UNIVERSIDAD NACIONAL DE COLOMBIA

Modelos Lineales Mixtos Generalizados aplicados a estudios de datos con estructura de familia

Trabajo presentado como requisito para optar el título de:
Magíster en Matemática Aplicada

Presentado por:
Luis Miguel Orozco Restrepo

Directora:
PhD. Nubia Esteban Duarte
Profesora Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales, Departamento de Matemáticas
Manizales, Colombia
2020

Agradecimientos

Primero que todo agradezco a Dios quien me da fuerza y sabiduría para enfrentar cada paso en mi vida, para cumplir con cada meta que me he trazado, con su ayuda todo se ha visto reflejado en mis proyecciones de vida.

De manera especial agradezco a mi Mamá, quien ha sido mi guía, la persona que a pesar de las adversidades ha sido un ejemplo a seguir. Ella es emprendedora, lucha día a día junto a mí y ha estado a mi lado en los momentos que más he necesitado y celebrando mis triunfos hasta el día de hoy. También agradezco a mi familia, a mi pareja, a los profesores, a todos mis amigos y demás personas que me han acompañado durante este proceso, ya que cada uno de ellos me ha brindado su acompañamiento para luchar por mis ideales, por esto les digo gracias.

Quiero agradecer inmensamente a la directora de esta tesis, profesora Nubia Esteban Duarte, puesto que sin su disposición, conocimiento, paciencia y dedicación no hubiese sido posible realizar este trabajo. Gracias profesora por su tiempo, confianza y sobre todo por el acompañamiento que he tenido durante esta construcción de mi proyecto de vida profesional.

Resumen

Modelos Lineales Mixtos Generalizados aplicados a estudios de datos con estructura de familia

El campo de los Modelos Lineales Mixtos y Modelos Lineales Mixtos Generalizados ha tenido un gran desarrollo en los últimos años. No obstante, existen campos de aplicación en los cuales es necesaria una amplia fundamentación teórica y práctica como lo es en estudios de datos con estructura de familia dentro del área de Genética.
En este trabajo, dentro de los Modelos Lineales Mixtos se presenta la respectiva fundamentación teórica para datos de familia y se realiza una aplicación utilizando un conjunto de datos reales del Proyecto "Corazones de Baependi" (Processo Fapesp 2007/58150-7) del laboratorio de Genética y Cardiología Molecular (Incor/USP), cuyo objetivo es identificar genes asociados a factores de riesgo cardivascular. Oliveira et al. (2008).
La teoría de Modelos Lineales Mixtos es extendida a Modelos Lineales Mixtos Generalizados, en el sentido de utilizar una función de enlace que contenga la información del parentesco de los individuos dentro de cada familia. Se resalta el hecho de que ya existe un programa llamado *SOLAR* que presenta resultados sin ninguna base teórica explícita asociada. Un ejemplo ilustrativo comparando las salidas del programa *SOLAR* con el programa *R* es establecido a través de un conjunto de datos que pertenecen a una población de indios *Xavantes*, (Brasil) en la cual se resalta que el sobrepeso y la obesidad son determinantes de alto riesgo para la diabetes.

Palabras Clave: Modelo Lineal Mixto, Modelo Lineal Mixto Generalizado, Modelo Lineal Generalizado, Modelo Mixto Poligénico, Marcadores Moleculares, Datos de familia.

Abstract

Generalized Mixed Linear Models applied to data studies with family structure

The Mixed Linear Models and Generalized Mixed Linear Models fields have had a great development in recent years. However, there are fields of application in which a broader theoretical and practical foundation is necessary, such as in data studies with a family structure within the area of Genetics. In this work, within the Mixed Linear Models the respective theoretical foundation for family data is presented and application is done using a set of real data from the "Hearts of Baependi"(Processo Fapesp 2007/58150-7) from the Laboratory of Genetics and Molecular Cardiology (Incor/USP), which its objective is to identify genes associated with cardiovascular risk factors. Oliveira et al. (2008).
The theory of Mixed Linear Models is extended to Generalized Mixed Linear Models, in the sense of using a link that contains the information of the relationship of individuals within each family. The fact that there is already a program called *SOLAR* that presents results without any explicit associated theoretical base is highlighted. An illustrative example comparing the outputs of the *SOLAR* program with the *R* program is established through a data set belonging to a population of *Xavantes* indigenous, (Brazil) in which it is highlighted that overweight and obesity are high-risk determinants for diabetes.

Key words: Mixed Linear Model, Generalized Mixed Linear Model, Generalized Linear Model, Polygenic Mixed Model, Molecular Markers, Family data.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

El campo de los Modelos Lineales Mixtos y Modelos Lineales Mixtos Generalizados ha tenido un gran desarrollo en los últimos años, en especial desde la creación e introducción de rutinas computacionales intensivas. En el área de Estadística se ha implementado el uso de estos modelos no sólo en trabajos de investigación sino en el quehacer de empresas y centros de investigación, siendo así que este tipo de modelos se ha posicionado como una poderosa herramienta de análisis estadístico. Sin embargo, existen campos de aplicación en los cuales es necesaria una amplia fundamentación teórica y práctica principalmente de los Modelos Lineales Mixtos Generalizados específicamente en estudios de datos con estructura de familia cuando se analizan variables respuesta de tipo binario, dentro del área de Genética.

La historia de la Estadística con aplicaciones en Genética es bastante antigua. El artículo de Fisher R. (1918) fue de gran importancia histórica para la genética de poblaciones porque proporcionó la primera demostración de que genes mendelianos múltiples podrían ser responsables por los patrones observados de transmisión de caracteres cuantitativos. Cabe destacar que el artículo mencionado dio origen a muchos trabajos, tal como el libro publicado en 1925 que trajo muchas contribuciones al Análisis de Componentes de Varianza, o equivalentemente al análisis de Modelos Lineales Mixtos, que sirvió de base para muchos estudios en Genética Cuantitativa.

Posteriormente, considerando las correlaciones generadas por datos de individuos de la misma familia, varios autores utilizaron el método de Máxima Verosimilitud para modelar datos agrupados en familias, en que los núcleos familiares son considerados independientes (Elston R. & Stewart J. 1971;

Lange K. et al, 1976). Cabe resaltar que inicialmente los estudios se restringían a estudios de pares de hermanos de diferentes familias (Haseman J. & Elston R. 1972). Posteriormente, otros autores como Amos C. & Elston R. (1989); Schork (1993); Blangero & Almasy (1997) extendieron el modelo de Haseman - Elston para realizar estudios de familias completas, lo cual da la posibilidad de utilizar estruturas de familias más complejas permitiendo la utilización de **coeficientes de parentesco** o en términos generales la **matriz de parentesco** (ver Tabla 2.2) para tener covarianzas más informativas lo cual genera estimativas más precisas en el ajuste de modelos. Amos C. (1994) propuso un modelo de Componentes de Varianza con efectos mixtos, dando posibilidad a análisis de los efectos de covariables y aumentando el poder de las pruebas resultantes. Otros métodos fueron desarrollados para estudiar familias extendidas de tamaño y complejidad arbitrarios, incluyendo análisis de interacciones gen por ambiente bien como análisis multivariadas. (Blangero J. 1993; Kraft P. & De Andrade M. 2003; Schork N. 1993; Blangero, J. & Almasy, L. (1997); De Andrade et al, 1997).

Cabe anotar que las investigaciones genéticas de los anteriores autores, en su mayoría se han publicado utilizando el modelo de Componentes de Varianza o Modelos Mixots cuando la variable respuesta es continua. Una parte del presente trabajo se fundamenta en esta teoría (considerando los análisis cuando la variable respuesta es continua). Cuando la variable respuesta es binaria, todavía se necesitan estudios para formalizar la teoría y también realizar aplicaciones con paquetes bien fundamentados en el programa *R*.

Diferentes fases experimentales caracterizan el análisis de datos de familia cuando la variable respuesta es binaria, dando lugar a un Modelo Lineal Mixto Generalizado. Modelo mixto en el sentido de acoplar la matriz de correlación existente en datos de familia por cuenta del parentesco de individuos. Modelo Generalizado en el sentido de utilizar una función de enlace para llevar en consideración la variable respuesta binaria.

Como ya fue anotado, los modelos mencionados cuando la variable respuesta es cuantitativa continua, se han planteado varios estudios por varios autores y se tiene una teoría muy bien fundamentada para los diferentes análisis incluyendo cálculos del **Coeficiente de Correlación Intraclase**, que en términos genéticos este coeficiente es ampliamente conocido como **Coeficiente de Herdabilidad**. En el caso de estudios con datos de familia en Moldelos Lineales Mixtos Generalizados no hay estudios específicos, fundamentados teoricamente, para el cálculo de dicho coeficiente que es relevan-

te en estudios genéticos con datos de familia ya que nos indica la proporción de variabilidad que es debida a componentes genéticos. En estudios del área de genética, el cálculo de este coeficiente necesario e indispensable, ya que indica si es factible incluir covariables genéticas en un modelo o no.

El estudio de este tipo de modelos de respuesta binaria para datos de familia es relevante ya que cuando se realizan análisis de individuos sin estructura familiar se desprecian componentes informativas de bases genéticas que pasan de generación en generación. Por otro lado, existen muchas variables que necesitan ser analizadas bajo esta estrutura, por ejemplo, si se tiene o no una determinada enfermedad. Para este caso, hay un programa que es ampliamente conocido en el área de genética, llamado SOLAR[1]. Este programa realiza los cálculos del coeficiente de correlación intraclase o herdabilidad para variables respuesta binaria, pero no se tiene la fundamentación teórica para la obtención de estos resultados. Por otro lado, dentro del análisis estadístico es indispensable tener estimativas de parámetros, mismo que no sean significativos, lo cual el programa **SOLAR** no los libera. Es también en ese sentido que se enfoca este trabajo estableciendo la fundamentación teórica, investigando lo que está atrás de los resultados arrojados por este programa. Por otro lado, se utilizarán los paquetes que existen en el programa **R** sobre Modelos Lineales Mixtos y Modelos Lineales Mixtos Generalizados. Además estableceremos la relacion entre los paquetes del programa R y el programa SOLAR.

1.1. Justificación

El presente trabajo surge de la necesidad de aplicar los Modelos Lineales Mixtos o (Modelo de Componentes de Varianza), así como Modelos Lineales Mixtos Generalizados a datos con estructura de familia y covariables genéticas representadas en plataformas de marcadores moleculares SNP (ver sección 1.2), que son datos de alta dimensión y así contribuir con investigaciones en enfermedades cardiovasculares. También, existe la necesidad de formalizar la teoría asociada a Modelos Lineales Mixtos Generalizados cuando el objeto de estudio son datos con estructura de familia para así analizar fenótipos de características binarias (si/no; por ejemplo tiene o no tiene una determinada enfermedad, caracterizando posibles riesgos). Destacamos que los análisis para individuos no relacionados (sin estructura familiar)

[1] http://www.solar-eclipse-genetics.org/

está muy bien fundamentada, siendo frecuente su utilización en diferentes campos del saber siendo posible que se pierda información que sí se puede obtener cuando los datos tienen estructura de familia.

Para las aplicaciones del presente trabajo, existen datos de investigaciones que dan la oportunidad de fundamentar la teoría y realizar las respectivas aplicaciones, específicamente, se utilizarán datos del Proyecto "Corazones de Baependi" (Processo Fapesp 2007/58150-7) del laboratorio de Genética y Cardiología Molecular (Incor/USP) Oliveira et al. (2008), Giolo et al. (2009). También, simplemente para ilustración de la teoría de Modelos Generalizados, se utilizará un conjunto de datos que pertenecen a una población de indios *Xavantes*, (Brasil) en la cual se resalta que el sobrepeso y la obesidad son determinantes de alto riesgo para la diabetes.

Dado lo anterior, el presente trabajo tiene como objetivo general:

Estudiar la fundamentación teórica y práctica de los Modelos Lineales Mixtos con aplicaciones específicas en datos de familia y marcadores moleculares SNPs, variable respuesta continua. Por otro lado, formalizar la teoría asociada con Modelos Lineales Mixtos Generalizados para estudios con datos que tienen estructura familiar, principalmente para variables respuesta binarias. Aplicar la teoría a conjuntos de datos reales.

Los objetivos específicos son:

1. Fundamentar la teorías asociada con Modelos Lineales Mixtos. Aplicar a conjuntos de datos reales.

2. Investigar la teoría que es utilizada en el Programa *SOLAR* para realizar los análisis en estudios principalmente con respuesta binaria. Relacionar la teoría con los Modelos Lineales Mixtos Generalizados.

3. Investigar sobre los paquetes existentes en el programa *R* que han sido establecidos para modelos Lineales Generalizados y adaptarlos para incorporar la información pertinente a datos de familia.

1.2. Conceptos fundamentales en Genética

1.2.1. Conceptos básicos

Tomando algunos conceptos descritos en Griffiths A. et al. (2000):

- **Genética** La genética es la ciencia que estudia la variación y la transmisión de rasgos (caracteres o características) de una generación a la otra. En esta definición, variación se refiere a variación genética; es decir, el rango de posibles valores para un carácter cuando es influenciado por la herencia. La herencia es la transmisión de caracteres o rasgos de los padres a su descendencia vía el material genético (localizado en el núcleo de cada célula del cuerpo a excepción de las células reproductoras, entre otras). Esta transmisión toma lugar en el momento de la fertilización en la reproducción, cuando un espermatozoide se une con un óvulo para producir un nuevo individuo con una composición genética única.

- **Medio ambiente** El medio ambiente es generalmente entendido como los alrededores físicos del individuo, luz, temperatura, ventilación y otros parámetros que pueden contribuir al desarrollo físico; es decir, es la combinación de todos los factores, con excepción de los genéticos, que pueden afectar la expresión de los genes (proceso mediante el cual la información almacenada en el ADN es usada para dirigir la síntesis de un producto génico específico como proteínas, RNA, etc).

- **Gen** Es la unidad física básica de herencia que consiste en una secuencia de ADN en una locación específica en un cromosoma.

- **ADN** Acido desoxirribonucleico, molécula que conforma el código genético.

- **Cromosoma** Uno de muchos hilos de ADN y proteínas asociadas presentes en el núcleo de cada célula.

- **Locus** La localización específica de un gen en un cromosoma. Loci es el plural de locus.

- **Alelo** Forma alternativa de un gen. Alelos múltiples cuando hay más de dos alelos posibles en un locus.

- **Genótipo y Fenótipo** El genótipo de un individuo representa el gen (o grupo de genes) responsable por un rasgo o carácter en particular. En un sentido más general, el genótipo describe todo el grupo de genes que un individuo ha heredado. El fentipo es el valor que toma un rasgo; es decir, es lo que puede ser observado o medido. Por ejemplo, el fenótipo puede ser el peso y la estatura de los individuos, el porcentaje de grasa.

Existe una diferencia importante entre genótipo y fenótipo. El genótipo es esencialmente una característica fija del organismo; permanece constante a lo largo de la vida del individuo y no es modificado por el medio ambiente. Cuando solamente uno o un par de genes son responsables por un rasgo, el genótipo permanece generalmente sin cambios a lo largo de la vida del individuo (ejemplo color de pelo). En este caso, el fenotipo otorga una buena indicación de la composición genética del individuo. Sin embargo, para algunos rasgos, el fenotipo cambia constantemente a lo largo de la vida del individuo como respuesta a factores ambientales. En este caso, el fenótipo no es un indicador directo confiable del genótipo. Esto generalmente se presenta cuando muchos genes se encuentran involucrados en la expresión de un rasgo.

Como descrito anteriormente, el fenótipo es el conjunto de características que un individuo posee, que en general es el resultado entre las interacciones entre el genótipo (constitución genética del individuo) y el ambiente (Falconer & Mackay, 1996). Esta relación se puede expresar mediante la ecuación:

$$Y = \mu + G + E, \tag{1.1}$$

donde

- Y representa los valores fenotípicos.
- μ es la media poblacional de la respuesta
- G es el valor o efecto genotípico
- E es el componente residual.

Esto muestra que el genótipo no se expresa en su totalidad en el fenótipo, sino que se ve modificado por el ambiente. En algunos casos es posible encontrar interacción entre genótipo y ambiente.

Los valores Y , G y E se pueden expresar en cualquier unidad que represente una propiedad Biológica que pueda ser medida de forma discreta o continua, tal como peso, tenor de grasa, presión arterial, etc. Cuando no existe efecto del ambiente (modelado en el factor E) sobre determinada característica el fenótipo presentado es idéntico para todos los individuos que poseen el mismo genótipo, diferenciándose por los alelos en G.

- **Herdabilidad** El conocimiento de la proporción de la variabilidad que es de origen genético es un parámetro de mucho interés, es conocido como **herdabilidad**, ampliamente conocido en Estadística como el **Coeficiente de Correlación Intraclase**(Falconer & Mackay, 1996; Gutierrez J. 2010).

  Teniendo en cuenta la herdabilidad en el sentido estricto, es definida como la proporción de la variabilidad fenotípica que es de origen genético aditivo:

  $$h_g^2 = \frac{\sigma_G^2}{\sigma_Y^2}. \tag{1.2}$$

  La ecuación anterior es un importante resultado porque expresa que la herdabilidad depende solamente de la varianza genética aditiva. Significa que la herdabilidad de un carácter es la proporción de la varianza genética aditiva en relación a la varianza fenotípica total. Esto permite medir cuánto de la variabilidad fenotípica de un carácter en una población dada, es probable que se transmita a sus descendientes, puesto que se considera que la parte aditiva es la que se hereda de forma directa, ya que los otros componentes (dominancia, epistasía) son interacciones entre genes que se tendrían en cuenta en un sentido más amplio de herdabilidad. Un punto importante sobre el componente de varianza genotípica es que estos representan los efectos estadísticos acumulados de todos los genes que afectan el carácter. Pocas inferencias sobre el modo de herencia real del carácter son posibles a partir de los componentes de varianza, principalmente en relación al número de genes envueltos en sus efectos individuales (Harlt D. & Clark A. 2010).

  - **Dominancia**: Es importante destacar que el efecto genético de un gen puede ser descompuesto en dos componentes: Efecto genético aditivo y efecto genético de Dominancia. El primero puede ser predicho linealmente por medio del número de alelos de un cierto tipo que definen el genótipo, mientras que el efecto de Dominancia es el valor fenotípico que no puede ser explicado linelamente (residuo genético debido al efecto de interacción entre los alelos A1 y A2 en un mismo punto o loco).
  - **Epistasía**: Es la interacción entre diferentes genes al expresar un determinado carácter fenotípico, es decir, cuando la expresión de uno o más genes dependen de la expresión de otro gen. El

término Epistasía fue utilizado originalmente por William Bateson para describir situaciones en que el efecto fenotípico de un gen interfería en la expresión de otro gen (Bateson, 1909). Un ejemplo clásico es el de una especie de gallinas blancas, en que los efectos de un gen para el color de las plumas son ocultos por un alelo dominante de otro locus, que impide el depósito de cualqueir pigmento. A lo largo de los años, los geneticistas moleculares ampliaran la definición de epistasía, para incluir todo tipo de interacción entre los alelos de diferentes genes.

- **Marcadores Moleculares** Los marcadores moleculares son puntos de referencia del genoma cuya posición es conocida y la constitución genotípica del individuo puede ser identificada. Son estratégicamente dispuestos a lo largo del genoma y debidamente codificados como variables predictoras o covariables. Estadísticamente se consideran como el muestreo del genoma. Corresponden a variaciones en la secuencia de ADN obtenidos por medio de técnicas moleculares, siendo posible conocer si en ese determinado punto un individuo es homozigoto o heterozigoto.

  En la Figura (1.1) se ilustra un cromosoma, como un arreglo lineal de marcadores cuya distancia entre ellos es dada en centimorgans (ilustrado en la columna izquierda). Cada marcador ocupa una posición fija en el cromosoma y tiene un nombre específico (columna de la derecha).

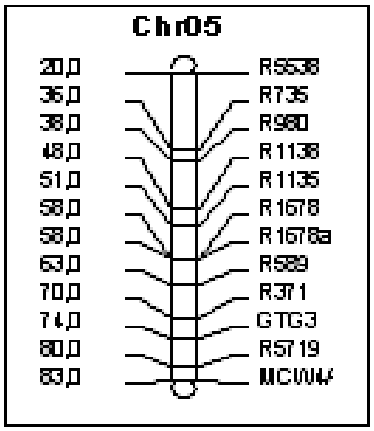


Figura 1.1: Ilustración de marcadores moleculares

Existen varios tipos de marcadores moleculares, como son los microsatélites que son compuestos por varios nucleotídeos, multialélicos con varias formas de respuesta genotípica.

Se centrará la atención en los marcadores tipo SNP (Single Nucleotide Polymorphism) o polimorfismos de un único nucleotídeo que representan áreas donde largas secuencias de ADN difieren entre los individuos en apenas un nuclotídeo, como es ilustrado en la Figura 2. Por exemplo, Maria tiene en su genoma la secuencia AAATTTCGCCG**G**TA, esa misma secuencia en Juan puede ser AAATTTCGCCG**T**TA. Observamos que hubo una alteración en un único nucleotídeo lo que puede ser la diferencia entre Maria y Juan tener un gen que produce una proteina defectuosa o no.

La Figura (1.2) está disponible en `httt://en.wikipedia.org/wiki/Single_nucleotide_polymorphism.`

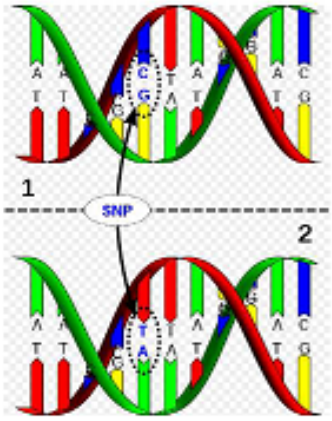


Figura 1.2: Ilustración de un marcador molecular conocido como SNP

Cada SNP se representa mediante la siguiente codificación:

$$SNP = \begin{cases} 2, & \text{si el individuo es homozigoto } AA, \\ 1, & \text{si el indivíduo es heterozigoto } Aa, \\ 0, & \text{si el indivíduo es homozigoto } aa\ , \end{cases} \tag{1.3}$$

Como se puede observar, el marcador SNP es dialélico (ocurriendo los alelos A o a, como ilustrado). En general en cada posición o *locus* el alelo A es el que ocurre con menor frecuencia.

Cabe resaltar que para genotipar marcadores microsatélites los costos son demasiado altos, puesto que el proceso se realiza marcador por marcador, mientras que para los marcadores tipo SNP se han desarrollado tecnolgías para genotipar en masa, reduciendo los costos.

En general, actualmente existe la capacidad de identificar un grande número de marcadores moleculares y también de realizar la genotipagen de los individuos en grande escala, lo cual posibilita la identificación de polimorfismos que son genéticamente ligados a los genes que afectan algún carácter cuantitativo. En muchos organismos ya es conocida la secuencia del ADN genómico y pueden ser reconocidos genes candidatos a través de modelos estadísticos que, tal vez, puedan estar afectando un carácter cuantitativo.

Capítulo 2

Modelos Lineales Mixtos

En esta sección será presentada, de forma resumida, la parte teórica de las metodologías estadísticas que serán estudiadas, analizadas y fundamentadas matemáticamente a través del desarrollo del presente trabajo.

Un modelo mixto permite analizar una variable aleatoria Y modelando simultáneamente el valor esperado del fenómeno estudiado y su variabilidad, es decir se modela un efecto fijo a todos los sujetos de estudio y otro efecto aleatorio asociado a cada uno de los sujetos.

Es posible obtener un modelo mixto de la agrupación de un modelo en dos etapas, que no es más que la adaptación al modelado lineal simple que permite aproximar los perfiles asociados a cada sujeto así: La primera parte es ajustar una regresión asociada a cada sujeto por separado y la segunda es ajustar una regresión a los coeficientes asociados a cada sujeto en función de variables conocidas (efectos fijos) (Correa J. & Salazar J. 2016; Penrose L. 1938):

1. **Primera Etapa:** Asuma que Y_{ij} es la respuesta para el i-ésimo sujero $X_{ij}, i = 1, \ldots, m$ y $j = 1, \ldots, n$. Así, se tiene que $\mathbf{Y_i} = (\mathbf{Y_{i1}}, \ldots, \mathbf{Y_{in_i}})^\top$ es el valor de respuestas para el sujeto i. De esta manera, el modelo $\mathbf{Y_i} = \mathbf{Z_i}\beta_\mathbf{i} + \epsilon_\mathbf{i}$, $i = 1, \ldots, m$, describe la variabilidad dentro de los sujetos, donde $\mathbf{Z_i}$ es una matriz $n_i \times q$ de covariables conocidas, β_i es un vector q-dimensional con los coeficientes de regresión sujeto-especificos y $\epsilon_\mathbf{i}$ se asume $N(0, \Sigma_i)$ donde Σ_i es una matriz de varianzas y covarianzas.

2. **Segunda Etapa:** En esta estapa se modela la variabilidad entre los sujeros. Esta variabilidad puede modelarse, si los $\beta' s$ se relacionan con

variables conocidas, obteniendo; $\beta_i = \mathbf{K_i}\beta + \gamma_\mathbf{i}$ para $i = 1, 2, \ldots, m$, siendo K_i una matriz de orden $q \times p$ de covariables conocidas, β un vector con dimensión p de parámetros de regresión desconocidos y $\gamma_i \sim N(0, D)$, con D, la matriz de varianzas y covarianzas $q \times q$.

De esta manera un modelo en dos etapas dado por:

$$\begin{aligned} Y_i &= Z_i\beta_i + \epsilon_i \\ \beta_i &= K_i\beta + \gamma_i \quad \text{Para } i = 1, 2, \ldots, n \end{aligned}$$

Se puede resumir en un sólo modelo mixto (MLM) dado por:

$$\begin{aligned} Y_i &= \underbrace{Z_i K_i}_{X_i}\beta + Z_i\gamma_i + \epsilon_i \\ Y_i &= X_i\beta + Z_i\gamma_i + \epsilon_i \end{aligned}$$

Para este modelo se tiene $Z_iK_i = X_i$, para $\gamma_i, \gamma_2 \ldots, \gamma_n$ y $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son mutuamente independientes, en donde:

- β: Hace referencia a los efectos fijos.
- γ_i y ϵ_i: Hace referencia a los efectos aleatorios.
- D y Σ: Contienen las componentes de varianza.

2.1. Forma Matricial

El modelo lineal mixto puede ser escrito en la siguiente forma:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\gamma_i + \epsilon_i \; i = 1 \ldots, c \tag{2.1}$$

En que

- $\mathbf{Y}_i$ representa un vector $(n_i \times 1)$ de respuestas de la i-ésima unidad.
- β es un vector $(p \times 1)$de parámetros de los efectos fijos.
- $\mathbf{X}_i$ es una matriz $(n_i \times p)$ de especificación de los efectos fijos, esta matriz es conocida.
- γ_i es un vector $(q \times 1)$ de variables latentes, comunmente denominadas efectos aleatórios, los cuales reflejan el efecto individual de la i-ésima unidad.

- $\mathbf{Z}_i$ es una matriz $(n_i \times q)$ de especificación de los efectos aleatorios, tambiém conocida y de rango completo.
- ϵ_i es un vector $(n_i \times 1)$ de errores aleatorios.

Escribiendo, $\mathbf{Y} = (\mathbf{Y}_1^\top, ..., \mathbf{Y}_c^\top)^\top$; $\mathbf{X} = (\mathbf{X}_1^\top, ..., \mathbf{X}_c^\top)^\top$; $\mathbf{Z}$=diag$(\mathbf{Z}_1, ..., \mathbf{Z}_c)$; $\gamma = (\gamma_1^\top, ..., \gamma_c^\top)^\top$ y $\epsilon = (\epsilon_1^\top, ..., \epsilon_c^\top)^\top$, el modelo (2.1) puede ser reescrito matricialmente como:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \tag{2.2}$$

En general, se asume que $E[\gamma] = \mathbf{0}$, $E[\epsilon] = \mathbf{0}$, con matriz de covarianzas:

$$Cov \quad \begin{pmatrix} \gamma \\ \epsilon \end{pmatrix} = \begin{pmatrix} \Delta & \mathbf{0}_{cq \times n} \\ \mathbf{0}_{n \times cq} & \Sigma \end{pmatrix}, \tag{2.3}$$

Donde $\mathbf{0}_{k_1 \times k_2}$ representa una matriz nula de orden $k_1 \times k_2$, $\mathbf{\Delta}$ y $\mathbf{\Sigma}$ son matrices positivas definidas, cuadradas cuyas dimensiones son cq y $n = \sum_{i=1}^{c} n_i$ y corresponden a las matrices de covarianzas de los vectores aleatorios γ e ϵ, respectivamente.

En el modelo (2.2), los efectos fijos son utilizados para modelar el valor esperado de la variable respuesta $\mathbf{Y}$, esto es $E[Y] = \mathbf{X}\beta$ mientras que los efectos aleatorios son utilizados para modelar la estructura de covarianza $cov[Y] = \mathbf{V} = \mathbf{Z\Delta Z} + \mathbf{\Sigma}$. Usualmente, se asume que γ y ϵ siguen una distribución normal cq y n-variada, respectivamente.

En el modelo (2.2), escribiendo $\xi = \mathbf{Z}\gamma + \epsilon$, se obtiene el modelo marginal.

$$\mathbf{Y} = \mathbf{X}\beta + \xi, \tag{2.4}$$

con ξ teniendo una distribución normal n-variada con vector de medias $\mathbf{0}_n$ y matriz de covarianzas $\mathbf{V} = \mathbf{Z\Delta Z}^\top + \mathbf{\Sigma}$.

Colocando un parámetro de dispersión común, se tiene que $\mathbf{\Delta}$=$\sigma^2\mathbf{D}$ e $\mathbf{\Sigma}$=$\sigma^2\mathbf{R}$, con $\mathbf{D}$ y $\mathbf{R}$ denotando matrices positivas definidas. Así,

$$\mathbf{V} = (\mathbf{ZDZ}^\top + \mathbf{R})\sigma^2. \tag{2.5}$$

Diferentes estructuras para $\mathbf{D}$ y $\mathbf{R}$ pueden ser encontradas en la literatura (Searle et al., 1992; Verbeke G. & Molenberghs 1997).

Cuando $\mathbf{R}$ es una matriz diagonal, el modelo (2.2) es denominado **modelo de independencia condicional**; si $\mathbf{R} = \mathbf{I}_n$ y $\mathbf{\Delta}$=$\mathbf{0}_{cq \times cq}$ el modelo

(2.2) corresponde al modelo homoscedástico usual.

Para el modelo lineal mixto el **mejor estimador lineal insegado** "**BLUE**-best linear unbiased estimator"para β y el **mejor predictor lineal insegado** "**BLUP**-best linear unbiased preditor"para el vector de efectos aleatorios γ puede ser obtenido por mínimos cuadrados por medio de la solución del siguiente sistema de ecuaciones, ampliamente conocidas en la literatura como **Ecuaciones de Henderson**:

$$\begin{pmatrix} \mathbf{X^\top R^{-1} Y} \\ \mathbf{Z^\top R^{-1} Y} \end{pmatrix} = \begin{pmatrix} \mathbf{X^\top R^{-1} X} & \mathbf{X^\top R^{-1} Z} \\ \mathbf{Z^\top R^{-1} X} & \mathbf{Z^\top R^{-1} Z + D^{-1}} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}. \tag{2.6}$$

El BLUE de β es dado por

$$\hat{\beta} = (\mathbf{X^\top V^{-1} X})^{-1} \mathbf{X^\top V^{-1} Y} \tag{2.7}$$

con varianza

$$V(\hat{\beta}) = (\mathbf{X^\top V^{-1} X})^{-1} \tag{2.8}$$

Se define la matriz $\mathbf{M}$ como

$$\mathbf{M} = \sigma^2 \mathbf{V}^{-1}, \tag{2.9}$$

en que $\mathbf{V}$ fue descrita en (2.5). Así se tiene que la matriz $\mathbf{M}$ puede ser escrita como $\mathbf{M} = (\mathbf{ZDZ^T} + \mathbf{R})^{-1}$ y por tanto la matriz $\mathbf{M^{-1}}$ es dada por:

$$\mathbf{M}^{-1} = \sigma^{-2} \mathbf{V} = (\mathbf{ZDZ}^\top + \mathbf{R}). \tag{2.10}$$

Esta matriz $\mathbf{M}$ hace parte de un predictor lineal y es útil en el desarrollo de diferentes teorías mediante particiones apropiadas (Duarte et al., 2014).

El BLUE de β en función de la matriz $\mathbf{M}$ es dado por:

$$\hat{\beta} = (\mathbf{X^\top M X})^{-1} \mathbf{X^\top M Y}. \tag{2.11}$$

El BLUE y el BLUP satisfacen la siguiente ecuación

$$\mathbf{X^\top R^{-1} Y} = \mathbf{X^\top R^{-1} X} \hat{\beta} + \mathbf{X^\top R^{-1} Z} \hat{\gamma} \tag{2.12}$$

$$\mathbf{Z^\top R^{-1} Y} = \mathbf{Z^\top R^{-1} X} \hat{\beta} + (\mathbf{Z^\top R^{-1} Z + D^{-1}}) \hat{\gamma} \tag{2.13}$$

El BLUP del vector de efectos aleatorios γ es dado por

$$\begin{aligned} \hat{\gamma} &= (\mathbf{Z^\top R^{-1} Z + D^{-1}})^{-1} \mathbf{Z^\top R^{-1}} (\mathbf{Y} - \mathbf{X} \hat{\beta}) \\ &= \mathbf{C^{-1} Z^\top R^{-1}} (\mathbf{Y} - \mathbf{X} \hat{\beta}), \end{aligned} \tag{2.14}$$

con $\mathbf{C} = \mathbf{D^{-1}} + \mathbf{Z}^\top \mathbf{R^{-1}} \mathbf{Z}$. O alternativamente, $\hat{\gamma}$ puede ser escrito como:

$$\hat{\gamma} = \mathbf{D}\mathbf{Z}^\top \mathbf{M} \left(\mathbf{Y} - \mathbf{X}\hat{\beta} \right) = \mathbf{D}\mathbf{Z}^\top \mathbf{V^{-1}} \left(\mathbf{Y} - \mathbf{X}\hat{\beta} \right). \tag{2.15}$$

En la siguiente sección, el Modelo Lineal Mixto será descrito específicamente para análisis de datos con estructura familiar, es decir, donde se pueda incluir la entre individuos (padre, madre e hijos), como se explica en la Tabla 2.1.

2.2. Modelo Mixto Poligénico

En análisis de asociación basados en familias, el Modelo Lineal Mixto (Modelo de Componentes de Varianza) que en el área de genética es ampliamente conocido como **Modelo Mixto Poligénico**. Este modelo es muy útil para analizar datos basados en familia llevando en consideración la relación que hay entre individuos modelando efectos fijos y efectos aleatorios.

Para este modelo en la ecuación (2.2), el término $\mathbf{Z}\gamma$ es usualmente denotado por g, un vector de efectos genéticos aleatorios de los individuos siendo utilizada la estrutura familiar para modelar el efecto genético.

$$\mathbf{Y} = \mathbf{X}\beta + \underbrace{\mathbf{Z}\gamma}_{g} + \epsilon. \tag{2.16}$$

Bajo los mismos supuestos de un Modelo Linear Mixto (Verbeke G. & Molenberghs G. 2000), para finalidad de mapeamento de genes podemos describir el modelo poligénico (Amos C. 1994, Blangero, J. & Almasy, L. 1997):

$$\mathbf{Y}_f = \mathbf{X}_f \beta + \mathbf{g}_f + \mathbf{e}_f \;\; f = 1, \ldots, F. \tag{2.17}$$

- $\mathbf{Y}_f$ representa un vector $(n_f \times 1)$ de respuestas de la $f-$ ésima familia.
- β es un vector $(p \times 1)$ de parámetros que representan los efectos fijos.
- $\mathbf{X}_f$ representa una matriz $(n_f \times p)$ la cual es conocida y de rango completo.
- $\mathbf{g}_f$ indica el vector de efectos aleatorios poligénicos los cuales pueden surgir de los efectos que no se pueden medir, más específicamente, efectos de muchos genes que son los responsables por la correlación o el parentesco entre los individuos de la misma familia, en otras palabras, es el resumen de la información de muchos genes.

- $\mathbf{e}_f$ representa el vector de errores aleatorios. Note que en (2.17) la matriz de efectos aleatorios poligénicos es la Identidad.
  Los efectos aleatorios, $\mathbf{g}_f$ y $\mathbf{e}_f$, se asume que son no correlacionados, con media cero y varianza σ_g^2 y σ_e^2, respectivamente.

Una vez que el modelo es escrito de esa forma, las covarianzas entre las variables respuesta para los individuos i e $i^{'}$ de las familias f y $f^{'}$ es dada por:

$$cov(y_{if}, y_{i'f'}) = \begin{cases} \sigma_g^2 + \sigma_e^2 & \text{para } i = i' \\ 2\phi_{ii'}\sigma_g^2 & \text{para } i \neq i' \text{ y } f = f' \text{ relacionado} \\ 0 & \text{para } i \neq i' \text{ y } f \neq f' \text{ no relacionados} \end{cases} \tag{2.18}$$

El parámetro $2\phi_{ii'}$ es definido como el coeficiente de relacionamiento entre los indivíduos i e $i^{'}$ siendo dado por $(\frac{1}{2})^r$, en que r representa el grado de relacionamento.

De lo definido anteriormente, la varianza de Y denotada por σ_y^2 es la suma $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$.

Considerando todas las familias, la matriz de covarianza para el Modelo Poligénico dado en (2.17) puede ser definida como:

$$\mathbf{V} = \mathbf{2\Phi}\sigma_{\mathbf{g}}^{\mathbf{2}} + \mathbf{I}\sigma_{\mathbf{e}}^{\mathbf{2}} \tag{2.19}$$

con $2\mathbf{\Phi}$ una matriz bloco diagonal (bajo independencia entre familias) con elementos $2\phi_{ii'}$ correspondiendo a los datos o informaciones de cada familia; $\mathbf{I}$ la matriz identidad de orden $n \times n$.

La **herdabilidad**, que es el mismo **coeficiente de correlación intraclase** es un concepto importante en Genética siendo definida como la proporción de la varianza total que es debida a componentes genéticos. Llevando en consideración el modelo (2.17), la herdabilidad es dada por:

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}. \tag{2.20}$$

La interpretación de esta proporción es de gran utilidad, por ejemplo, si h_g^2 es pequeña se puede inferir que la variable no es esencialmente regulada por

factores genéticos.
Alternativamente, es posible definir la herdabilidad residual como:

$$h_e^2 = \frac{\sigma_e^2}{\sigma_g^2 + \sigma_e^2}. \tag{2.21}$$

Los estimadores de los componentes de varianza σ_g^2 y σ_e^2 son obtenidos por los métodos de Máxima Verosimilitud Restricta (Searle S. et al., 1992; Demidenko, 2004; Blangero J. & Almasy L. 1997; De Andrade et al., 1999).

En el modelo poligénico (2.17) la hipótesis de interés es:

$$H_0 : \sigma_g^2 = 0 \quad \text{vs} \quad H_1 : \sigma_g^2 > 0$$

Lo que corresponde a probar si el efecto poligénico es significante. Bajo H_0 se tiene el estadístico de razón de verosimilitud generalizada que es asintóticamente distribuída como una mezcla 1/2:1/2 de χ_1^2 y χ_0^2, respectivamente (Self G. & Liang K., 1987).

A seguir se presenta un ejemplo para ilustrar la matriz de parentesco entre individuos de dos familias hipotéticas a través de un **heredograma**, herramienta altamente usada en el estudio de familias, siendo un diagrama simplificado de la genealogía de una familia en la cual se muestra la relación existente entre los miembros de la familia. Por medio de este esquema se puede estudiar cómo se hereda un determinado rasgo o enfermedad. Un **heredograma** también es conocido como **pedigree** o **árbol genealógico**.

Ejemplo 1. Se crea la siguiente base de datos de dos familias aleatorias compuesta por 13 integrantes (8 individuos en la primera familia y 5 individuos para la segunda familia), donde cada individuo tiene la información de sus padres y su respectivo género:

La pimera familia está compuesta por 8 integrantes, con cada uno de los miembros etiquetados desde 101 hasta 108 (columna 2), y etiquetando el género (género masculino "1" y femenino "2"). De igual forma, la familia 2 está compuesta por 5 integrantes etiquetados desde 201 a 205.

En el heredograma de la Figura 2.1 se puede verificar que hay tres generaciones de padres a hijos, esto quiere decir que cuando hay una relación de padre a hijo se tiene un coeficiente de parentesco $(\frac{1}{2})^{r=1} = 0{,}50$, por ende cuando se tiene una segunda generación, es decir, de padres de los padres

Tabla 2.1: Datos con estructura familiar para dos familias

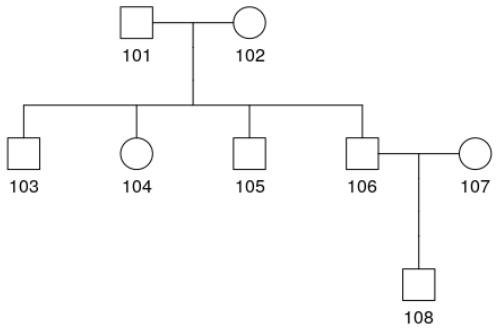
Familia	Individuo	Padre	Madre	Género
1	101	0	0	1
1	102	0	0	2
1	103	101	102	1
1	104	101	102	2
1	105	101	102	1
1	106	101	102	1
1	107	0	0	2
1	108	106	107	1
2	201	0	0	1
2	202	0	0	2
2	203	201	202	1
2	204	201	202	2
2	205	201	202	2

(abuelos) a los hijos tenemos $(\frac{1}{2})^{r=2} = 0{,}25$ y así se continúa con cada una de las generaciones que la familia posea; la diagonal principal posee 1 ya que corresponde a la relación genética del mismo individuo, a esta matriz se le conoce como la matriz de parentesco 2Φ. Para esta base de datos de dos familias, la matriz de parentesco 2Φ se vusualiza en la tabla 2.2.

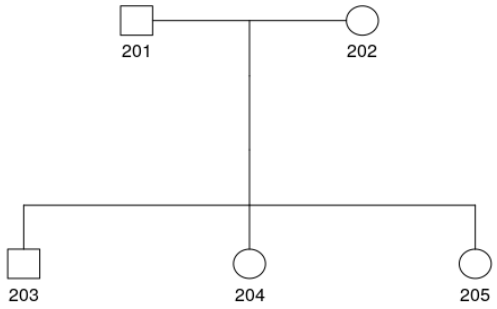
Tabla 2.2: Matriz de Parentesco para las familias descritas en la Tabla 2.1

	101	102	103	104	105	106	107	108	201	202	203	204	205
101	1.00	.	0.50	0.50	0.50	0.50	.	0.25	.	.	.	.	.
102	.	1.00	0.50	0.50	0.50	0.50	.	0.25	.	.	.	.	.
103	0.50	0.50	1.00	0.50	0.50	0.50	.	0.25	.	.	.	.	.
104	0.50	0.50	0.50	1.00	0.50	0.50	.	0.25	.	.	.	.	.
105	0.50	0.50	0.50	0.50	1.00	0.50	.	0.25	.	.	.	.	.
106	0.50	0.50	0.50	0.50	0.50	1.00	.	0.50	.	.	.	.	.
107	.	.	.	.	.	.	1.00	0.50	.	.	.	.	.
108	0.25	0.25	0.25	0.25	0.25	0.50	0.50	1.00	.	.	.	.	.
201	.	.	.	.	.	.	.	.	1.00	.	0.50	0.50	0.50
202	.	.	.	.	.	.	.	.	.	1.00	0.50	0.50	0.50
203	.	.	.	.	.	.	.	.	0.50	0.50	1.00	0.50	0.50
204	.	.	.	.	.	.	.	.	0.50	0.50	0.50	1.00	0.50
205	.	.	.	.	.	.	.	.	0.50	0.50	0.50	0.50	1.00

Ahora, cuando se tienen la información de varias familias, $n \geq 2$ que es el caso que interesa, se tendrá una matriz diagonal por bloques, donde cada uno de los bloques va a representar la matriz de parentesco de cada una de



(a) Heredograma de la primera familia, Tabla 2.1



(b) Heredograma de la segunda familia, Tabla 2.1

Figura 2.1: Heredogramas de cada familia de la Tabla 2.1

las familias:

$$2\Phi = \begin{pmatrix} F_1 & 0 & \cdots & 0 & 0 \\ 0 & F_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & F_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & F_n \end{pmatrix} \tag{2.22}$$

Cada F_i representa la matriz de parentesco para la familia i con $i = 1, \cdots, F$, suponiendo que se cuenta con una información de F familias.

Retomando el Modelo Poligénico descrito en 2.17 se pude definir también:

$$Cov[\epsilon] = I_n \sigma_e^2 \tag{2.23}$$

Donde σ_e^2 es la componente de varianza del error del modelo.

También, el **coeficiente de correlación intraclase** o **herdabilidad** fue escrito como:

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

Así, se puede definir la matriz de varianza descrita en 2.5, del Modelo Mixto general, para el modelo poligénico de la siguiente forma:

$$V = 2\Phi\sigma_g^2 + I_n\sigma_e^2 \tag{2.24}$$

$$= (2\Phi h_g^2 + I_n h_e^2)\sigma, \tag{2.25}$$

donde $\sigma^2 = \sigma_g^2 + \sigma_e^2$ y $h_g^2 + h_e^2 = 1$.

Capítulo 3

Modelo Mixto Poligénico: Aplicaciones

Para las aplicaciones de la teoría que se ha fundamentado sobre Modelos Lineales Mixtos, se utilizaron datos del Proyecto "Corazones de Baependi" (Processo Fapesp 2007/58150-7) del laboratorio de Genética y Cardiología Molecular (Incor/USP) Oliveira et al. (2008), Giolo et al. (2009). El objetivo general del proyecto es encontrar determinantes genéticos que modulan o regulan enfermedades cardiovasculares.

3.1. Descripción de los datos

Una descripción detallada del estudio de Baependi, puede ser encontrado en Egan K. et al. (2016).

Entre diciembre de 2005 y enero de 2006, un total de 81 familias fueron muestreadas en el municipio de Baependi, Minas Gerais, Brasil (aproximadamente, 1700 indiviudos). El tamaño de esas familias varía entre 3 y 156 miembros con una media de 21 individuos por familia; la edad media de los individuos fue de 44 años, variando de 18 a 100 años. La distribución de la muestra en relación al género fue 57 % mujeres (43 % hombres). En esa primera fase del estudio, varios fenótipos de los individuos de la población fueron evaluados, por ejemplo, glicemia, colesterol total, presión arterial, fracción de lipoproteinas, triglicéridos, entre otras. También, el ADN fué extraido de las muestras de sangre de cada individuo y almacenado para posterior genotiparlo.

3.2. Justificación de un análisis con datos genéticos

Considerando los trabajos de Oliveira et al. (2008) e Giolo et al. (2009), Horimoto A. et al. (2011 a, b), los resultados de los análisis realizados por los autores, indican efectos poligénicos significantes en las variables respuesta evaluadas, en otras palabras, la herdabilidad de los fenótipos fue considerada significativa.

Algunas de las variables analizadas en el estudio están descritas en la Tabla (3.1) junto con las estimativas de herdabilidad y los componentes de varianza, obtenidos mediante el ajuste de modelos poligénicos específicos. Para la aplicacion se usó la variable $PASmedia$ que corresponde a la media de medidas de presión arterial sistólica en milimetros de mercurio (mmHg).

Tabla 3.1: Herdabilidad de algunas variables del Proyecto

Y	Covariables	Herdabilidad	σ_g^2	σ_e^2
ln(SBP)	Ninguna	0.1532	0.1617	0.8941
	Sexo	0.1563	0.1602	0.8649
	Sexo, Edad	0.2582	0.2035	0.5847
ln(DBP)	Ninguna	0.1492	0.1478	0.8428
	Sexo	0.1489	0.1472	0.8414
	Sexo, Edad	0.2052	0.1867	0.7231
ln(LDL)	Ninguna	0.2693	0.2589	0.7024
	Sexo	0.2682	0.2560	0.6984
	Sexo, Edad	0.2638	0.2421	0.6758
ln(HDL)	Ninguna	0.3132	0.4126	0.9046
	Sexo	0.3123	0.4062	0.8945
	Sexo, Edad	0.3217	0.4174	0.8800
PASmedia	Ninguna	0.1689	63.368	311.716
	Sexo	0.1729	63.338	302.933
	Sexo, Edad	0.2837	81.819	206.616

Donde las otras variables de la tabla anterior son:

- **SBP**: Sistolic blood pressure - presión arterial sistólica (mmHg).
- **DBP**: Diastolic blood pressure - presión arterial diastólica (mmHg).
- **LDL**: Colesterol Malo.
- **HDL**: Colesterol Bueno.

Considerando los valores obtenidas de herdabilidad, los resultados justifican análisis de determinantes genéticos que regulan estos fenótipos.

Así, 1120 individuos fueron genotipados, obteniéndose una muestra del genoma de más de 900.000 marcadores moleculares SNP. Algunos estudios con esta cantidad de marcadores fueron realizados y están siendo analizados los resultados, más desde el punto de vista biológico para posterior publicación.

El proyecto también cuenta con un banco de datos imputados con el objetivo de refinar el genoma con marcadores moleculares. Este nuevo conjunto de marcadores tiene aproximadamente 6 millones de SNPs, siendo una fuente muy valiosa para estudios genéticos, entre ellos la búsqueda de los determinantes genéticos que influyen en los fenótipos bajo estudio.

Cabe resaltar que los bancos de datos a ser analizados es conformado por el conjunto de fenótipos y la plataforma de SNPs los cuales son de altísima dimensión, siendo necesarias herramientas computacionales de alto desempeño fuera de algoritmos paralelizados, para disminuir el tiempo de procesamiento y obtener los resultados deseados.

Para las aplicaciones se tomó el fenótipo $PASmedia$ (mmHg) como variable respuesta que corresponde a la media de medidas de presión arterial sistólica de los individuos de la población. También se tomó la plataforma de SNPs imputados, los cuales representan la muestra del genoma humano a través de los 22 cromosomas. El interés específico en el estudio del presente trabajo es encontrar determinantes genéticos o SNPs que posiblemente estén influenciando esta variable que representa una enfermedad cardiovascular que aqueja a gran parte de la población.

Para esta variable $PASmedia$ (mmHg) se obtienen los siguientes resultados acerca del comportamiento de la variable:

- Mínimo: 79.7
- Primer Cuartil (Q_1): 112.3
- Segundo Cuartil (Q_2) o Mediana: 124.7
- Media: 126.9
- Tercer Cuartil (Q_3): 137.3
- Máximo: 216.3

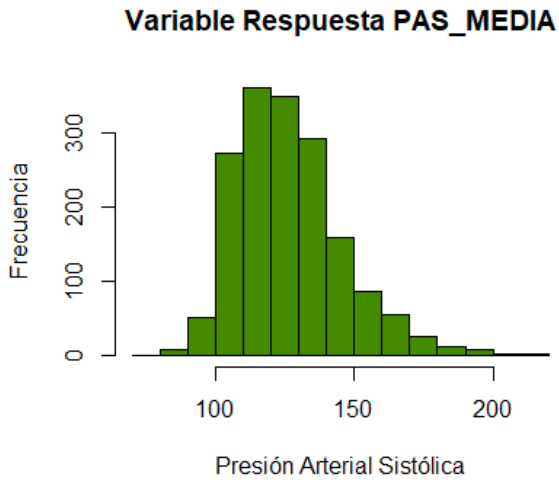


Figura 3.1: Histograma del Comportamiento de $PASmedia$ (mmHg)

Donde, de forma general podemos verificar valores extemos, por ejemplo se tiene un máximo de 216.3 lo cual es muy superior a los valores normales, ya que en los humanos una presión sistólica normalmente debe estar por debajo de 120 mm Hg.

3.3. Descripción del modelo poligénico utilizado para la obtención de resultados

Dado que el interés específico en el estudio es conocer cuáles SNPs están influenciando la variable $PASmedia$ (mmHg), se realizó el ajuste del modelo mixto poligénico descrito en (2.17). Con el objetivo de capturar el efecto genético, además de cada SNP (un SNP en cada ajuste), otras covariables fueron incluídas: género, edad, índice de masa corporal y 4 covariables para corrección de la estructura genética. Estas 4 covariables, corresponden específicamente a Componentes Principales, pues, específicamente en el área de genética, las Componentes Principales son usadas para obtener escores en una determinada población con el objetivo de controlar la estratificación poblacional, con el objetivo de encontrar resultados más confiables en los modelos y las respectivas estimativas de los parámetros.

Denotando por β_{snp} el efecto del SNP sobre la variable $PASmedia$ (mmHg), la hipótesis que se prueba es

$$H_0 : \beta_{snp} = 0 \tag{3.1}$$
$$H_1 : \beta_{snp} \neq 0, \tag{3.2}$$

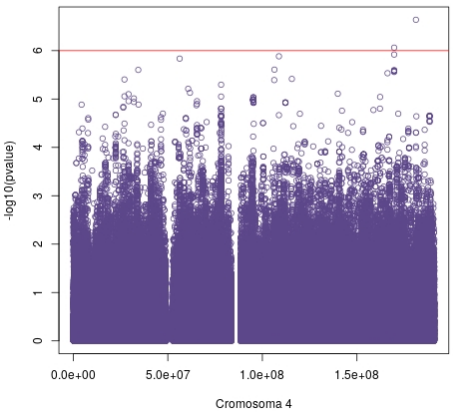
Para cada SNP su respectivo p-valor asociado con el parámetro β_{snp} fue calculado. En la literatura se recomienda que el punto de corte del p-valor sea menor que 5×10^{-8} (Barsh G. et al., 2012) para considerar que un SNP es significativamente asociado con la variable respuesta, en este caso con $PASmedia$ (mmHg).

3.4. Resultados

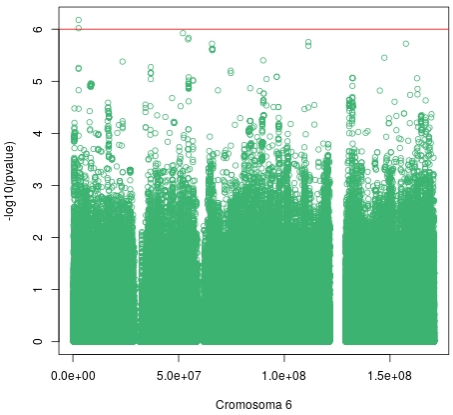
Después de ajustar el modelo descrito y haber dejado rodando el algoritmo que está descrito en el Apéndice B, durante mucho tiempo, debido a la cantidad de covariables genéticas o SNPs a través de los cromosomas, se obtuvieron los resultados que serán descritos a seguir. Los correspondientes p-valores para los SNPs de los cromosomas estudiados (4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 y 19) se encuentran en las figuras (3.4), (3.3), (3.4), (3.5), (3.6), (3.7), (3.8), (3.9), (3.10) y (3.11)

En el área de aplicaciones en genética este tipo de gráfico es conocido como **gráfico de Manhattan**. En cada gráfico, la línea horizontal representa el punto de corte de 1×10^{-6} (tomando logaritmo a los p-valores). En el eje X se representa la posición de cada uno de los SNPs, dada en centimorgans. Por su parte, en el eje Y se representa el valor de logaritmo en la base 10 del p-valor. El hecho de tomar logaritmo, es simplemente para visualización de los resultados y para tener una idea intuitiva de la región donde están los marcadores que alcanzan o ultrapasan el límite.

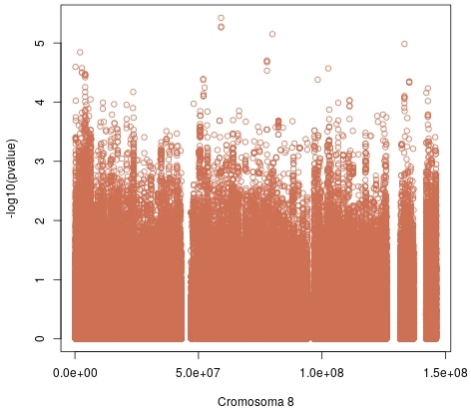
Por otro lado para los cromosomas 7, 9, 10, 11, 12, 17, 18 y 19 se han incluido dos gráficos que se denominan **Regional Plot** para dos SNP que poseen los p-valores más altos en el respectivo cromosoma. En este gráfico cada punto representa un SNP (marcador genético) y en la coordenada Y se indica el nivel de significancia para cada SNP (p-valor). Para cada SNP la correlación o desequilibrio de ligación medido como r^2 se indica por la estructura del color. Finalmente, la anotación genética, o los genes que están en la región discriminada son mostrados en la parte inferior del grafico.



(a) Cromosoma 4

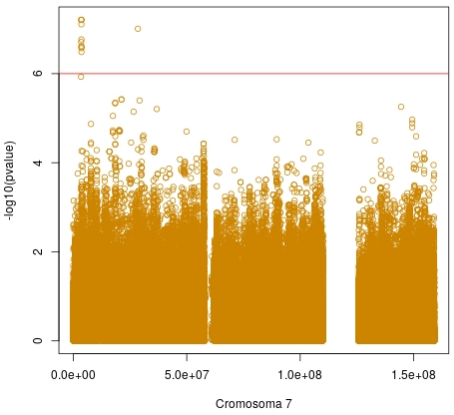


(b) Cromosoma 6

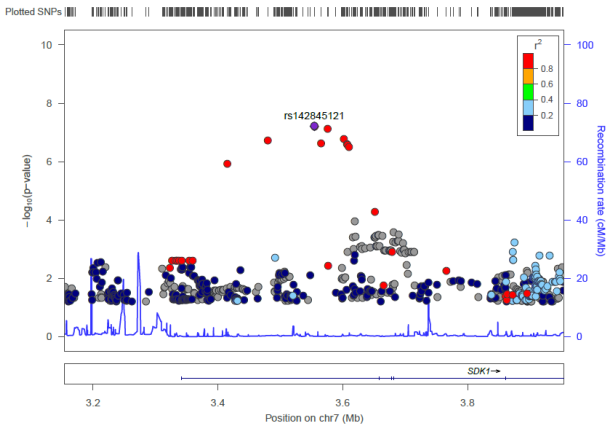


(c) Cromosoma 8

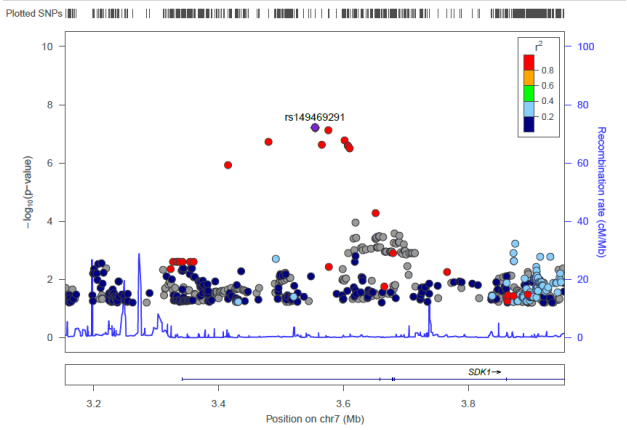
Figura 3.2: Posiciones de los SNPs más significativos para los cromosomas 4, 6 y 8.



(a) Cromosoma 7

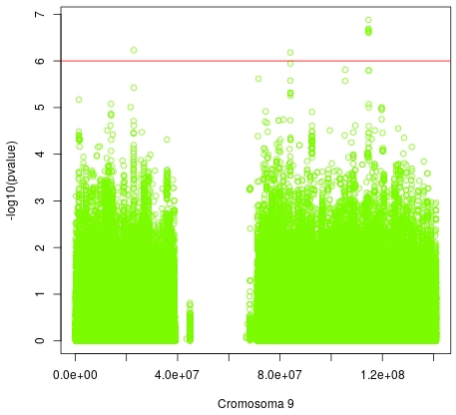


(b) Regional Plot 1 Cromosoma 7

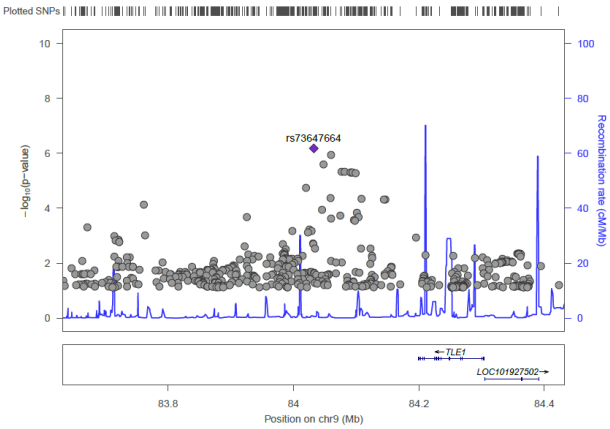


(c) Regional Plot 2 Cromosoma 7

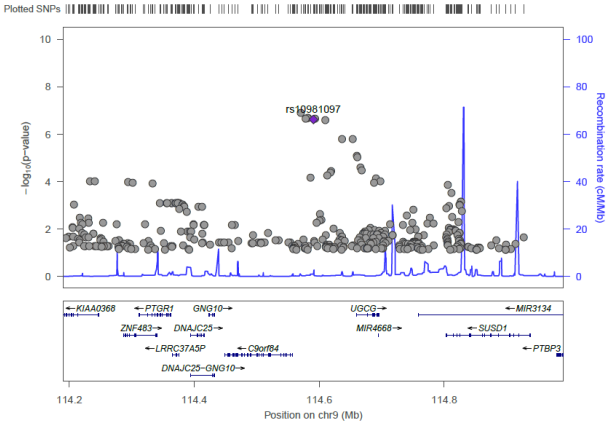
Figura 3.3: Posiciones de los SNPs más significativos para el cromosoma 7 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



(a) Cromosoma 9

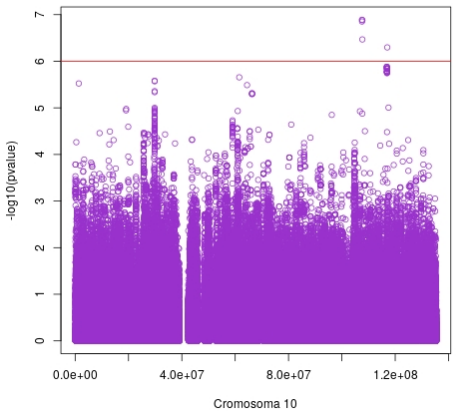


(b) Regional Plot 1 Cromosoma 9

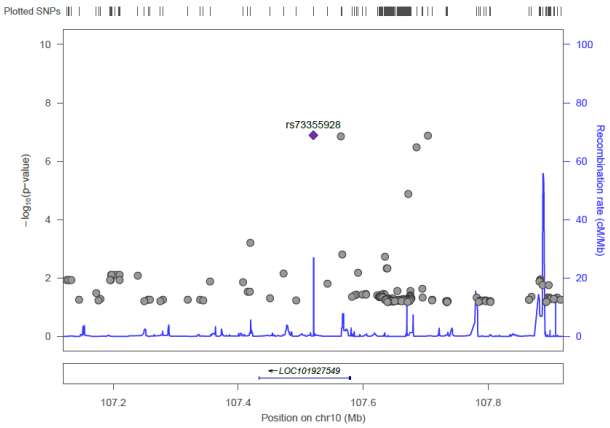


(c) Regional Plot 2 Cromosoma 9

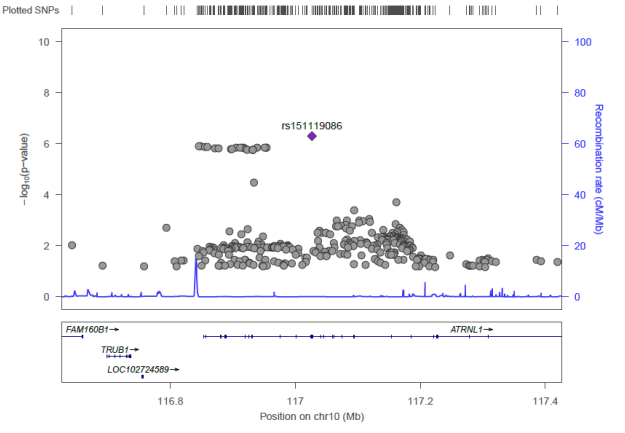
Figura 3.4: Posiciones de los SNPs más significativos para el cromosoma 9 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



(a) Cromosoma 10

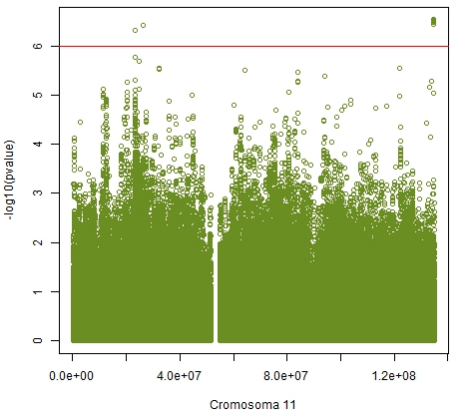


(b) Regional Plot 1 Cromosoma 10

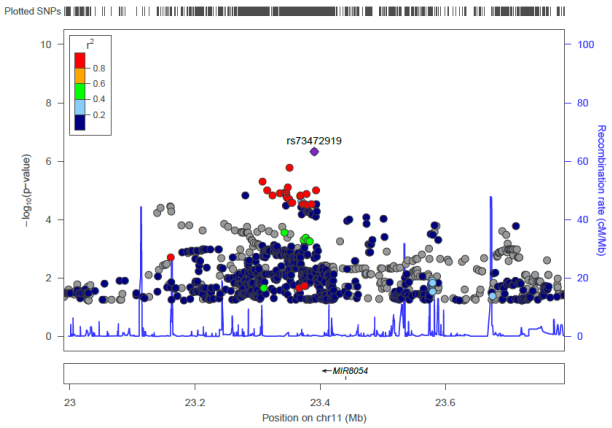


(c) Regional Plot 2 Cromosoma 10

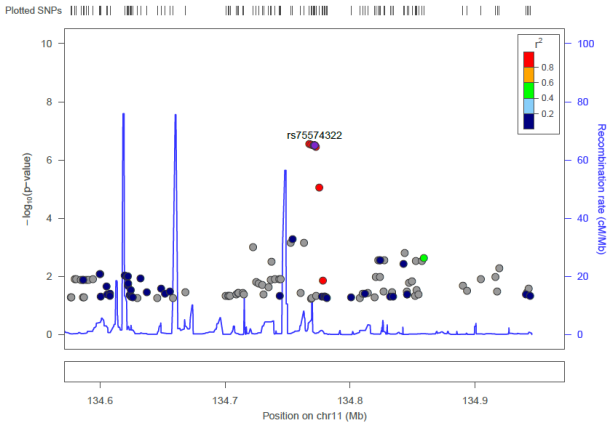
Figura 3.5: Posiciones de los SNPs más significativos para el cromosoma 10 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



(a) Cromosoma 11

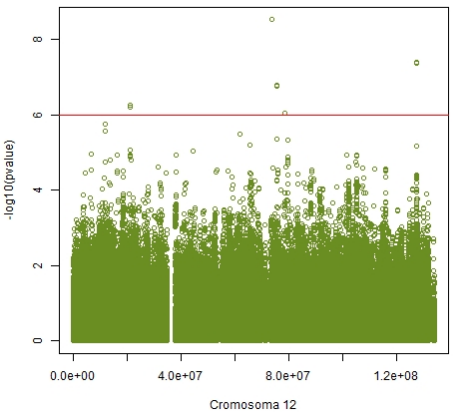


(b) Regional Plot 1 Cromosoma 11

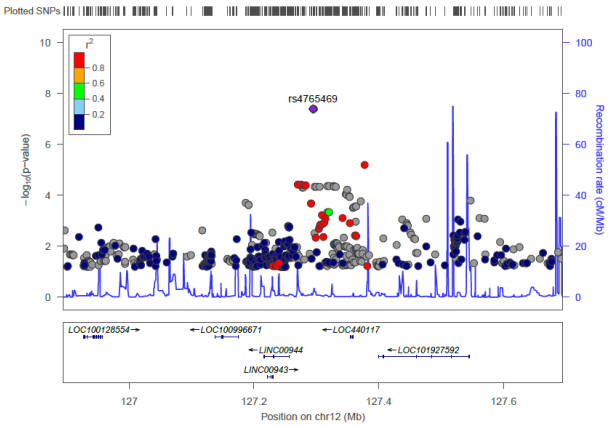


(c) Regional Plot 2 Cromosoma 11

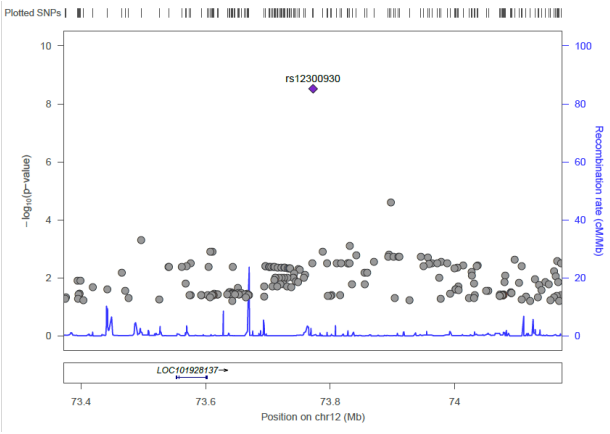
Figura 3.6: Posiciones de los SNPs más significativos para el cromosoma 11 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



(a) Cromosoma 12

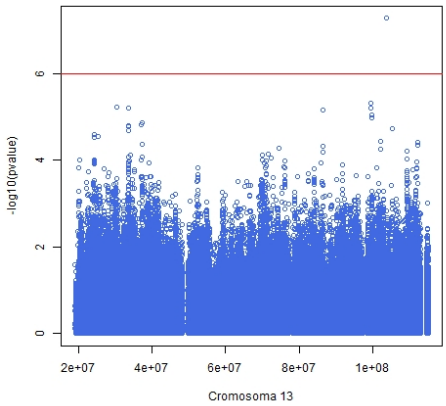


(b) Regional Plot 1 Cromosoma 12

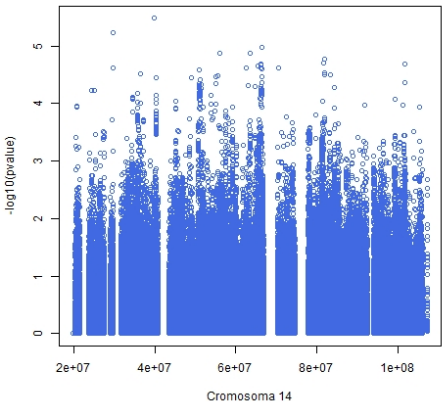


(c) Regional Plot 2 Cromosoma 12

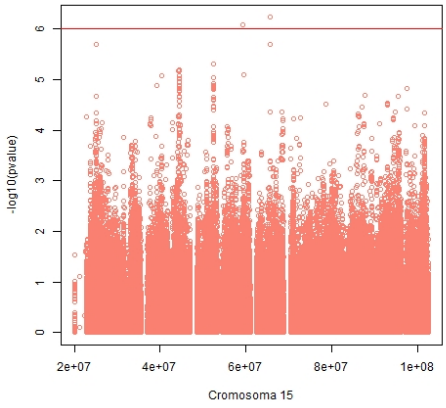
Figura 3.7: Posiciones de los SNPs más significativos para el cromosoma 12 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



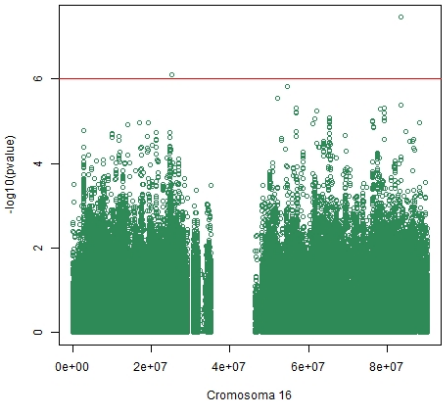
(a) Cromosoma 13



(b) Cromosoma 14

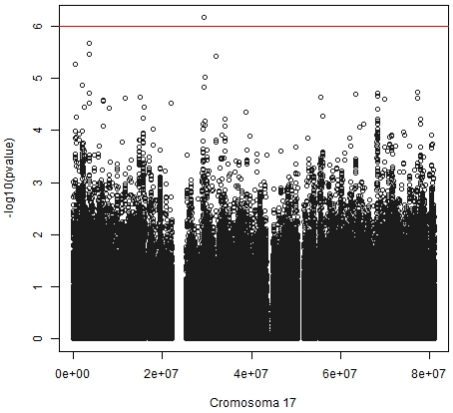


(c) Cromosoma 15

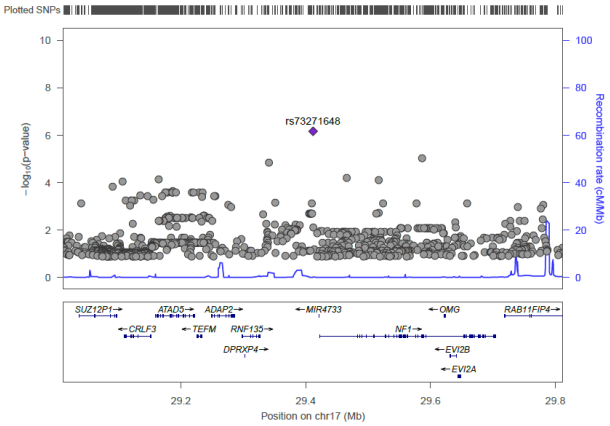


(d) Cromosoma 16

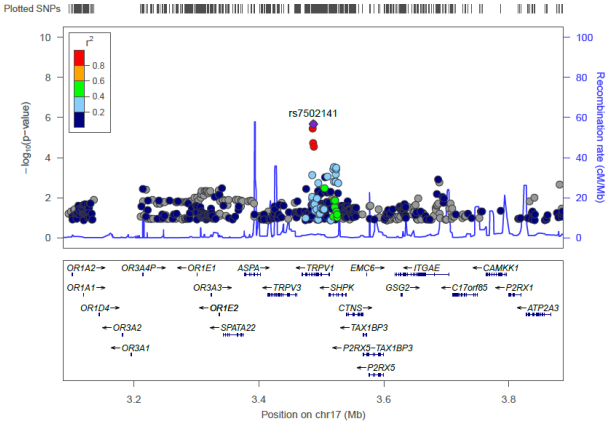
Figura 3.8: Posiciones de los SNPs más significativos para los cromosomas 13, 14, 15 y 16.



(a) Cromosoma 17

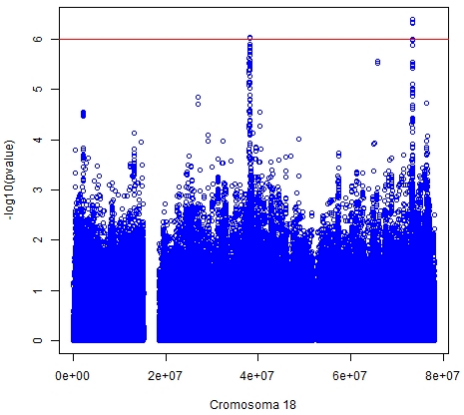


(b) Regional Plot 1 Cromosoma 17

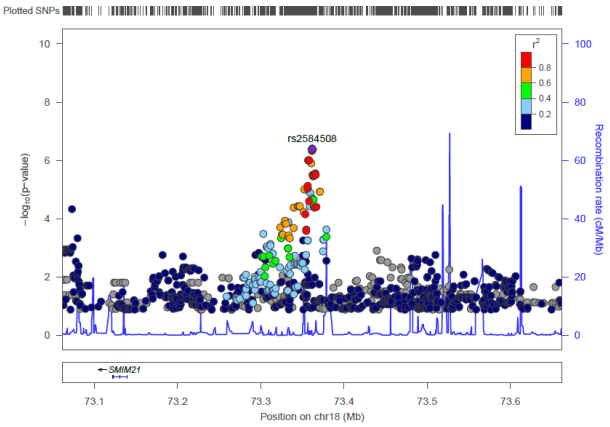


(c) Regional Plot 2 Cromosoma 17

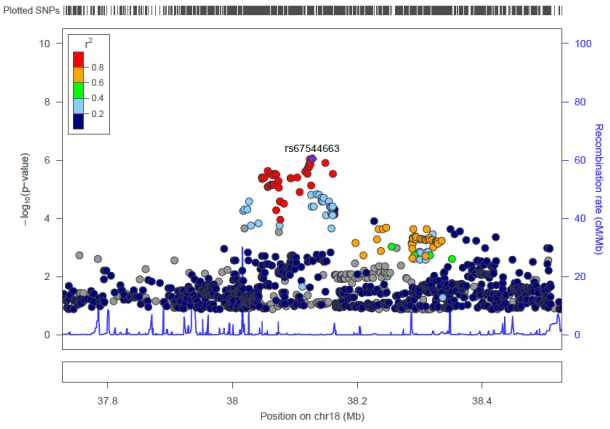
Figura 3.9: Posiciones de los SNPs más significativos para el cromosoma 17 y Regional Plot Correspondiente a dos SNP con mayor p-valor



(a) Cromosoma 18

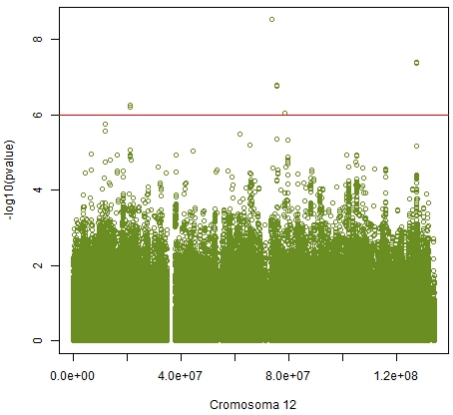


(b) Regional Plot 1 Cromosoma 18

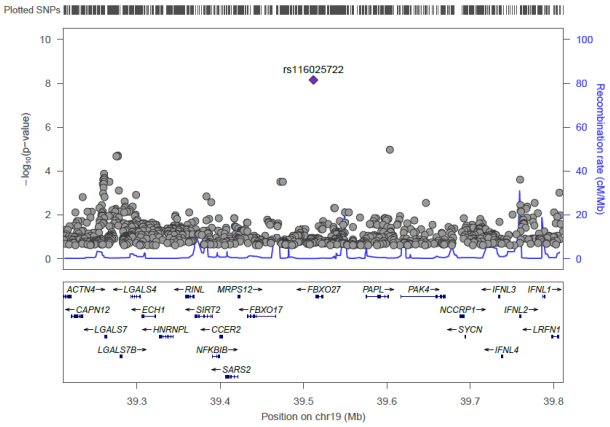


(c) Regional Plot 2 Cromosoma 18

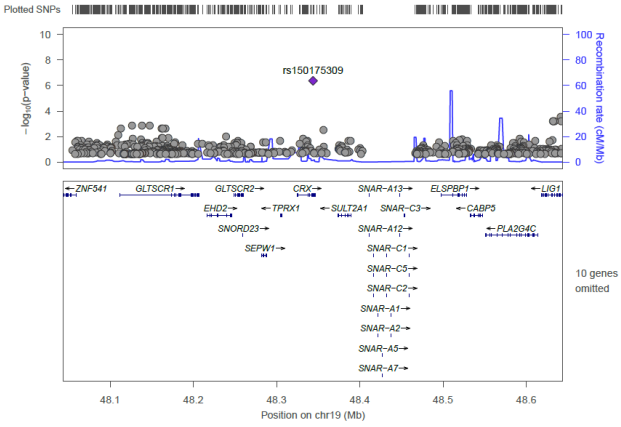
Figura 3.10: Posiciones de los SNPs más significativos para el cromosoma 18 y Regional Plot Correspondiente a dos SNP con mayor p-valor.



(a) Cromosoma 19



(b) Regional Plot 1 Cromosoma 19



(c) Regional Plot 2 Cromosoma 19

Figura 3.11: Posiciones de los SNPs más significativos para el cromosoma 19 y Regional Plot Correspondiente a dos SNP con mayor p-valor.

En las tablas (3.2) y (3.3) los SNPs con p-valores menores que 1×10^{-6} son presentados, indicando el cromosoma al cual pertenecen, la posición así como el parámetro β_{snp} y su respectiva desviación.

Para estas tablas se tienen 6 columnas con los resultados del ajuste del modelo en que los SNPs tuvieron el logaritmo del p-valor mayor que 1×10^{-6} junto con otros resultados, específicamente:

- **CHR**: Cromosoma correspondiente.
- **BP**: Posición del SNP, dado en pares de bases.
- **SNP**: Nombre del SNP correspondiente.
- **beta**: Valor del β asociado al SNP.
- **desvio**: desviación estándar.
- **p**: $p - valor$ correspondiente al SNP en el modelo ajustado.

	CHR	BP	SNP	beta	desvio	p
1	4	181179162	rs187593148	21.70	4.19	2.3004e-07
2	4	169674905	rs140713942	50.44	10.25	8.67293e-07
3	6	2663067	rs76561277	13.39	2.69	6.63078e-07
4	6	2663536	rs115734042	13.24	2.70	9.58746e-07
5	**7**	**3554458**	**rs142845121**	**38.96**	**7.20**	**6.17494e-08**
6	**7**	**3554411**	**rs149469291**	**38.96**	**7.20**	**6.17494e-08**
7	**7**	**3554959**	**rs139304917**	**38.93**	**7.19**	**6.23605e-08**
8	**7**	**3575857**	**rs189292863**	**38.19**	**7.11**	**7.79484e-08**
9	**7**	**28478543**	**rs190010851**	**32.89**	**6.17**	**9.81556e-08**
10	7	3601985	rs189815424	38.25	7.31	1.70976e-07
11	7	3479867	rs78611018	37.60	7.23	1.95155e-07
12	7	3565174	rs80002219	37.81	7.32	2.41864e-07
13	7	3606933	rs150085342	37.85	7.34	2.53397e-07
14	7	3607794	rs76200863	37.82	7.35	2.68458e-07
15	7	3610056	rs141063055	37.57	7.36	3.26822e-07
16	9	114570723	rs12379756	42.61	8.08	1.31415e-07
17	9	114582239	rs7874520	38.01	7.32	2.04951e-07
18	9	114581260	rs10981090	37.97	7.32	2.10324e-07
19	9	114578471	rs76894475	37.85	7.31	2.23437e-07
20	9	114593316	rs114356224	37.81	7.31	2.32368e-07
21	9	114590748	rs10981097	37.80	7.32	2.3851e-07
22	9	114609664	rs10981117	37.69	7.31	2.53587e-07
23	9	22913412	rs76605761	4.58	0.92	5.89832e-07
24	9	84032343	rs73647664	12.89	2.59	6.61767e-07
25	10	107520417	rs73355928	17.56	3.33	1.28887e-07
26	10	107703734	rs111616864	16.96	3.21	1.31773e-07
27	10	107564406	rs73357953	17.52	3.33	1.40336e-07
28	10	107685495	rs56686074	15.38	3.02	3.4046e-07
29	10	117026701	rs151119086	18.26	3.63	5.04377e-07
30	11	134767434	rs61280035	24.33	4.74	2.92379e-07
31	11	134769159	rs11822045	24.32	4.75	3.03201e-07
32	11	134770749	rs58519850	24.22	4.74	3.20246e-07
33	11	134771189	rs79581908	24.21	4.74	3.21969e-07
34	11	134771880	rs75140384	24.20	4.74	3.23278e-07
35	11	134771879	rs76686000	24.20	4.74	3.23278e-07
36	11	134771776	rs75574322	24.20	4.74	3.23647e-07
37	11	134772779	rs77580034	24.09	4.74	3.68301e-07
38	11	26089942	rs190849611	36.38	7.16	3.72612e-07
39	11	23390521	rs73472919	6.20	1.23	4.69403e-07

Tabla 3.2: P valores de los marcadores moleculares que superan el 1*e-06 (1)

	CHR	BP	SNP	beta	desvio	p
40	**12**	**73772439**	**rs12300930**	**42.50**	**7.17**	**3.01321e-09**
41	**12**	**127294750**	**rs4765469**	**-12.92**	**2.36**	**4.1427e-08**
42	**12**	**127295750**	**rs1355552**	**-12.90**	**2.35**	**4.19612e-08**
43	12	75561010	rs76593584	25.39	4.84	1.59725e-07
44	12	75559563	rs74703229	25.36	4.84	1.62717e-07
45	12	75557931	rs114855629	25.36	4.84	1.66199e-07
46	12	75556005	rs116165464	25.45	4.87	1.69854e-07
47	12	21395277	rs56397921	24.95	4.98	5.51865e-07
48	12	21402026	rs74064273	24.83	4.98	6.21916e-07
49	12	78352291	rs114220039	16.62	3.39	9.17455e-07
50	**13**	**103817933**	**rs113060440**	**37.10**	**6.82**	**5.32176e-08**
51	15	65561094	rs115886642	26.80	5.37	5.87252e-07
52	15	59326555	rs4775112	-39.28	7.97	8.34559e-07
53	**16**	**83385947**	**rs74804557**	**42.06**	**7.63**	**3.47736e-08**
54	16	25223852	rs111309315	26.41	5.35	7.99681e-07
55	17	29411741	rs73271648	17.43	3.51	6.86946e-07
56	18	73361902	rs2584508	-2.59	0.51	4.06154e-07
57	18	73361703	rs2584507	-2.59	0.51	4.06297e-07
58	18	73362253	rs2849802	-2.58	0.51	4.509e-07
59	18	73361191	rs2584300	-2.58	0.51	4.75373e-07
60	18	38128002	rs67544663	3.40	0.69	9.28294e-07
61	18	38124106	rs4465667	3.34	0.68	9.49964e-07
62	**19**	**39511773**	**rs116025722**	**31.90**	**5.51**	**7.02051e-09**
63	19	48343828	rs150175309	43.93	8.69	4.26528e-07

Tabla 3.3: P valores de los marcadores moleculares que superan el 1*e-06 (2)

3.4.1. Conclusiones del problema aplicado

1. De acuerdo con los resultados de las tablas hay presencia de p-valores por encima de 1×10^{-7} en los cromosomas 7, 12, 13, 16 y 19 lo cual representa hallazgos relevantes que pueden ayudar en investigaciones genéticas. Se destacan las regiones marcadas en negro p-valores por encima de 1×10^{-8}.

2. Es importante anotar que los SNPs significativos estadísticamente, indican una posible región que debe ser analizada con mucho cuidado para determinar con veracidad las variantes genéticas que modulan la variable respuesta.

3. Podemos concluir que a través de los gráficos se presentan 2 regiones del cromosoma 7 con una alta correlación entre los SNP que tienen un mayor p-valor, al igual que en el cromosoma 11, cromosoma 12 y principalmente en el cromosoma 18. Esto muestra evidencias de altas correlaciones entre los SNP de estas regiones, es posible que tengan un gen que está asociado PASmedia (presión arterial sistólica). Cabe resaltar que en la región discriminada del cromosoma 18 existe un gen llamado SMIM21 el cual ya ha sido descubierto es muchos estudios, por ello recomendamos que esta región pueda ser estudiada desde el punto de vista genético con más detalle.

4. Cabe resaltar que un estudio de este tipo solamente es una fase inicial de un estudio genético, pero que ayuda enormemente al geneticista a concentrarse en estudios posteriores de los pocos SNPs (si los hay!!) significativos estadísticamente, ya que se espera que apunten a los genes que regulan la variable en estudio. Son varias metodologías, desde el punto de vista genético, que se realizan como un análisis posterior a este estudio.

5. No se pudieron terminar los análisis de los cromosomas restantes por falta de un servidor y tiempo, ya que estos analisis tuvieron una duración aproximadamente de 6 meses en una máquina convencional.

Capítulo 4

Modelos Lineales Generalizados

Nelder J. & Wedderburn R. (1972) muestran una serie de técnicas específicas, comúnmente estudiadas separadamente, para ser formuladas de manera unificada, como una clase de modelos de regresión. A esta teoría unificadora de modelado estadístico, dieron el nombre de **Modelos Lineales Generalizados** (MLG). Estos modelos involucran una variable respuesta univariada, variables explicativas y una muestra aleatoria de n observaciones independientes:

- Una variable respuesta o componente aleatorio del modelo, de una distribución perteneciente a la familia de distribuciones exponenciales biparamétricas en el componente aleatorio del modelo que engloba las distribuciones normal, gamma y normal inversa para datos continuos; binomial para proporciones; Poisson y binomial negativa para conteos:

$$f(y;\theta,\phi) = \exp\{\phi[y\theta - b(\theta)] + c(y,\phi)\} \tag{4.1}$$

  En donde $b(.)$ y $c(.)$ son funciones conocidas, además cuando ϕ es conocida la función anterior es idéntica a la familia exponencial de forma canónica.

- Las variables explicativas entran de una forma de estructura lineal, constituyendo el componente sistemático del modelo.

- Una **ligación o enlace** entre los componentes aleatorios y sistemático es a través de una función adecuada, como por ejemplo, logarítmica para los modelos log-lineales, llamada **función de ligación o enlace**.

Ejemplos

- **Distribución Normal:** Sea Y una variable aleatoria con distribución normal de media μ y varianza σ^2. Una función de densidad de Y se expresaría de la siguiente forma:

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu^2)\right] = \exp\left[\frac{1}{\sigma^2}\left(\mu y - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\ln 2\pi\sigma^2 + \frac{y^2}{\sigma^2}\right)\right]$$

donde:

$$\begin{aligned}\theta &= \mu \\ b(\theta) &= \frac{\mu^2}{2} = \frac{\theta^2}{2} \\ \phi &= \sigma^2 \\ c(y,\phi) &= \frac{1}{2}\ln 2\pi\phi - \frac{y^2}{\phi^2}\end{aligned}$$

Ademas vemos que:

$$\begin{aligned}E(Y) &= \mu = \theta = \frac{\partial}{\partial\theta}\left(\frac{\theta^2}{2}\right) = b'(\theta) \\ V(Y) &= \sigma^2 = (\phi)(1) = \phi b''(\theta)\end{aligned}$$

- **Distribución Poisson:** Sea Y una variable aleatoria con distribución Poisson $Y \sim P(\mu)$ tenemos que la función de probabilidad está dada por:

$$\frac{e^{-\mu}\mu^y}{y!} = \exp\left(y\ln\mu - \mu - \ln y!\right)$$

Con ello podemos tomar cada una de las componentes de la familia exponencial:

$$\begin{aligned}\theta &= \ln\mu \Rightarrow \mu = e^\theta \\ b(\theta) &= \mu = e^\theta \\ \phi &= 1 \\ E(Y) &= b'(\theta) = e^\theta = \mu \\ V(Y) &= \frac{\partial\mu}{\partial\theta} = \frac{\partial}{\partial\theta}(e^\theta) = e^\theta = \mu\end{aligned}$$

- **Distribución Binomial:** Sea y la proporción de éxitos en n el número de ensayos independientes cada uno con probabilidad μ, entonces $ny \sim B(n, \mu)$ su función de densidad está dada por:

$$\binom{n}{ny}\mu^{ny}(1-\mu)^{n-ny} = \exp\left(\ln\binom{n}{ny} + ny\ln\left(\frac{\mu}{1-\mu}\right) + n\ln(1-\mu)\right)$$

Donde tenemos las componentes de la familia exponencial:

$$\begin{aligned}
\theta &= \ln\left(\frac{\mu}{1-\mu}\right) \\
b(\theta) &= \ln(1+e^{\theta}) \\
\phi &= n \\
E(Y) &= b'(\theta) = \frac{e^{\theta}}{e^{\theta}+1} = \mu \\
V(Y) &= \frac{e^{\theta}}{(e^{\theta}+1)^2} = \mu(1-\mu)
\end{aligned}$$

- Ahora se puede realizar el mismo procedimiento en las distribuciones Gama, Normal Inversa y se obtienen los siguientes resultados:

Tabla 4.1: Elementos de las distribuciones pertenecientes a la familia exponencial

Distribución	$b(\theta)$	θ	ϕ	$V(\mu)$
Normal	$\frac{\theta^2}{2}$	μ	σ^2	1
Poisson	e^{θ}	$\ln\mu$	1	μ
Binomial	$\ln(1+e^{\theta})$	$\ln\left\{\frac{\mu}{1-\mu}\right\}$	n	$\mu(1-\mu)$
Gama	$-\ln(-\theta)$	$-\frac{1}{\mu}$	$\frac{1}{(CV)^2}$	μ^2
Normal Inversa	$-\sqrt{-2\theta}$	$-\frac{1}{2\mu^2}$	ϕ	μ^3

Con lo anterior se tiene la siguiente definición:

Definición 1. Los modelos lineales generalizados pueden ser usados cuando se tiene una única variable aleatoria Y asociada a un conjunto de variables explicativas $x_1, \ldots, x_p$, para una muestra de n observaciones (y_i, x_i) en que $x_i = (x_{i1}, \ldots, x_{ip})^T$ del vector columna de variables explicatorias, donde:

1. **Componente aleatorio**: Representado por un conjunto de variables aleatorias independientes $Y_1, \ldots, Y_n$ provenientes de una misma distribución que hace parte de una familia de distribuciones (4.1) con medias $\mu_1, \ldots, \mu_n$.

$$E(Y_i) = \mu_i, \qquad i = 1, \ldots, n.$$

Siendo $\theta > 0$ un parámetro de dispersión y el parámetro θ_i denominado el parámetro canónico. Una función de densidad de probabilidad de Y_1 dada por

$$f(y_i; \theta_i, \phi) = \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}$$

Siendo $b(.)$ y $c(.)$ funciones conocidas, además bajo las condiciones de regularidad:

$$E(Y_i) = \mu_i = b'(\theta_i)$$
$$Var(Y_i) = \phi^{-1}b''(\theta_i) = \phi^{-1}V_i$$

Donde $V_i = V(\mu_i) = d\mu_i/d\theta_i$ es denominada función de varianza y depende únicamente de la media μ_i. El parámetro natural θ_i puede ser expresado como:

$$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i) \tag{4.2}$$

Siendo $q(\mu_i)$ una función conocida de medida μ_i. Dada una relación funcional para una funcion de varianza $V(\mu)$, y un parámetro canonico y obtenido de la ecuación 4.2 y la distribución se determina en la función exponencial 4.1.

2. **Componente Sistemático**: las variables explicativas entran en la forma de una suma lineal de sus efectos:

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_j = \boldsymbol{x}_i^T\boldsymbol{\beta} \qquad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{4.3}$$

Siendo $\boldsymbol{X} = (x_1, \ldots, x_n)^T$ una matriz del modelo, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ un vector de parametros y $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ un predictor lineal.

3. **Función de Enlace o Ligación**: Es una función que relaciona el componente aleatorio con el componente sistemático, es decir, realiza la media al predictor lineal, esto es:

$$\eta_i = g(\mu_i) \tag{4.4}$$

Siendo $g(.)$ una función monotona y diferenciable.

4.1. Funciones de Enlace canónicas

Dado ϕ conocido, el logaritmo de una función de verosimilitud de un MLG con respuestas independientes puede ser expresado de la forma (McCullagh, P. & Nelder J. 1989):

$$\begin{aligned} f(y_i;\theta_i,\phi) &= \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\} \\ \prod_i f(y_i;\theta_i,\phi) &= \prod_i \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\} \\ \mathscr{L}(\beta) &= \exp\left\{\sum_i \phi[y_i\theta_i - b(\theta_i)] + \sum_i c(y_i,\phi)\right\} \end{aligned}$$

Tomamos el logaritmo a ambos lados de la igualdad para obtener la función de verosimilitud, con ϕ conocida:

$$L(\beta) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i,\phi) \tag{4.5}$$

El caso canónico sucede cuando el parametro (θ) coincide con el predictor lineal, es decir, cuando

$$\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

Tenemos que:

$$L(\beta) = \sum_{i=1}^{n} \phi\left\{ y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right\} + \sum_{i=1}^{n} c(y_i,\phi)$$

Definiendo un estádistico $S_j = \phi\sum_{i=1}^{n} y_i x_{ij}$:

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{p} s_j\beta_j - \phi\sum_{i=1}^{n} b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{i=1}^{n} c(y_i,\phi)$$

Por el teorema de factorización de Neyman-Fisher ($f(y;\theta,\phi) = g(t,\theta)h(y_1,\ldots,y_n)$), además usando el teorema de Lehmann-Scheffé el estadístico $\boldsymbol{S} = (S_1,\ldots.S_p)^T$ es suficiente mínimo para el vector $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$.

Además dos conseguencias importantes que son la esperanza y la varianza de la función de verosimilitud de un MLG es:

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0$$
$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) + E\left(\frac{\partial L}{\partial \theta}\right)^2 = 0$$

Donde tomamos la función de verosimilitud general:

$$L(\theta; y) = \phi(y\theta - b(\theta)) + c(y; \phi)$$

Así se tiene como resultado para la esperanza:

$$\phi(\mu - b'(\theta)) = 0$$
$$\mu = b'(\theta)$$

Ahora se define la varianza:

$$\phi^2 Var(Y) - \phi b''(\theta) = 0$$
$$Var(Y) = b''(\theta)\phi^{-1}$$

Las funciones de enlace corresponden a estos esdísticos que son llamados enlaces canónicos y desempeñan un papel importante en la teoría.

Tabla 4.2: Funciones de enlace canónicas

Distribución	Función de enlace canónicas $\theta = \eta$
Normal	$\mu = \eta$
Binomial	$\log\left(\frac{\mu}{1-\mu}\right) = \eta$
Poisson	$\log\ \mu = \eta$
Gama	$\mu^{-1} = \eta$
Normal Inversa	$\mu^{-2} = \eta$

Como la función de enlace es biyectiva podremos invertirla, obteniendo:

$$\mu = g^{-1}(\eta) = g^{-1}(\boldsymbol{x'\beta})$$

Además en el caso binomial, se tienen otras funciones de enlace

- **Función de enlace Probit:** Sea μ la media de éxitos de una sucesión de una distribución normal, la función de enlace probit está definida por:

$$\Phi^{-1}(\mu) = \eta \tag{4.6}$$

Donde $\Phi(.)$ es la función de distribución acumulada de la normal estándar, ya que se asume X tiene una distribución normal estándar, con media $\mu \in \mathbb{R}$ y varianza $\sigma^2 > 0$:

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Entonces, tomando $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$$\pi_i = P(X \leq x_i) = P\left(Z \leq -\frac{\mu}{\sigma} + \frac{1}{\sigma}x_i\right) = P(Z \leq \beta_1 + \beta_2 x_i)$$

Para $\beta_1 = -\frac{\mu}{\sigma}$ y $\beta_2 = \frac{1}{\sigma}$, así:

$$\pi_i = \Phi(\beta_1 + \beta_2 x_i)$$
$$probit(\pi_i) = \Phi^{-1}(\pi_i) = \beta_1 + \beta_2 x_i$$

- **Función de enlace Logit:** La función de densidad de la distribución logística está definida por:

$$f(y) = \frac{\exp(y)}{[1+\exp(y)]^2}$$

Donde $-\infty < y < \infty$, con ello podemos tomar $y = \frac{x-\mu}{\tau}$, que es similar a la distribución normal en forma, con colas ligeramente más largas:

$$f(y) = \frac{\exp\left(\frac{x-\mu}{\tau}\right)}{[1+\exp\left(\frac{x-\mu}{\tau}\right)]^2}$$

Con $E(X) = \mu$ y varianza $\sigma^2 = Var(X) = \frac{\pi^2\tau^2}{3}$, tomando entonces $\beta_1 = -\frac{\mu}{\tau}$ y $\beta_2 = \frac{1}{\tau}$:

$$f_X(x; \beta_1, \beta_2) = \frac{\beta_2 e^{\beta_1+\beta_2 x}}{(1+e^{\beta_1+\beta_2 x})^2}$$

Ademas se tiene entonces la función de distribución acumulada:

$$F(y) = \frac{e^y}{1+e^y}$$

Así

$$\pi_i = P(X \leq x_i) = F(x_i) = \frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}}$$
$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$$

- **Función de enlace Complemento Log-Log:** También conocida como la distribución del valor extremo, se tiene la función de densidad dada por:

  $$f(y) = \exp\{y - \exp(y)\}$$

  Donde $-\infty < y < \infty$ por ello, la función de acumulada está dada por:

  $$F(y) = 1 - \exp\{-\exp(y)\}$$

  Análogamente a los dos ejemplos anteriores, se tiene:

  $$\pi_i = P(X \leq x_i) = F(x_i) = 1 - \exp[-\exp(\beta_1 + \beta_2 x_i)]$$

  Entonces para el modelo binomial la función de enlace Complemento Log-Log está definido como:

  $$\log[-\log(1-\pi_i)] = \beta_1 + \beta_2 x_i \tag{4.7}$$

Tabla 4.3: Funciones de enlace de la Binomial

Distribución	Función de enlace
Logit	$\eta = \log\left(\frac{\mu}{1-\mu}\right)$
Probit	$\eta = \Phi^{-1}(\mu)$
Complemento Log-Log	$\eta = \log(-\log(1-\mu))$

Como consecuencia de estas 3 funciones de enlace especiales, se tiene que:

1. La distribución binomial de Y_i, pertenece a la familia exponencial, con $\mu_i = m_i\pi_i$
2. Las variables explicativas vienen en forma de una suma lineal de sus efectos sistemáticos:

   $$\eta_i = \sum_{j=1}^{2} x_{ij}\beta_j = \boldsymbol{x_i^T\beta} \tag{4.8}$$

   Tomando $\boldsymbol{x_i^T} = (1, x_i)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ y η_i predictor lineal.

4.2. Función Desvío o Deviance

Se trata de adaptar un modelo a un conjunto de observaciones con el objetivo de sustituir por un conjunto de valores estimados $\hat{\mu}$ para un modelo con un cierto número de parámetros. Los $\mu's$ no serán exactamente iguales a los $y's$ siendo que la idea principal es saber cuanto difieren. Se debe observar que una pequeña diferencia puede ser torelable, mientras que una gran diferencia no lo puede ser.

Para discriminar la diferencia entre modelos, se deben introducir medidas de discrepaciancias para medir el ajuste del modelo. Nelder & Wedderburn (1972) propusieron como medida de discrepancia **Función desvío o deviance** (Cordeiro G. & Demétrio C. (2008); Paula GA. (2013); McCullagh, P. & Nelder J. (1989)), y está dada por:

$$S_p = 2(\hat{L}_n - \hat{L}_p) \tag{4.9}$$

con L_n y L_p son los máximos del logaritmo de la función de verosimilitud para los modelos saturado y actual, respectivamente. Se puede ver que el modelo saturado se utiliza como base para medir el ajuste del modelo en investigación. Las funciones son:

$$\hat{L}_n = \sum_{i=1}^{n} \phi\{y_i\overline{\theta_i} - b(\overline{\theta_i})\} + \sum_{i=1}^{n} c(y_i, \phi)$$
$$\hat{L}_p = \sum_{i=1}^{n} \phi\{y_i\hat{\theta}_i - b(\hat{\theta}_i)\} + \sum_{i=1}^{n} c(y_i, \phi)$$

Donde $\overline{\theta} = q(y_i)$ y $\hat{\theta} = q(\hat{\mu})$ las estimaciones de máxima verosimilitud del parámetro canónico en el modelo saturado y modelo actual, respectivamente.

$$S_p = \phi D_p = 2\phi \sum_{i=1}^{n} [y_i(\overline{\theta_i} - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\overline{\theta_i})] \tag{4.10}$$

en donde S_p y D_p son denominados desvío escalonado y desvío, respectivamente.

Por ejemplo se puede realizar el cálculo de las principales distribuciones analizadas anteriormente:

- **Normal:** Sea $Y_1, \ldots, Y_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$, entonces $\mu_i = X_i^\top \beta$, especificamente:

$$\begin{aligned} S_p &= \frac{1}{\sigma^2} \sum_{i=1}^{n} 2 \left[y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (2y_i^2 - 2\hat{\mu}_i y_i - y_i^2 + \hat{\mu}_i^2) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \end{aligned}$$

- De la misma forma podemos completar los desvíos para cada una de las distribuciones pertenecientes a la familia exponencial:

Tabla 4.4: Deviance para modelos de distribución de la familia exponencial

Modelo	Deviance
Normal	$D_p = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$
Binomial	$D_p = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right]$
Poisson	$D_p = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$
Binomial Negativa	$D_p = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (y_i + k) \log\left(\frac{\hat{\mu}_i + k}{y_i + k}\right) \right]$
Gamma	$D_p = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Normal Inversa	$D_p = 2 \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$

4.3. Forma Matricial

Lo importante de los Modelos Lineales Generalizado es determinar 3 partes del modelo que son: Variable Respuesta, Matriz del modelo y función

de enlace, se considerará el método de máxima verosimilitud (MV) para estimar los parametros $\beta's$ del modelo (Cordeiro G. & Demétrio C. (2008)).

El vector score está formado por las derivadas parciales de primer orden del logaritmo de la función de verosimilitud, esto es:

$$\ell(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi) \tag{4.11}$$

Donde:

$$\begin{aligned}\theta_i &= q(\mu_i)\\ \mu_i &= g(\eta_i)\\ \eta_i &= \sum_{r=1}^{p} x_{ir}\beta_r\end{aligned}$$

Ahora se toma la derivada parcial de primer orden, regla de la cadena y las implicaciones anteriores, se tiene:

$$U_r = \frac{\partial \ell(\beta)}{\partial \beta_r} \tag{4.12}$$

$$= \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} \tag{4.13}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ir} \qquad r = 1, \ldots, p \tag{4.14}$$

Con ello se tiene que para solucionar $U_r = 0$ para $r = 1, \ldots, p$ son ecuaciones no lineales, entonces se debe solucionar por un método numérico, en este caso y el comunmente mas usado es el Método de Newtton-Raphson, basado en la serie de taylor para aproximar la primera derivada parcial del vector score. (Paula GA. 2013).

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}$$

Donde para el caso multivariado,sea $F : \mathbb{R}^n \to \mathbb{R}^n$ es continua y diferenciable y denotamos como $\boldsymbol{J}$ como la Matriz Jacobiana $(\partial F/\partial x_j)$. Dando un vector inicial x_0, el método de Newton para encontrar una solución a $F(x) = 0$ está definida como:

$$J(x_i)(x_{i+1} - x_i) = -F(x_i) \qquad i \geq 0$$

Ahora tratando de conocer la solución a $\boldsymbol{U(\beta) = \partial \ell(\beta)/\partial \beta} = 0$, adaptamos el método de Newton Raphson:

$$\boldsymbol{\beta^{(m+1)}} = \beta^{(m)} + (J^{(m)})^{-1} U^{(m)} \tag{4.15}$$

Ahora se tiene que aclarar que cuando las derivadas parciales mixtas de la matriz Jacobiana se pueden evaluar facilmente, el método de Newton Raphson es bastante rápido, pero a menudo en los MLG no ocurre esto, por ello se realizá el método de Fisher–scoring que consiste en utilizar las esperanzas de las derivadas parciales mixtas de segundo orden $E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_s}\right)$ en vez de usar simplemente la derivada parcial mixta de segundo orden $\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_s}$, a esta matriz se llamará $\boldsymbol{K}$; Entonces:

$$\boldsymbol{\beta^{(m+1)}} = \beta^{(m)} + (K^{(m)})^{-1} U^{(m)} \tag{4.16}$$

Donde:

$$k_{r,t} = -E\left[\frac{\partial^2 \ell(\beta)}{\partial \beta_r \partial \beta_t}\right] = E\left[\frac{\partial \ell(\beta)}{\partial \beta_r}\frac{\partial \ell(\beta)}{\partial \beta_t}\right]$$

Entonces se puede escribir el método de la siguiente manera:

$$K^{(m)}\beta^{(m)} = K^{(m)}\beta^{(m)} + U^{(m)} \tag{4.17}$$

Ahora tomando la ecuación 4.14 se tiene:

$$\begin{aligned} k_{r,t} &= E(U_s U_t) \\ &= \phi^{-2} \sum_{i=1}^{n} E[y_i - \mu_i]^2 \frac{1}{V^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ir} x_{it} \\ &= \phi^{-1} \sum_{i=1}^{n} \frac{1}{V} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ir} x_{it} \\ &= \phi^{-1} \sum_{i=1}^{n} w_i x_{ir} x_{it} \end{aligned}$$

Donde $w_i = \frac{1}{V}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ también llamado el peso. Con ello la matriz de la información de Fisher para β tiene la forma:

$$K = \phi^{-1} X^T W X \tag{4.18}$$

Donde $W = diag\{w_1, w_2, \ldots, w_n\}$ que es una matriz de peso diagonal que contiene información sobre la distribución y la función de enlace utilizada. Así el vector score puede ser escrito

$$U = \frac{1}{\phi} X^T W G(y - \mu) \tag{4.19}$$

Con $G = diag\{d\eta_1/d\mu_i, \ldots, d\eta_n/d\mu_n\} = diag\{g'(\mu_1), \ldots, g'(\mu_n)\}$, G es una matriz diagonal con las derivadas de primer orden de la función de enlace. Sustituyendo K por U en la equación 4.17.

$$\begin{aligned}
X^T W^{(m)} X\beta^{(m+1)} &= X^T W^{(m)} X\beta^{(m)} + X^T W^{(m)} G^{(m)}(y - \mu^{(m)}) \\
X^T W^{(m)} X\beta^{(m+1)} &= X^T W^{(m)} [\underbrace{X\beta^{(m)}}_{\eta^{(m)}} + G^{(m)}(y - \mu^{(m)})] \\
X^T W^{(m)} X\beta^{(m+1)} &= X^T W^{(m)} [\eta^{(m)} + G^{(m)}(y - \mu^{(m)})]
\end{aligned}$$

Por último definimos $z = \eta + G(y - \mu)$

$$\beta^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} z^{(m)} \tag{4.20}$$

Existen varios criterios de convergencia, pero uno de los más usados en este método es:

$$\sum_{i=1}^{p} \left(\frac{\beta_i^{(m+1)} - \beta_i^{(m)}}{\beta_i^{(m)}} \right)^2 < \epsilon \tag{4.21}$$

4.4. Residuos de los Modelos Lineales Generalizados

Los métodos para analizar los residuos de los MLG poseen pequeños cambios con respecto a los modelos de regresión, se entiende como matriz de proyección sobre un subespacio columna X está definido como (Lee Y., Nelder J. & Pawitan Y. 2017):

$$H = X(X^T X)^{-1} X^T$$

Donde cada elemento de la matriz puede definirse como:

$$h_{i,i} = x_i^T (X^T X)^{-1} x_i$$

Con $x_i^T = (x_{i1}, \ldots, x_{ip})$.
Ahora para ajustar mejor la matriz de proyección a MLG se debe cambiar X por $W^{1/2}X$, entonces, se define H como la matriz de proyección para MLG, que depende de las variables explicativas, la función de enlace η y la función de varianza.

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2} \tag{4.22}$$

Los residuos en general son de mucha importancia ya que detectan la presencia de valores atípicos que deben estudiarse (que para el investigador pueden o no ser importantes). El residuo R_i define una distancia entre los valores de observación y ajustado por el modelo:

$$R_i = h_i(y_i, \hat{y}_i)$$

La matriz H es importante para el análisis de residuos de los MLG y tiene las propiedades:

$$tr(H) = p$$
$$0 \leq h_{ii} \leq 1$$

La función h_i debe ser escogida de forma que pueda satisfacer las propiedades:

$$E(R_i) = 0$$
$$V(R_i) = c$$
$$Cov(R_i, R_j) = 0 \qquad i \neq j$$

Donde c es una constante, los residuos de Pearson tienen estructura de covarianza dada aproximadamente por la matriz de proyección:

$$I - H = I - W^{1/2}ZW^{1/2} \qquad \text{Donde } Z = X(X^TWX)^{-1}X^T$$

Donde Z es una matriz de covarianza asintótica de $\hat{\eta}$, donde tomamos la ecuación 4.20 de forma general:

$$\hat{\beta} = (X^T\widehat{W}X)^{-1}X^T\widehat{W}\hat{z} \tag{4.23}$$

Donde $\hat{z} = \hat{\eta} + \widehat{H}(y - \hat{\mu})$. Tomando la matriz Z se tiene:

$$\hat{z} - \hat{\eta} = (I - \widehat{Z}\widehat{W})\hat{z}$$

Suponiendo Z y W son apróximadamente constantes entonces se puede escribir:

$$Cov(\hat{z} - \hat{\eta}) \approx (I - ZW)Cov(\hat{z})(I - ZW)^T$$
$$Cov(\hat{z} - \hat{\eta}) \approx W^{-1/2}(I - H)W^{-1/2} \qquad \text{puesto que } Cov(\hat{Z}) = W^{-1}$$

Así

$$Cov[\widehat{W}^{1/2}(\hat{z} - \hat{\eta})] \approx I - H \tag{4.24}$$

4.5. Matriz de Varianza

Se ha definido el Modelo Lineal Generalizado GLM como:

$$y = \mu + \epsilon \tag{4.25}$$

μ relacionado con una suma lineal de los efectos fijos $X\beta$ (componente lineal) por una función de enlace $g(.)$:

$$g(\mu) = X\beta \tag{4.26}$$

La matriz de varianza para GLM puede ser escrito como:

$$Var(y) = Var(\epsilon) = V \tag{4.27}$$

Como se tiene que GLM es un modelo de efectos fijos, se supone que las observaciones son incorrelacionadas por tanto la matriz V es una matriz con estructura diagonal. Los términos de la matriz son iguales a las variaciones de cada observación dada la distribución. Como se está trabajando en datos con respuesta binaria, entonces una matriz de varianza para n observaciones Bernoulli donde $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ (Brown H. & Prescott R. 2006):

$$V = \begin{pmatrix} \mu_1(1-\mu_1) & 0 & \cdots & 0 & 0 \\ 0 & \mu_2(1-\mu_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mu_{n-1}(1-\mu_{n-1}) & 0 \\ 0 & 0 & \cdots & 0 & \mu_n(1-\mu_n) \end{pmatrix} \tag{4.28}$$

Por tanto la matriz de varianza se puede escribir de forma general:

$$Var(y) = ab''(\theta) = ag'^{-1}(\mu) \tag{4.29}$$

Con ello se tiene que de forma matricial se tiene esta matriz de la forma:

$$V = AB \tag{4.30}$$

Donde

- $A = diag\{a_i\}$.
- $B = dig\{b''(\theta_i)\} = diag\{g'^{-1}(\mu_i)\}$

Para el problema que se desea invstigar y fundamentar, como la variable respuesta es de tipo binario, se utilizarán distribuciones de Bernoulli es decir de respuesta binaria $A = I$, entonces $V = B$. Habiendo fundamentado la teoría de Modelos Lineales Generalizados, en la próxima sección se abordarán los modelos Lineales Mixtos Generalizados, el propósito ahora es incorporar la matriz de parentesco definida en la sección de Modelo Lineales Mixtos como fundamentación teórica para estudiar datos con estructura de familia.

4.6. Modelos Lineales Mixtos Generalizados

Los modelos lineales mixtos generalizados (MLMG) (los cuales constituyen una combinación natural de dos líneas de modelación, los modelos lineales mixtos y los modelos lineales generalizados) se ha convertido en una rama del análisis estadístico que proporciona las herramientas necesarias para dar respuesta a diversos tipos de problemas que envuelven una matriz de correlación. (Brown H. & Prescott R. 2006).

Definición 2. El modelo GLMM puede ser definido por:

$$y = \mu + \epsilon \tag{4.31}$$

Como en GLM, μ es el vector de valores esperados de las observaciones y está vinculado a los parámetros del modelo mediante la función de enlace g^{-1}.

$$g^{-1}(\mu) = X\beta + Z\gamma \tag{4.32}$$

Donde X y Z son matrices de los elementos fijos y aleatorios respectivamente. Además β y γ son los vectores de los parámetros de efectos fijos y aleatorios como en el modelo mixto. Se supone que los efectos aleatorios, γ sigue una distribución normal.

$$\gamma \sim N(0, G)$$

Así la matriz de varianza puede ser definido por:

$$Var(y) = V = Var(\mu) + R \tag{4.33}$$

Donde R es la matriz de la varianza residual. Además tomando $Var(\mu)$ como en GLM se escribe:

$$V \approx BZGZ'B + R \tag{4.34}$$

Donde $B = diag\{b''(\theta_i)\} = diag\{g^{-1}(\mu_i)\}$

Un caso particular para datos binarios se tiene que $B = diag\{\mu_i(1-\mu_i)\}$

4.7. Resumen de los Modelos.

Modelo Lineal Mixto

$$\begin{aligned} y &= X\beta + Z\gamma + \epsilon \\ \gamma &\sim N(0, G) \\ Var(\epsilon) &= R \\ Var(y) &= V = ZGZ' + R \end{aligned}$$

Modelo Lineal Mixto Generalizado

$$\begin{aligned} y &= \mu + \epsilon \\ g(\mu) &= X\beta + Z\gamma \\ \gamma &\sim N(0, G) \\ Var(\epsilon) &= R \\ Var(y) &= V = Var(\mu) + R \approx BZGZ'B + R \end{aligned}$$

Donde:

- y es la variable dependiente.
- ϵ Variable de error.
- X Matriz de elementos fijos.
- Z Matriz de elementos aleatorios.
- β Parámetros de los elementos fijos.
- γ Parámetros de los elementos aleatorios.
- R Matriz de covarianza residual.
- G Matriz de covarianza.
- V Matriz de varianza.
- μ Valores esperados.
- g Función de enlace.

- B Matriz diagonal para datos binarios $B = diag\{\mu_i(1-\mu_i)\}$

Para los datos con estructura familiar, haciendo referencia al modelo poligénico, donde la matriz de varianza V era descrita como:

$$V = 2\Phi\sigma_g^2 + I_n\sigma_e^2$$

donde 2Φ fue definida como la matriz de parentesco. Note que en el Modelo Lineal Mixto Generalizado, la conexión para formalizar la teoría con datos de familia es a través de la matriz $\mathbf{G} = 2\Phi$ cuya descripción fue definida en el Capítulo dedicado a Modelos Mixtos.
Cabe anotar que se intentó acabar de formalizar la teoría para poder llegar al cálculo análitico de herdabilidad en este tipo de estudios de datos de familia y variables respuesta binaria.

4.8. Aplicaciones

En esta aplicación el objetivo es analizar los programas en que es posible realizar aplicaciones de los Modelos Mixtos Generalizados y comparar resultados. En el Capítulo de la Introducción se mencionó el hecho de que el Programa *SOLAR* establece los resultados del ajuste del Modelo Mixto Generalizado para datos de familia. Esos resultados se van a comparar con resultados del Programa R para establecer conclusiones muy importantes que se deberían tener en cuenta a la hora de analizar este tipo de datos.
Específicamente, se comparan las salidas del programa *SOLAR* versus el Programa R a través del paquete *lme4qtl*, cuando se analiza una variable respuesta binaria. Es sumamente relevante considerar la codificación de las variables, ya que, dependiendo de su respectiva codificación las salidas de los dos programas arrojan valores con distinto signo, cambiando totalmente la interpretación. Se enfatizará en la ventajas de usar el programa R, pues hay limitaciones en el uso del programa *SOLAR* (que ha sido ampliamente usado hasta ahora).

Simplemente para ilustrar este hecho, se tomará una base de datos que está en estudio por algunos investigadores. La base consiste de 953 individuos que pertenecen a una población de Indios Xavantes, (Brasil), para los cuales fueron tomadas algunas medidas que están ilustradas en la Tabla (4.5). Se resalta que el sobrepeso y la obesidad son determinantes de alto riesgo para la diabetes, que se está volviendo un problema de salud pública no solamente en adultos sino también en niños y adolescentes.

Para algunas poblaciones se ha descubierto que el FTO es causal de sobrepeso y el objetivo del estudio es evaluar si la variante genética FTO (rs9939609) es asociada con medidas de sobrepeso u obesidad en la población de Indios Xavantes.

Para los análisis, se van a tomar las comparaciones entre las salidas de los software *SOLAR* y *R*, para ello en *R* se va a usar el paquete `lme4qtl`[1]. (Cabe anotar que la última versión de este paquete fue publicado el 19 de Septiembre del 2019 por ello aún no cuenta con la documentación completa en el CRAN de R).

Tabla 4.5: Ilustración del banco de datos de los fenótipos de los Indios Xavantes.

	FID	IID	FA	MO	genero	Edad	sobrepeso	imc	Obesidad	FTO	FTO_Aditivo
1	1	1	9981	9991	1	60	NA	43.7	0	AT	1
2	1	2	118	43	1	43	0	22.5	0	TT	0
3	1	3	9983	9993	2	49	0	24.2	0	TT	0
4	1	4	9984	9994	2	75	0	21.3	0	TT	0
5	1	5	160	9995	2	72	1	25.1	1	TT	0
6	1	6	48	9996	2	47	NA	34.6	0	TT	0
7	1	7	9987	9997	2	68	1	19.5	1	TT	0
8	1	8	9988	9998	2	74	1	25.8	1	TT	0
9	1	9	9989	9999	2	78	NA	32	0	TT	0
10	1	10	99810	99910	2	46	1	29.4	1	TT	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
948	1	1701	1428	1429	2	22	1	25.7	0	TT	0
949	1	1702	1401	1431	1	22	1	28.9	1	TT	0
950	1	1703	9981000	9991000	1	24	1	26.5	1	TT	0
951	1	1704	1428	1429	1	26	1	26.8	1	TT	0
952	1	1705	1420	1417	1	20	0	23.6	0	TT	0
953	1	1706	9981003	9991003	1	21	1	25.6	1	TT	0

Donde en la base de datos se consideraron las siguientes variables:

- **FID**, **IID**,**FA**,**MO** corresponden a la etiquetas de la estructura familiar, específicamente: familia, individuo, padre y madre. Cabe recordar que de esta estructura familiar es obtenida la matriz de parentesco, 2Φ.
- **genero**: Género de cada uno de los individuos (1 para el género masculino y 2 para el género femenino).
- **Edad**: Edad en años de cada uno de los individuos.

[1] https://github.com/variani/lme4qtl

- **sobrepeso**: Variable binaria (0 no posee sobrepeso y 1 posee sobrepeso).
- **imc**: Variable continua que representa el índice de masa corporal.
- **Obesidad**: Variable binaria (0 no posee obesidad y 1 posee obesidad).
- **FTO**: Marcador molecular codificado según el número de alelos de riesgo. A se denota como el alelo de riesgo, esto se representa como 0 para alelos TT, 1 para alelos AT o TA y 2 para alelos AA.

Resultados

Modelo 1 Para el primero modelo se tomó las siguientes variables, como variable respuesta *Sobrepeso*, y como covariables, *sexo*, *edad* y $FTOAditivo$ para verificar si en realidad está afectando el sobrepeso de las personas en estudio, donde tambien el sistema incorpora la matriz de parentesco 2Φ:

$$Sobrepeso \sim Sex + Edad + FTOAditivo \tag{4.35}$$

- *SOLAR*

  Esta es la salida en el programa SOLAR correspondiente al modelo tratado.

  Donde la base de datos (Pedigree) es la correspondiente a la información de cada uno de los individuos como el ID e información de identificacion de los padres, ademas se tiene también la base de datos `Phenotypes` que corresponde a la base de datos de los fenotipos de los individos;

```
Pedigree:    dbsnp.ped
Phenotypes:  xavantes_1.phen
Trait:       sobrepeso             Individuals:  472

H2r is           0.1773953  p = 0.1735676  (Not Significant)
H2r Std. Error:  0.1998086

sex          p = 0.8422097  (Not Significant)
idade        p = 0.0338713  (Significant)
FTO_Aditivo  p = 0.0226122  (Significant)
```

```
The following covariates were removed from final models:
sex

Kullback-Leibler R-squared is 0.0176259

Output files and models are in directory sobrepeso/
Summary results are in sobrepeso/polygenic.out
Loglikelihoods and chi's are in sobrepeso/polygenic.logs.out
Best model is named poly and null0 (currently loaded)
Final models are named poly, spor, nocovar
Constrained covariate models are named no<covariate name>
solar> parameters

mean          =-0.5866527425 Lower -8           Upper 8
mean         se 0.0685969201 score  0

sd             = 1           Lower  0           Upper 2.283815759
e2             = 0.8226047294 Lower  0.03       Upper 1
e2           se 0.1998085533 score  0
h2r            = 0.1773952705 Lower  0           Upper 1
h2r          se 0.1998085533 score  0

bidade         = 0.0055489595 Lower -0.015432099 Upper 0.0154320989
bidade       se 0.0025883150 score  0
bFTO_Aditivo   = 0.4476494385 Lower -1.25        Upper 1.25
bFTO_Aditivo se 0.1956009761 score 0
```

Cada una de las salidas de el solar se interpretan de la siguiente manera:

- Comenzando estos resultados toma los P-valores de la herdabilidad que se representa como `H2r` y el valor del mismo (donde en este caso no es significante). Además de cada p-valor de las variables fijas para mirar si son realmente significantes o no en este modelo. Se puede observar que la variable `sex` no es significativa y las variable `idade` y `FTO_Aditivo` si son significativas.
- mean: corresponde al β_0 del modelo con un valor aproximado de -0.5866527425.
- e2: corresponde a la varianza del componente de error del modelo.

- h2r: Es la herdabilidad del modelo, hay que recordar que e2+H2r=1
- *bidade*, FTO_Aditivo: son los $\beta's$ correspondientes a cada una de las variables fijas (como se observa no presenta el β de la variable `sex` ya que esta variable no es significativa en el modelo.)

- **R**

  Esta es la salida en el programa R (con el paquete *lme4qtl*) correspondiente al modelo tratado, la cual si solo es una sola base de datos, tomamos la función de enlace binomial (probit) y exactamente el mismo modelo que en SOLAR, con variable respuesta `sobrepeso`:

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]
Family: binomial  ( probit )
Formula: sobrepeso ~ sex + idade + FTO_Aditivo + (1 | IID)
   Data: nuevo

     AIC      BIC   logLik deviance df.resid
   573.9    594.7   -281.9    563.9      467

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8530 -1.2768  0.5747  0.6294  1.0711

Random effects:
Groups Name        Variance  Std.Dev.
IID    (Intercept) 1.489e-08 0.000122
Number of obs: 472, groups:  IID, 472

Fixed effects:
               Estimate   Std. Error z value Pr(>|z|)
(Intercept)    0.881381   0.224696   3.923   8.76e-05 ***
sex           -0.030498   0.122632  -0.249   0.8036
idade         -0.005406   0.002629  -2.056   0.0397 *
FTO_Aditivo   -0.446841   0.191406  -2.335   0.0196 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
```

```
              (Intr) sex.x  idad.x
sex.x        -0.787
idade.x      -0.510 -0.043
FTO_Aditv.x -0.084 -0.003 -0.022
```

- **Observaciones**

  - Una observación muy importante en el momento de analizar estos dos software es que cuando se ajustan estos tipos de modelos en SOLAR, si una variable no es significativa el programa no presenta los valores de esta variable, por ejemplo en este modelo no presenta los resultados de la variable "genero" ya que no es significativa para este modelo por su $p-valor$. En nuestros análisis, así la variable no sea estadísticamente significativa, es muy importante tener los valores estimados, intervalos de confianza, entre otros resultados. Afortunadamente en el programa R si se pueden visualizar todas las aproximaciones de los parámetros del modelo.

  - Hay una presencia en cambios de signos de los modelos de SOLAR a R lo cual puede ocasionar una mala interpretación de los resultados. Según estos resultados, las salidas del SOLAR se interpretarían así: **El gen FTO aumenta el riesgo de sobrepeso en la población.**

    Por otro lado, en el mismo modelo, las salidas del R se interpretarían así: **El gen FTO disminuye el riesgo de sobrepeso en la población.**

    A simple vista se obtiene una contradicción, por lo cual es muy importante que en este tipo de análisis de variables respuesta binaria, se tenga en cuenta una referencia a la hora de realizar interpretaciones, de lo contrario, se puede llegar a conclusiones totalmente equivocadas. Por ello se ha realizado un cambio en la variable binaria respuesta, es decir cambiar la relación de 1 tener sobrepeso a interpretarla como 0 tener sobrepeso.

    Una vez realizado este proceso, se ajusta el modelo con la nueva variable, en el programa R, (la variable binaria sobrepesoT como la nueva variable) y se obtienen los siguientes resultados:

    ```
    SALIDAS DE R CON LA VARIABLE BINARIA 0 -- 1 POR 1 -- 0
    ```

```
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial  ( probit )
Formula: sobrepeso.xt ~ sex.x + idade.x + FTO_Aditivo.x + (1 | IID)
   Data: nuevoT

     AIC      BIC   logLik deviance df.resid
   573.9    594.7   -281.9    563.9      467

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0711 -0.6294 -0.5747  1.2768  1.8530

Random effects:
 Groups Name        Variance  Std.Dev.
 IID    (Intercept) 2.111e-09 4.594e-05
Number of obs: 472, groups:  IID, 472

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.881379   0.224692  -3.923   8.76e-05 ***
sex.x          0.030497   0.122632   0.249   0.8036
idade.x        0.005406   0.002629   2.056   0.0397 *
FTO_Aditivo.x  0.446842   0.191405   2.335   0.0196 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) sex.x  idad.x
sex.x       -0.787
idade.x     -0.510 -0.043
FTO_Aditv.x -0.084 -0.003 -0.022
convergence code: 0
boundary (singular) fit: see ?isSingular
```

Se observa que se tienen muy buenas aproximaciones a los mismo resultados que se tenían anteriormente con las salidas de SOLAR, pero con el signo contrario que está acorde con los resultados que da el programa SOLAR.

- También se tiene que las aproximaciones de los $\beta's$ tienen muy buena aproximación:

Tabla 4.6: Comparación entre $\beta's$ SOLAR vs $\beta's$ R, Modelo 1

Variable	$\beta's$ SOLAR	$\beta's$ R
intercepto	-0.5866527425	-0.881379
Sex		0.030497
Edad	0.0055489595	0.005406
FTO Aditivo	0.4476494385	0.446842

- Desafortunadamente para este paquete en R no se ha implementado la herdabilidad para datos binarios (para datos continuos funciona muy bien), esto es algo por mejorar y se puede seguir estudiando y trabajando en ello ya que es un dato muy importante para estudios en genética.

Modelo 2 Como en el modelo anterior para el segundo modelo se ha tomado las siguientes variables:

$$Obesidad \sim Sex + Edad + FTOAditivo \tag{4.36}$$

Se realiza este modelo con el fin de verificar el enunciado principal de virificar que la variable `FTO_Aditivo` si o no influye en el sobrepeso y en la obesidad de las personas.

- **SOLAR**

```
Pedigree:    dbsnp.ped
Phenotypes:  xavantes_1.phen
Trait:       Obesidade             Individuals:  618

H2r is           0.5360691  p = 0.0010443  (Significant)
H2r Std. Error:  0.1876579

sex          p = 0.4581813      (Not Significant)
idade        p = 8.2771726e-11  (Significant)
FTO_Aditivo  p = 0.3306246      (Not Significant)

The following covariates were removed from final models:
```

```
sex
FTO_Aditivo

solar> parameters
mean         =-0.7788478562 Lower -8             Upper 8
mean        se 0.0653442663 score  0
sd           = 1            Lower  0             Upper 2.0977552752
e2           = 0.4639308773 Lower  0.03          Upper 1
e2          se 0.1876578761 score  0
h2r          = 0.5360691227 Lower  0             Upper 1
h2r         se 0.1876578761 score  0
bidade       = 0.0187302603 Lower -0.0154320988  Upper 0.1388888889
bidade      se 0.0028836973 score  0
```

- **R**

```
eneralized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']

 Family: binomial  ( probit )
Formula: Obesidadet ~ sex.x + idade.x + FTO_Aditivo.x + (1 | IID)
   Data: nuevoT

     AIC      BIC   logLik deviance df.resid
   628.4    650.5   -309.2    618.4      609

Scaled residuals:

    Min      1Q  Median      3Q     Max
-1.1851 -0.5270 -0.4512 -0.3600  2.9318

Random effects:

 Groups Name        Variance Std.Dev.
 IID    (Intercept) 0        0
Number of obs: 614, groups:  IID, 614

Fixed effects:

               Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)   -1.429074   0.227641  -6.278 3.44e-10 ***
sex.x         -0.083111   0.115204  -0.721    0.471
idade.x        0.018746   0.003038   6.171 6.78e-10 ***
FTO_Aditivo.x  0.150462   0.163793   0.919    0.358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

            (Intr) sex.x  idad.x
sex.x       -0.778
idade.x     -0.584  0.023
FTO_Aditv.x -0.118  0.020  0.003
convergence code: 0
boundary (singular) fit: see ?isSingular
```

- **Observaciones**
  - Se tiene que para este caso la variable `FTO_Aditivo` no es significativa, lo cual implica que no afecta la variable respuesta que es `obesidad`.
  - Para el Modelo 2 también se realizo el cambio de signo de la variable respuesta para poder obtener igualdad en los resultados arrojados por *SOLAR* y *R*.
  - En este modelo también se presenta una muy buena aproximación de los $\beta's$, como se puede observar:

Variable	$\beta's$ SOLAR	$\beta's$ R
intercepto	-0.7788478562	-1.429074
Sex		-0.083111
Edad	0.0187302603	0.018746
FTO Aditivo		0.150462

Tabla 4.7: Comparación entre $\beta's$ SOLAR vs $\beta's$ R, Modelo 2

De este ejercicio, podemos concluir que los dos programas se pueden utilizar para obtener resultados de análisis de Modelos Mixtos Generalizados, pero con mucho cuidado ya que si no se tiene en cuenta la parametrización se puede llegar a errores que repercutirán notablemente en una investigación.

Capítulo 5

Conclusiones

- Como aplicación del Modelo Lineal Mixto, los resultados hasta ahora obtenidos de la variable $PASmedia$ son muy satisfactorios y relevantes para posteriores investigaciones biológicas.
- Las aplicaciones del Modelo Lineal Mixto fueron realizadas con paquetes diferentes al lme4qtl debido a que cuando se implementaron no se conocía aún este paquete. Para los próximos análisis se recomienda su uso, dado que presenta un bajo costo computacional y mayor rapidez en obtener los resultados. Observase que con este paquete se pueden realizar análisis para variables respuesta continuas y binarias, ya considerando los Modelos Lineales Generalizados.
- Muchos investigadores hasta ahora han estado usando el programa $SOLAR$ con sus diferentes limitaciones (por ejemplo obtener los $\beta's$ de las variables no significativas o intervalos de confianza de estos $\beta's$), por ello se ha creado la necesidad de estudiar en el programa R, diferentes paquetes e implementarlos para los diferentes análisis.
- En el caso de estudios con datos de familia en Modelos Lineales Mixtos Generalizados no hay estudios específicos para el cálculo del coeficiente de herdabilidad que es relevante en estudios genéticos con datos de familia, ya que nos indica la proporción de la variable respuesta que es debida a componentes genéticos, para decidir si es factible incluir variables genéticas en el modelo bajo análisis o no. Se esperan más análisis en esta dirección.

Apéndice A

Paquetes en R

Hasta el momento se han encontrado varios paquetes en R para facilitar el trabajo de los Modelos Lineales Mixtos Generalizados como:

- **lme4qtl**: Este paquete es el que se ha desarrollado a lo largo de este trabajo, es una extension del paquete `lme4`; Este paquete se ha desarrollado para utilizar la estructura de covarianza de los efectos aleatorios de modelos de respuesta binaria. Como visto, este paquete incorpora la matriz de parentesco 2Φ, por tal razón es importante para el estudio y análisis de datos con estructura familiar.

  Este paquete tiene las iniciales de `lme4qtl`= Linear Mixed Effects Models for Quantitative Trait Loci Mapping, por esto una de las bases de programación de este paquete son los modelos poligénicos para respuesta continua (muy buen desarrollo), para respuesta binaria (falta aun la implementación de los analisis de Coeficientes de Correlación Intraclase).

  Ziyatdinov A. et al. (2018) presenta una muy buena diferencia en la eficiencia de SOLAR vs lme4qtl de su rapidez, para una base de datos de 934 individuos (Base de datos de la viñeta GAIT2):

  Aun la información en el CRAN de R está pendiente por su publicación, pero se puede descargar intalar de la siguiente manera:

```
# install.packages("devtools")
devtools::install_github("variani/lme4qtl")
```

Tabla A.1: Comparación de rapidez **SOLAR** vs lme4qtl, Ziyatdinov A. et al. (2018)

Modelo	SOLAR (días)	lme4qtl (días)
$APTT \sim 1 + (1\|ID)$	1.2	1.6
$APTT \sim AGE + SEX + (1\|ID)$	1.6	1.6
$APTT \sim 1 + (1\|HHID) + (1\|ID)$	5.6	1.7
$APTT \sim AGE + SEX + (1\|HHID) + (1\|ID)$	8.2	1.7

- **Genetics:** Este paquete incluye clases y métodos para el manejo de datos genéticos donde los datos de fenótipos tienen estructura de familia..

- **Solarius:** Por ultimo se este paquete es una interfaz entre SOLAR y R, lo cual se hace que sea un poco más de dificil acceso ya que SOLAR solo está presente para el sistema operativo UBUNTU o MAC, lo cual hace que en un sistema operativo Windows se tenga que manejar una máquina virtual lo cual hace que el proceso requiera mas costo computacional además de que tiene los mismos resultados de SOLAR y su desconocimiento en las variables que no son significativas para el modelo que esté analizando.

  Por el momento el CRAN de R no cuenta con la informacion ya que en Marzo del 2020 se ha eliminado para correcciones del paquete.

Apéndice B

Rutina R para Modelo Lineal Mixto

Para la aplicacion del Modelo Lineal Mixto el proyecto Corazones de Baepend del laboratorio de Genética y Cardiología Molecular (Incor USP) Oliveira et al. (2008), Giolo et al. (2009). Se realizó la siguiente rutina en R para la aplicación de los Modelos Lineales Mixtos.

Para comenzar se deben instalar los siguientes paquetes, ya que se toman funciones de ellas para desarrollar la rutina:

- coxme
- kinship2
- mvtnorm

```
#-----------------------------------------------------------------
# Inicio del script, ingreso y manejo de las bases de datos
#-----------------------------------------------------------------
for(k in 1:XXX){
# XXX Incluir el número total de los marcadores
# que se tiene en el cromosoma a estudiar

setwd("#Direción donde se tienen la informacion de los Cromosomas")
cromo<-paste("./chrXX.", k, sep="")
#XX indica el Cromosoma que está en estudio

# Se carga la base de datos de los insividuos
```

```
imputados<-read.table(cromo, h=F,sep=";")
nomesnps<-imputados[,2]

#Se verifica las dimensiones de la base de datos
r<-nrow(imputados)
c<-ncol(imputados)

# Se toman los genotipos de la base de datos de los individuos
genotipos<-imputados[,6:c]
genotipos<-as.data.frame(genotipos)
genotipos1<-t(genotipos)

#Se toma los nombres de cada uno de los SNPs, para guardarlos en un vector
# que se va a llamar nomesnps
colnames(genotipos1)<-nomesnps
N<-ncol(genotipos1)
banco<-genotipos1

#Dimensiones de la base de datos nueva (banco)
m1<-ncol(banco)
n1<-nrow(banco)
verif<-matrix(NA, ncol=1, nrow=m1)
dim(verif)

#Se comienza las iteraciones desde 1 hasta
#el numero de columnas de la base de datos banco
for(i in 1:m1){
banco[,i]<-as.numeric(as.vector(banco[,i]))
a1<-which(banco[,i] > 0)
na1<-NROW(a1)
verif[i]<-na1 }
a2<-which(verif > 10)
base1<-banco[,a2]

Indiv_chr = read.table('#base de datos con estructura familia, head=T)
#en este campo se puede tomar una base de datos que tenga estructura
#de familia como lo hemos definido en este trabajo, para este trabajo hemos
#tomado la base de datos de Corazones de Baependi

Genotipo = cbind(IID=indiv_chr, base1) # Crear el archivo con IID + snps
```

```
dim(Genotipo) #dimensión de la base Genotipo

#Ahora se incluye el la base de datos de los Fenotipos
fenotipo<-read.table("#Fenotipos", head=T,sep=",")
#Aqui se debe tomar los Fenotipos de la base de datos de estructura familiar
colfenotipo<-ncol(fenotipo)

#Se unifica la base de datos de Fenotipo con Genotipos bajo la variable IID
Dados<-merge(fenotipo,Genotipo, by= "IID",all.x=T)

#Se verifica el numero de filas y columnas de la base Dados
linhas = nrow(Dados)
colunas = ncol(Dados)
snps<-Dados[,(colfenotipo+1):ncol(Dados)]
col.gen<-ncol(snps)
gens = colnames(snps)
attach(Dados, warn.conflicts=F)
IID<-Dados$IID
FID<-Dados$FID

#Ahora se forma la matriz de parentesco
kmat<-makekinship(FID,IID,PAT,MAT)
ncov<-7
ajuste <-function(i){
require(coxme)
require(kinship2)
require(mvtnorm)

snps[,i]<-as.numeric(as.vector(snps[,i]))
snps[,i]

#---------------------------------------
# Realización del modelo
#---------------------------------------

#Se toman el modelo de estudio, tomando las variables SEX
# IDADE_QUES, IMC, V1, V2, V3, V4 y snps correspondiente
#Posterior a ellos se hace una verificacion.
res <- try(lmekin(PAS_MEDIA~ SEX+IDADE_QUES+IMC+V1+V2+V3+V4+snps[ , i],
data=Dados, random=~1|IID, varlist=list(2*kmat)))
```

```
if (class(res)!="try-error"){
ajuste = lmekin(PAS:MEDIA~ SEX+IDADE_QUES+IMC+V1+V2+V3+V4+snps[ , i],
 data=Dados, random=~1|IID, varlist=list(2*kmat))
source('./funcionJASON-deisy.R')
# Función para tomar los valores de ajuste del modelo

#----------------------------------------
# Cálculo del valor de herdabilidad
#----------------------------------------
#Se toman los valores del modelo antes realizado para guardalos en un
#vector de respuestas

#Varianza del error al cuadrado que es la componente
#que se necesita para la herdabilidad sigma_e^2
var.erro <- ajuste$sigma^2

#Varianca Genetica sigma_g^2 explicados en la teoria de modelos Mixtos
var.pol <- as.numeric(ajuste$vcoef)

#Ahora se calcula la herdabilidad del modelo
h2g.mod <- var.pol/(var.pol+var.erro)
h2g.mod

#Se extraen los elementos de los modelos
mod = coxme.extract(ajuste)
mod[[1]]
mod[[2]]
possnp<-ncov+2
beta<-mod[[1]][possnp,1]
desvio<-mod[[1]][possnp,2]
p<-mod[[1]][possnp,4]
num<-ajuste$n

#----------------------------------------
# Creación de tabla deresultados
#----------------------------------------
#Se crea la base de datos con los resultados del modelo
Resultado = cbind(gens[i],k, p, beta, desvio, var.pol, var.erro,
 h2g.mod,num)
```

```
return(Resultado) }

#De lo contrario incluimos NA en los resultados
else{
var.erro <- c("NA")
var.pol <- c("NA")
h2g.mod <- c("NA")
beta<-c("NA")
desvio<-c("NA")
p<-c("NA")
num<-c("NA")
Resultado = cbind(gens[i],k, p, beta, desvio, var.pol,
var.erro, h2g.mod,num)
return(Resultado) } }

#se configuran los nucleos de la maquina para tener un mejor rendimiento
np=1
tudo <-ls()
require(snow)
cl=makeCluster(np, type="SOCK", outfile = "arquivolog.txt")
clusterExport(cl,tudo)
saida<-clusterApplyLB(cl,1:col.gen,ajuste)
stopCluster(cl)
#saida
saidao <- saida[[1]]
for (i in 1 : length(saida)){
saidao <- rbind(saidao, saida[[i]]) }
Resultado <- saidao
row.names(saidao) <- saidao[,1]

#Se imprimen los resultados de ese Cromosoma en análisis
Result=paste("./ResultadocrXX/RcromXX",sep="") #XX numero del cromosoma
Result=paste(Result,".",sep="")
Result=paste(Result,k,sep="")
write.table(Resultado,file=Result,col.names=TRUE,row.names=FALSE, sep=" ")

#Se muestra el valor k del marcador en cuestion
print(k)

#Se limpia la consola para no albergar tantos datos.
```

```
rm(list=ls())}
```

Bibliografía

[1] **Almasy L. & Blangero J. (1997)**, Multipoint oligogenic linkage analysis of quantitative traits. Genet. Epidemiol., 14: 959-964. doi:10.1002/(SICI)1098-2272(1997)14:6¡959::AID-GEPI66¿3.0.CO;2-K

[2] **Amos, C. I. (1994).** Robust Variance-Components Approach for Assessing Genetic Linkage in Pedigrees. Am. J. Hum. Genet 54(3), 535-543.

[3] **Amos C. & Elston, R. (1989)** Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol, 6:349 - 360.

[4] **Barsh G., Copenhaver G., Gibson G., Williams S.(2012)** Guidelines for genome-wide association studies. PLoS Genet. 2012;8(7):e1002812. doi:10.1371/journal.pgen.1002812

[5] **Bateson W. (1909)** Medel's Principles of Heredity. Cambridge University Press, Cambridge, UK.

[6] **Blangero J. (2009).** Update to Blangero's "Statistical Genetic Approaches to Human Adaptability" (1993): A Unified Theory of Genotype × Environment Interaction. Human biology. 81. 547-50. 10.3378/027.081.0604.

[7] **Brown H. & Prescott R. (2006)** Applied Mixed Models in Medicine, John Wiley & Son, Ltd. Second ed.

[8] **Cordeiro G. & Demétrio C. (2008)**. Modelos Lineares Generalizados Departamento de Estadística e Informática, UFRPE y Departamento de Ciencias Exactas, ESALQ.

[9] **Correa J. & Salazar J. (2016)** Introducción a los Modelos Mixtos, Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, primera edicion.

[10] **De Andrade M., Amos, C. I. & Thiel, T. J. (1999).** Methods to estimate genetic components of variance for quantitative traits in family studies. Genetic. Epidemiol. 17, 64-76.

[11] **De Villemereuil P. (2012)**. Tutorial Estimation of a biological trait heritability using the animal model How to use the MCMCglmm R package.

[12] **Demidenko E. (2004)**. Mixed Models: Theory and Applications. New York: Wiley, DOI:10.1002/0471728438, ISBN:9780471728436, 9780471601616.

[13] **Díaz E., Bermúdez D. & Pineda W. (2007)**. Estimación de un modelo lineal generalizado mixto para datos de conteo con exceso de ceros., Universidad Santo Tomas, Facultad de Estadística.

[14] **Duarte N., Giolo R., Pereira A., De Andrade M. & Soler J. (2014)** Using the theory of added-variable plot for linear mixed models to descompose genetic effects in family data. Stat. Appl. Genet. Mol. Biol; 13(3): 359-378

[15] **Egan K. et al. (2016)** Timing and quality of sleep in a rural Brazilian family-based cohort, the Baependi Heart Study. Sci Rep 6, 39283. https://doi.org/10.1038/srep39283

[16] **Elston R. & Stewart J. (1971)**. A General Model for the Genetic Analysis of Pedigree Data, Department of Biostatics and the Genetics Curriculum, University of North Carolina, Chapel Hill, N.C., and Department of Genetics, Milton Road Cambridge.

[17] **Falconer & Mackay (1996)**; Introduction to Quantitative Genetics, Cuarta edición, 1996, Pearson Price-Hall p160-170.

[18] **Fisher R. (1918).** The correlation between relatives on the supposition of Mendelian Inheritance. Trans. R. Soc. Edinburg 52, 399-433.

[19] **Fisher R. (1925).** Statistical Methods for Research Workers. 1a ed. Oliver & Boyd, Edinburgh and London.

[20] **Giolo S., Pereira A., De Andrade M., Oliveira C., Krieger J. & Soler J. (2009).** Genetic analysis of age-at-onset for cardiovascular risk factors in a Brazilian family study. Human Heredity. 68(2), 131-138.

[21] **Griffiths A., Miller J., Suzuki D., Lewontin R. & Gelbart W. (2000)** An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Disponible en `https://www.ncbi.nlm.nih.gov/books/NBK21766/`

[22] **Gutierrez J. (2010)** Iniciación a la valoración genética animal. Metodología adaptada al EEES. Madrid, España. UCM Editorial Complutense. 368 p.

[23] **Harlt D. & Clark A. (2010)** Princípios de Genética de Populações, Artmed Editora, 4 ed, ISBN 8536323744, 9788536323749

[24] **Haseman J. & Elston R. (1972)** The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2, 3–19 (1972). https://doi.org/10.1007/BF01066731

[25] **Hazelton M. & Gurrin L. (2003)**, A note on genetic variance components in mixed models. Genet. Epidemiol., 24: 297-301. doi:10.1002/gepi.10242

[26] **Higham N. (2002)** Accurancy and Stability of Numerical Algorithms. Second edition. 460-461.

[27] **Horimoto A., Giolo S., Oliveira C., Alvim R., Soler J., De Andrade M., Krieger J. & Pereira, A. (2011)**. Heritability of physical activity traits in Brazilian families: the Baependi Heart Study. BMC medical genetics, 12, 155. https://doi.org/10.1186/1471-2350-12-155

[28] **Kleber M., Seppala I., et al (2013)** Genome-Wide Association Study Identifies 3 Genomic Loci Significantly Associated With Serium Levels of Homoarginine. The AtheroRemo Consortium, doi: 10.1161/CIRCGENETICS.113.000108

[29] **Kochunov P., Blangero J., et al** Tawes C. Catonsville, MD 21228, SOLAR-Eclipse: An Imaging Genetics Analyses Software `http://www.solar-eclipse-genetics.org/`

[30] **Kraft P. & De Andrade M. (2003)**, Group 6: Pleiotropy and multivariate analysis. Genet. Epidemiol., 25: S50-S56. doi:10.1002/gepi.10284

[31] **Lange K., Westilake J. & Spence M. (1976)**, Extensions to pedigree analysis III. Variance components by the scoring method. Annals of Human Genetics, 39: 485-491. doi:10.1111/j.1469-1809.1976.tb00156.x

[32] **Lee Y., Nelder J. & Pawitan Y. (2017)**; Generalized Linear Models with Random Effects, Chapman & Hall /CRC, Monographs on Statistics and Applied Probability 106.

[33] **McCullagh, P. & Nelder J. (1989)** Generalized linear models. Chapman and Hall, London New York.

[34] **Mousseau T. & Roff D. (1987)**. Mousseau TA, Roff DA. Natural selection and the heritability of fitness components. Heredity 59: 181-197. Heredity. 59 (Pt 2). 181-97. 10.1038/hdy.1987.113.

[35] **Nelder J. & Wedderburn R. (1972)**. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384. doi:10.2307/2344614

[36] **Oliveira M., Pereira A., De Andrade M., Soler J. & Krieger J. (2008).** Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. BMC Medical Genetics. 9:32, 881-888.

[37] **Paula GA. (2013)** Modelos de Rregressao com apoio computacional, Instituto de Matemática e Estatística, Universidade de Sao Paulo.

[38] **Penrose L. (1938).** Genetic linkage in graded human characters. Ann. Eugenics 6, 133-138.

[39] **Schork N. (1993)**, The Design and Use of Variance Component Models in the Analysis of Human Quantitative Pedigree Data. Biom. J., 35: 387-405. doi:10.1002 bimj.4710350402

[40] **Searle S., Casella G. & McCulloch (1992)** Cariance Components, Jhon Wiley & Sins INC., Publications, ISBN 139780470009598

[41] **Self G.& Liang K. (1987)** Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions, Journal of the American Statistical Association, 82:398, 605-610, DOI10.1080/01621459.1987.10478472

[42] **Verbeke G. (2000)** Linear Mixed Models for Longitudinal Data, Springer, New York, NY, 978-0-387-22775-7

[43] **Verbeke G. & Molenberghs G. (1997)** Linear Mixed Models in Practice: A SAS-Oriented Approach. Lecture Notes in Statistics 126. New York; Springer-Verlag.

[44] **Ziyatdinov A., Vásquez S., Brunel H., Martinez A., Aschard H. & Soria J. (2017)**; Supplementary materia for lme4qtl: linear mixed models with flexible covariance structure for genetic of related individuals.

[45] **Ziyatdinov A. et al. (2018)**, lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals, BMC Bioinformatics `https://github.com/variani/lme4qtl`