet* que usábamos anteriormente con la validación cruzada: la función *cv.glmnet*. Esta función nos indicará el valor de λ más adecuado en el sentido de minimizar CV_{10} de la Ecuación (2.8) y ajustará el modelo por máxima verosimilitud penalizada.

En este apartado, además de aplicar las tres técnicas de regularización escogiendo el parámetro de penalización por validación cruzada, vamos a comparar los tres métodos fijándonos en las variables consideradas, los valores del AUC y las tablas de clasificación. Para ello, empezaremos por dividir los datos aleatoriamente, de manera que tengamos un conjunto de entrenamiento con 1500 observaciones y otro de test con las 500 restantes. Ajustaremos los modelos aplicando la función *cv.glmnet* a los datos del conjunto de entrenamiento y aplicaremos los modelos ajustados al conjunto de test para computar los valores del AUC y las tablas de clasificación. La Figura 3.9 muestra un esquema de dicho proceso.

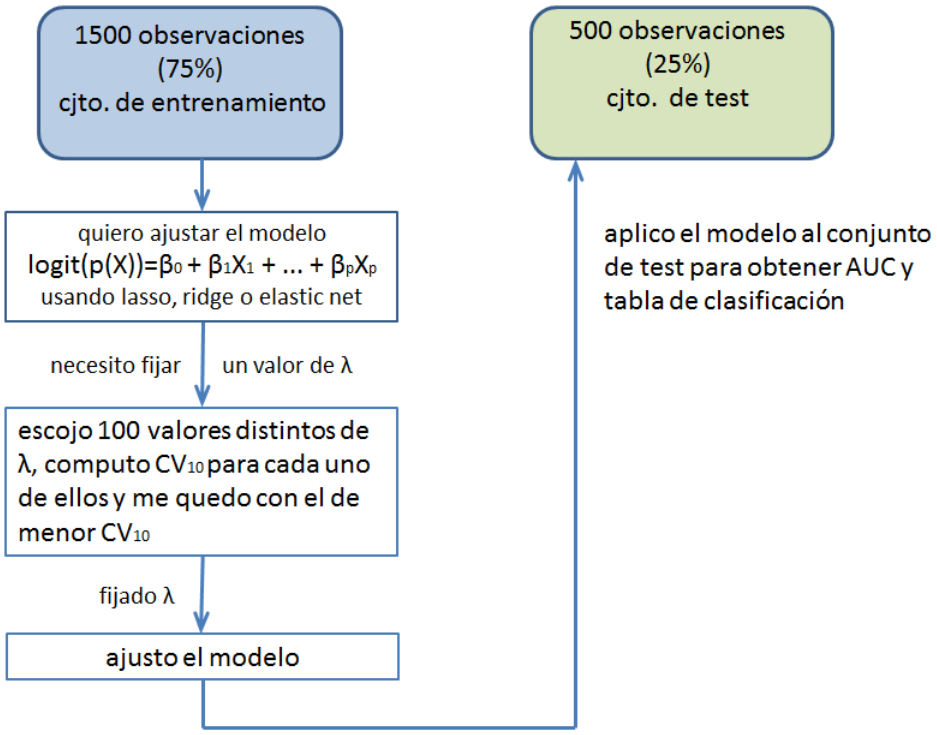


Figura 3.9: Proceso para comparar los métodos

Regresión lasso

Si llevamos a cabo ese proceso para $\alpha = 1$, es decir, para el método *lasso*, concluimos que el valor de α que minimiza la *deviance* de los residuos es $\lambda = 0{,}00922$. En la Figura 3.10 vemos como, en efecto, la menor *deviance* de validación cruzada CV_{10} se corresponde con el λ seleccionado.
Ajustando el modelo para dicho λ, obtenemos un modelo disperso cuyo vector de coeficientes $\hat{\boldsymbol{\beta}}^{lasso}$ tiene 15 elementos no nulos: el correspondiende a $\hat{\beta}_0$ y los correspondientes a las variables $V5$, $V30$, $V40$, $*V46$, $V67$, $V79$, $V88$, $V139$, $V160$, $*V164$, $V170$, $*V176$, $V214$ y $V220$. Es decir, tres de las variables consideradas por el modelo son SNPs verdaderamente causales. En la Figura 3.11 representamos los coeficientes estimados para cada variable.

Aplicando el modelo ya ajustado al conjunto de test, obtenemos la curva ROC de la Figura 3.11. En ella se muestran además el valor del AUC, el valor del umbral c que maximiza la suma de la sensibilidad y la especificidad y los valores de dichos estadísticos para ese punto de corte. La figura indica que el valor del AUC para el modelo logístico ajustado utilizando la técnica

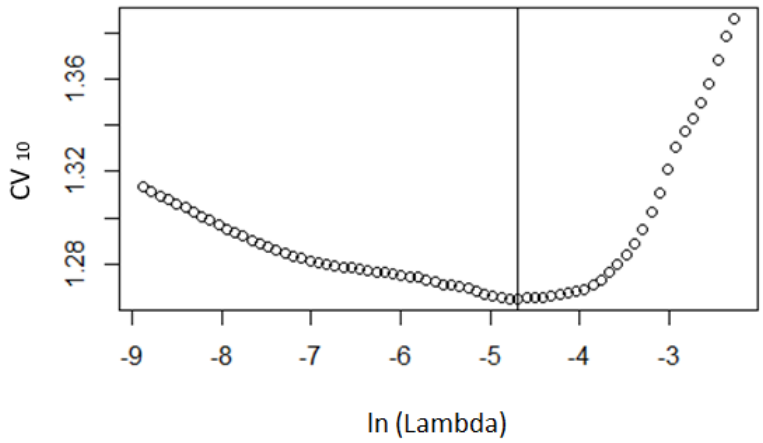
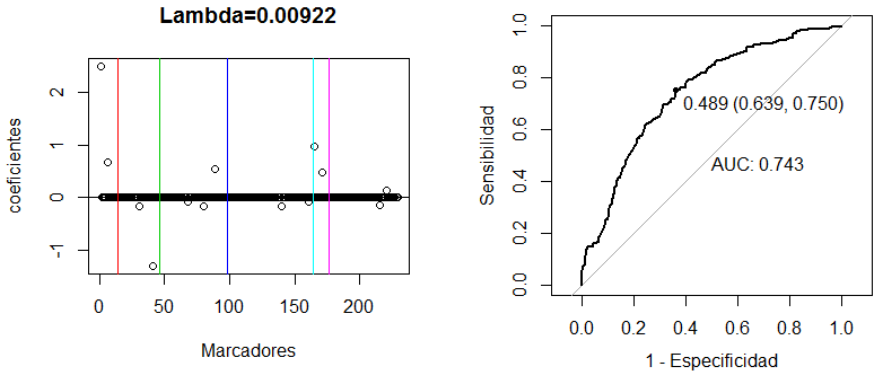


Figura 3.10: *Deviance* de validación cruzada en función de $ln(\lambda)$.



(a) Coeficientes (b) Curva ROC

Figura 3.11: Gráficos correspondientes a método *lasso* con $\lambda = 0{,}00922$

lasso para la penalización $\lambda = 0{,}00922$ es de 0,743. Si fijamos el umbral como $c = 0{,}489$, obtenemos los valores de la Tabla 3.2. Observamos que el modelo ha clasificado correctamente el 69,6 % de las observaciones fuera de muestra, con una sensibilidad de 0,750 y una especificidad de 0,639.

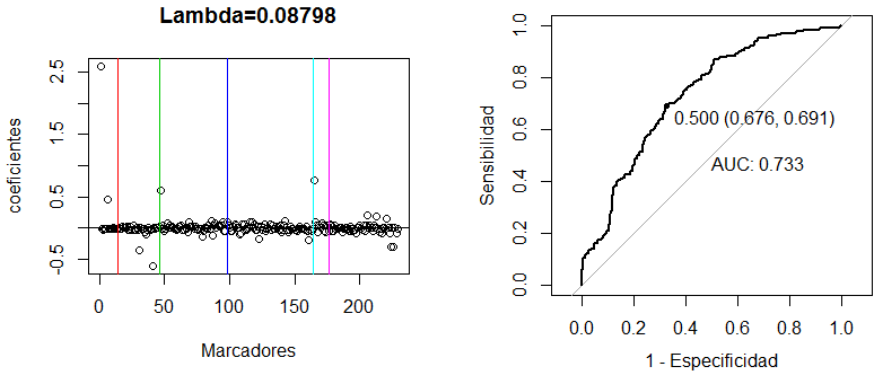
	$\hat{Y}$	
Y	0	1
0	156	88
1	64	192

Tabla 3.2: Tabla de clasificación para modelo *lasso* y $c = 0{,}489$

Regresión ridge

Repitiendo el procedimiento anterior con $\alpha = 0$, obtenemos un modelo que no realiza selección de variables, es decir, cuyo vector de coeficientes $\hat{\boldsymbol{\beta}}^{ridge}$ no tiene ningún elemento nulo. Aunque logremos un modelo con gran capacidad de predicción, tendremos complicaciones para decidir cuáles son las variables más relacionadas con la enfermedad, que es nuestro principal objetivo.

El valor óptimo para λ en este caso será 0,08798, dando lugar a un modelo cuyo AUC es de 0,733, algo menor que en el caso anterior.



(a) Coeficientes (b) Curva ROC

Figura 3.12: Gráficos correspondientes a método *ridge* con $\lambda = 0{,}08798$

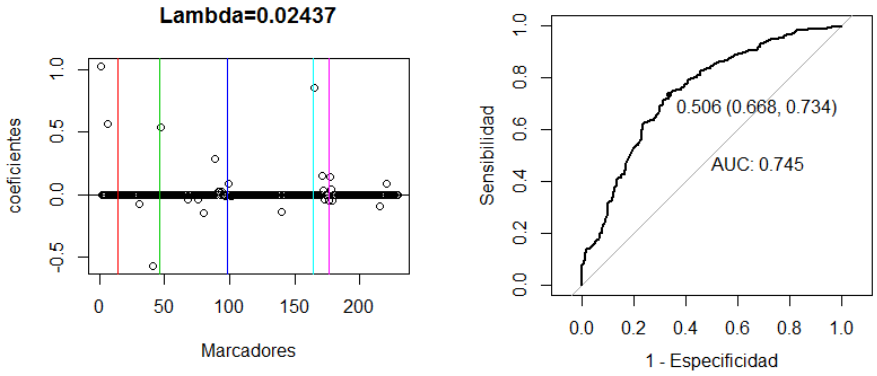
Lo ideal en este caso es considerar que un individuo está enfermo cuando la probabilidad de que lo esté supere el 0,5. La Tabla 3.3 muestra cómo es la clasificación en ese caso, con un 68,4 % de aciertos en la clasificación. Observamos que este modelo clasifica algo peor que el ajustado por *lasso*, se percibe sobre todo un aumento en los falsos negativos, es decir, en los casos en que predecimos que un individuo está sano y no lo está.

	$\hat{Y}$	
Y	0	1
0	165	79
1	79	177

Tabla 3.3: Tabla de clasificación para modelo *ridge* y $c = 0{,}5$

Regresión elastic net

Ajustamos ahora el modelo para $\alpha = 0{,}5$. En la sección 3.2.3 compararemos los resultados obtenidos para diferentes valores del parámetro α. El valor de λ más adecuado en este caso es 0,02437. En la Figura 3.13 se muestran los coeficientes estimados para cada marcador. En este caso, el modelo ha considerado 29 variables predictoras, localizando todos los SNPs causales excepto al correspondiente a $V14$.



(a) Coeficientes (b) Curva ROC

Figura 3.13: Gráficos correspondientes a método *elastic net* con $\lambda = 0{,}08798$

En cuanto a la capacidad predictiva del modelo, contamos con un AUC de 0,745. La tabla de clasificación al considerar $c = 0{,}506$ es la que se muestra en la Tabla 3.4 e indica que el modelo ha predicho correctamente un 70,2 % de los resultados fuera de muestra.

	$\hat{Y}$	
Y	0	1
0	163	81
1	68	188

Tabla 3.4: Tabla de clasificación para modelo *elastic net* y $c = 0{,}506$

Comparación de los resultados

Las conclusiones que podemos extraer después de ajustar los modelos para valores de penalización λ escogidos por validación cruzada son las siguientes:

- *cv.glmnet* nos ha proporcionado valores de λ muy adecuados con los que llevar a cabo los métodos de regularización sin necesidad de conocer previamente los marcadores causales.

- Las técnicas *lasso* y *elastic net* han dado lugar a modelos con una capacidad predictiva ligeramente mayor (0,743 y 0,745 frente a 0,733). Podemos observar esta diferencia en la Figura 3.14.

- Al igual que antes, los métodos *lasso* y *elastic net* han ayudado a la selección de variables, facilitando nuestro propósito de localizar los SNPs causales. Además, han mostrado eficiencia localizándolos.

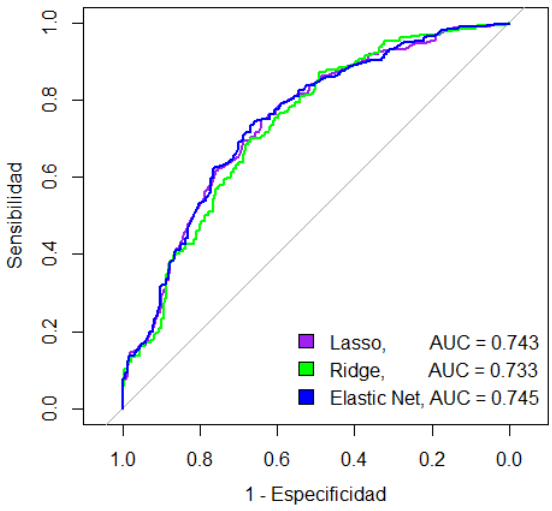


Figura 3.14: Curvas ROC para los distintos métodos.

Llegados a este punto, parece que el método *ridge* no servirá para cumplir nuestro propósito de localizar los SNPs relacionados con la enfermedad. Sin embargo, habrá que realizar un estudio más exhaustivo para decidir si el método *lasso* es el más apropiado para abordar el problema o si es preferible aplicar *elastic net*.

3.2.3. Comparativa de resultados en función del parámetro α seleccionado en el método elastic net

Vamos a comparar algunos aspectos de la regresión *lasso* y *elastic net* para diferentes valores de α. Para llevar a cabo esta parte del estudio hemos considerado los siguientes valores de α : 1 (equivalente a *lasso*), 0,75, 0,5, 0,3, 0,1, 0,05 y 0,01 (muy similar a *ridge*). A partir de la base de datos de 2000 observaciones, hemos generado aleatoriamente 100 subconjuntos de entrenamiento de tamaño $n = 1500$ y sus 100 correspondientes subconjuntos de test con 500 observaciones cada uno. Para cada uno de los siete valores de α considerados, hemos utilizado los subconjuntos de entrenamiento y la función *cv.glmnet* para seleccionar el parámetro de penalización λ por validación cruzada y ajustar el modelo. Así, hemos ajustado 100 modelos diferentes para cada valor del parámetro α.

Hemos calculado los siguientes valores para cada método considerado y los hemos recogido en la Tabla 3.5:

- Los AUC medios de cada modelo.
- La media y desviación típica del número de SNPs seleccionados por cada modelo en las 100 simulaciones. Consideramos que una variable ha sido seleccionada si su coeficiente β correspondiente es distinto de cero, sin embargo, éste puede ser un número muy cercano a cero.
- El número de veces que cada modelo ha seleccionado a cada uno de los SNPs causales a lo largo de todas las simulaciones.
- La media y la moda del número de SNPs causales localizados simultáneamente por cada modelo en las 100 simulaciones.
- El número de veces que cada modelo ha localizado todos los SNPs causales a la vez a lo largo de las simulaciones.
- El número medio de variables seleccionadas no causales.

Número de SNPs seleccionados

El número de variables (en este caso SNPs) predictoras considerado por el modelo aumenta a medida que α disminuye, es decir, a medida que la penalización de tipo l_1 pierde fuerza en favor de la de tipo l_2. Este aumento en el número de SNPs seleccionados se traduce también en un aumento de variables no causales seleccionadas: cuantas más variables consideremos, más variables poco relacionadas con la respuesta estaremos teniendo en cuenta.

α		**1**	**0,75**	**0,5**	**0,3**	**0,1**	**0,05**	**0,01**
AUC medio		0,707	0,707	0,706	0,705	0,703	0,702	0,700
Promedio de SNPs seleccionados y (sd)		17,50 (4,54)	27,23 (4,67)	36,27 (5,44)	44,88 (7,31)	74,22 (12,05)	101,36 (13,68)	178,18 (13,31)
Número de veces que ha sido seleccionado cada SNP causal (%)	$V14$	30	51	53	59	80	86	90
	$V46$	68	100	100	100	100	100	100
	$V98$	32	85	100	100	100	100	100
	$V164$	100	100	100	100	100	100	100
	$V176$	71	97	99	100	100	100	100
Media moda y sd de SNPs causales identificados		3,01 3 (0,88)	4,33 4 (0,64)	4,52 5 (0,50)	4,59 5 (0,49)	4,80 5 (0,40)	4,86 5 (0,35)	4,90 5 (0,30)
Número de veces que han sido identificados todos los SNPs causales (%)		3	42	52	59	80	86	90
Promedio de variables no causales seleccionadas y (sd)		14,49 (4,65)	22,90 (4,70)	31,75 (5,40)	40,29 (7,32)	69,42 (12,11)	96,50 (13,72)	173,28 (13,33)

Tabla 3.5: Resultados tras 100 simulaciones

Frecuencia con la que se seleccionan los SNPs causales

Cuantas más variables considera el modelo, más SNPs causales es capaz de localizar. De esta manera, el método *lasso* tiende a identificar tres de los cinco SNPs causales. Si damos el mismo peso a la norma l_1 y a la norma l_2, estaremos incluyendo en el modelo todos los SNPs causales en algo más de la mitad de los casos, aunque tendremos algunos problemas para identificar a la variable $V14$ como causal. A partir de $\alpha = 0{,}3$ únicamente tendremos dificultades para identificar la variable $V14$, el resto estarán siempre consideradas en el modelo.

Capacidad predictiva

Se observa una tendencia a que el AUC disminuya en la medida en que el valor de α se reduce. Sin embargo, esta disminución es insignificante y no sería adecuado rechazar uno de los modelos por este motivo. El motivo de esta tendencia puede ser el sobreajuste.

Conclusiones del análisis

Aunque el método *lasso* es el que menos variables no causales incluye en el modelo, es prácticamente incapaz de identificar los cinco SNPs causales si-

multáneamente. Esto se debe a que cuando un grupo de variables están muy relacionadas entre sí (en alto desequilibrio de ligamento), el método *lasso* escoge solamente una de ellas. Por lo tanto, si el método ha seleccionado una variable no causal pero que pareciera más influyente en la respuesta por casualidad, será muy complicado que la verdadera variable causal relacionada con ella pase a formar parte del modelo. Esta conclusión es coherente con la obtenida por Ayers y Cordell en [8].

En caso de estar interesados en identificar todos los SNPs causales, será más conveniente utilizar el modelo elastic net. Al pasar, por ejemplo, de $\alpha = 1$ a $\alpha = 0{,}75$, la capacidad de identificar todos los SNPs causales simultáneamente crece considerablemente. Es difícil escoger un valor de α concreto, ya que éste dependerá del número de SNPs no causales que aceptemos incluir en el modelo.

Conclusiones del trabajo

La realización de este trabajo me ha permitido introducirme en el mundo de la bioestadística y trabajar con problemas estadísticos que surgen en el campo de la genética. De esta manera, he podido observar algunas de las peculiaridades que presentan este tipo de problemas.
He entendido las limitaciones que tienen los métodos estadísticos estudiados durante el grado a la hora de abordar problemas de dimensiones altas como los estudios de asociación del genoma completo y he analizado herramientas capaces de solventar dichas limitaciones, centrándome en las técnicas de regularización. He trabajado también con paquetes y funciones de R nuevos para mí, desarrollando una mayor destreza para programar e interpretar resultados y gráficos.

He estudiado un tema aún abierto y en constante desarrollo, por lo que si bien no he podido demostrar cuál será la herramienta óptima para resolver el problema, he tenido acceso a una gran cantidad de información actual y a estudios cuya finalidad era comparar algunos de los diferentes métodos estadísticos utilizados en estudios de asociación genética.

Todo esto me ha ayudado a tener una visión más amplia de la estadística y a percatarme de la necesidad de desarrollar nuevas técnicas para abordar los distintos tipos de problemas que precisan de la estadística para ser resueltos.

Apéndice A

Principales comandos de R utilizados

A continuación mostramos una parte del código utilizado en R para obtener los resultados del capítulo 3.

El paquete de R más utilizado a lo largo del trabajo ha sido *glmnet*. En [12] se explican todas las características del mismo. Para utilizar las funciones del paquete, ha sido suficiente con cagar el paquete y los datos.

```
library(glmnet)

geno <- read.table("C:/glmnet/Genotypes.txt")
pheno <- read.table("C:/glmnet/Phenotypes.txt")
geno1 <- as.matrix(geno) #forma matricial
pheno1 <- pheno[,1] #forma vectorial
```

Estos son los comandos utilizados para ajustar los modelos *lasso*, *ridge* y *elastic net* con todos los datos y para 100 valores diferentes de λ:

```
fit_lasso <- glmnet(geno1,pheno1,family="binomial",alpha
   =1,nlambda=100)
fit_ridge <- glmnet(geno1,pheno1,family="binomial",alpha
   =0,nlambda=100)
fit_enet <- glmnet(geno1,pheno1,family="binomial",alpha
   =0.4,nlambda=100)
```

La función *cv.glmnet*, que escoge el mejor λ por validación cruzada se aplica de la siguiente manera:

```
lasso_cv <- cv.glmnet(geno1,pheno1,family="binomial",
    alpha=1)
ridge_cv <- cv.glmnet(geno1,pheno1,family="binomial",
    alpha=0)
enet_cv <- cv.glmnet(geno1,pheno1,family="binomial",alpha
    =0.4)
```

Para generar las gráficas que muestran la evolución de los coeficientes en función de λ hemos utilizado el siguiente comando:

```
plot(modelo ajustado, xvar="lambda", label=TRUE)
```

En cuanto a los comandos utilizados para medir la capacidad predictiva de los modelos, hemos usado el *xtabs* para crear las tablas de clasificación y el paquete *pROC* para generar las curvas ROC. Toda la información acerca de este paquete está en [13].

```
contin <- xtabs(~y+yhat, data=datos) #'datos' contiene
    los fenotipos reales, las probabilidades estimadas y
    los fenotipos estimados

library(pROC)

ROC <- roc(datos$y, datos$phat)
plot(ROC,legacy.axes=TRUE,print.thres=TRUE, print.auc=
    TRUE)
```

Bibliografía

[1] J. A. Riancho, (2012). Enfermedades complejas y análisis genéticos por el método GWAS. Ventajas y limitaciones. *Reumatología Clínica*, 8(2):56-57.

[2] J. M. Soria, J. F. Dilmb, J. Martnez-Gonzlez, M. Camacho, C. Rodrguez, J. M. Romero, S. Bellmuntb, J. R. Escudero,L. Vilac, (2010). Métodos de estudio de las enfermedades complejas: aneurismas de la aorta abdominal. *Angiología*, 62(2):58-64.

[3] D. J. Balding, (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781-791.

[4] G. James, D. Witten, T. Hastie, R. Tibshirani,(2015). *An Introduction to Statistical Learning with Applications in R*, 6^{th} printing.

[5] A. E. Hoerl, R. W. Keenard, (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12: 55-67.

[6] R. Tibshirani, (1996). Regression shrinkage and selection via the LASSO. *J R Statist Soc*, 58:267-288.

[7] H. Zou, T. Hastie, (2005). Regularization and variable selection via the elastic net. *J R Statist Soc* B, 67(2):301-320.

[8] K. L. Ayers, H. J. Cordell, (2010). SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. *Genet. Epidemiol.* 34:879891.

[9] P. Waldmann, G. Mszros, B. Gredler, C. Frst, J. Slkner, (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics* Vol.4, Article 270.

[10] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobe, K. Lange, (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6): 714-721.

[11] J. H. Friedman, T. Hastie, R. Tibshirani, (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1-22. URL https://www.jstatsoft.org/article/view/v033i01 .

[12] J. Friedman, T. Hastie, N. Simon, R. Tibshirani, (2016). Lasso and Elastic-Net Regularized Generalized Linear Models. Version 2.0-5. URL http://www.jstatsoft.org/v33/i01/ .

[13] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, M. Mller. (2015). Display and Analyze ROC Curves. Version 1.8. URL http://expasy.org/tools/pROC/ .

[14] S. Castro, L. Mateus. (2012). Técnicas de regularización para regresión en grandes dimensiones. Aplicación a estudios de asociación genética. URL: http://conferencias.unc.edu.ar/index.php/xclatse/clatse2012/ paper/ view/645/93

[15] P. Cacheiro. (2011). Métodos de selección de variables en estudios de asociación genética. Aplicación a un estudio de genes candidatos en Enfermedad de Parkinson. Proyecto Fin de Máster. URL: http:// eio. usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_608.pdf