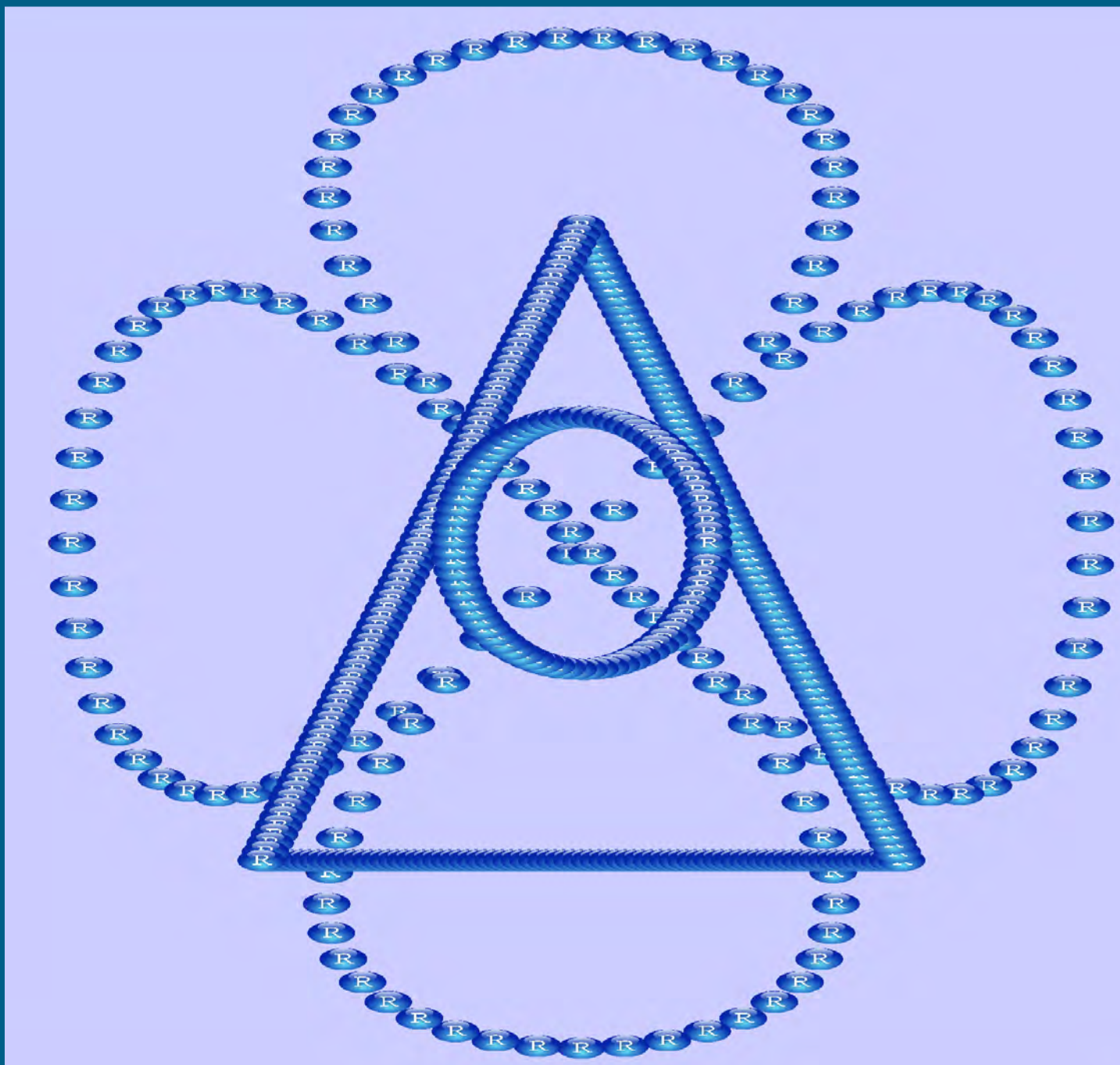

Modelos lineales para evaluación genética en animales

Carlos Eugenio Solarte Portilla
Carlos Alberto Martinez Niño
Mario Fernando Cerón-Muñoz
-Editores-

Universidad de Nariño

UNIVERSIDAD DE ANTIOQUIA
Facultad de Ciencias Agrarias

UNIVERSIDAD NACIONAL DE COLOMBIA

Modelos lineales para evaluación genética en animales

Editores:
Carlos Eugenio Solarte Portilla
Universidad de Nariño

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

Editorial UTP
Colección Trabajos de Investigación
2024

Modelos lineales para evaluación genética en animales / Editado por Carlos Eugenio Solarte Portilla, Carlos Alberto Martínez Niño y Mario Fernando Cerón Muñoz. – Pereira : Editorial Universidad Tecnológica de Pereira, 2024.
477 páginas. -- (Colección Libros resultado de investigación).

e-ISBN: 978-958-722-930-1

1. Genética cuantitativa 2. Mejoramiento Genético Animal 3. Parámetros genéticos 4. Genómica 5. Biotecnológia animal
CDD. 636.089

Modelos lineales para evaluación genética en animales

eISBN: 978-958-722-930-1

Universidad Tecnológica de Pereira
Vicerrectoría de Investigaciones, Innovación y Extensión
Editorial Universidad Tecnológica de Pereira
Pereira, Colombia

Coordinador editorial:
Luis Miguel Vargas Valencia
luismvargas@utp.edu.co
Teléfono (606) 313 7381
Edificio 9, Biblioteca Central Jorge Roa Martínez
Cra. 27 No. 10-02 Los Álamos, Pereira, Colombia
www.utp.edu.co

Montaje y producción
Tomás Flórez Calle
Universidad Tecnológica de Pereira

Pereira, Risaralda, Colombia.

Esta publicación fue generada gracias al desarrollo de las actividades de formación académica en pregrado y posgrado. El contenido fue elaborado en R-project con Sweave y LaTeX.

Contenido

SEGUNDA PARTE

Lista de figuras

Lista de tablas

PRÓLOGO

El Mejoramiento Genético Animal (MGA) es tal vez una de las áreas de mayor dificultad y complejidad a la que se enfrentan los estudiantes, profesores y profesionales de las Ciencias Animales, por la necesidad de combinar y manejar adecuadamente conceptos de genética y matemática, junto con el uso intensivo de programas de ordenador para la construcción, edición y análisis de grandes bases datos.

Adicionalmente, es un hecho evidente que, tanto la mayoría de los libros, como de los artículos científicos necesarios para el estudio, comprensión y aplicación del MGA se publican en lengua inglesa, y si bien en la actualidad el inglés se enseña masivamente en los programas escolares de secundaria y en los programas de pregrado, no es un secreto que en varios países de América Latina, el escaso manejo del idioma inglés sigue siendo un elemento limitante para acceder eficaz y oportunamente al conocimiento.

Algunos libros clásicos que presentan temáticas fundamentales para el MGA han sido traducidos al castellano y también se han publicado algunos textos en español, todos de gran valor y utilidad, puesto que se convierten en una muy buena introducción para estudiar y entender los principios que rigen el MGA. No obstante, estos textos no se enfocan en los modelos estadísticos empleados en evaluación genética animal, un tema de gran relevancia en las ciencias animales.

Cabe destacar, que para muchos estudiantes resulta complejo hablar de matrices, modelos mixtos y modelo animal. Es por ello que, resulta de utilidad disponer de libros que consideren la introducción a los modelos estadísticos más frecuentemente empleados en las evaluaciones genéticas de las poblaciones interés zootécnico, al igual que los fundamentos teóricos de mayor relevancia, tanto en la parte matemática como genética, y que incorpore ejemplos numéricos y uso de los programas computacionales.

Las evaluaciones genéticas constituyen el soporte fundamental e imprescindible para poder llevar a cabo programas MGA, por ese motivo, los estudiantes especialmente de pregrado, han manifestado la necesidad de tener a su alcance un texto especializado que contenga definiciones y conceptos, que faciliten la comprensión y aplicación de los procedimientos requeridos para seleccionar los animales de mayor mérito genético, con el propósito de incrementar el desempeño de las especies de interés zootécnico.

Por otra parte, es necesario mencionar, que entre las comunidades académicas y de criadores de diversas especies animales, existen concepciones y visiones diversas sobre el MGA y sus alcances, por lo que es necesario precisar algunas definiciones básicas e igualmente hacer una breve referencia sobre la evolución, alcances, limitaciones, dificultades y perspectivas del MGA.

Este libro se enfoca en la aplicación de modelos lineales en evaluación genética animal. Se inicia con un recuento del desarrollo de la evaluación genética, con énfasis en tres métodos de predicción que han sido de gran relevancia en el campo: el mejor predictor, el mejor predictor lineal y el mejor predictor lineal insesgado; para luego presentar varios casos del modelo animal que obedecen a diferentes tipos de acción génica, condiciones zootécnicas, estructuras y tipos de datos. Se inicia con el caso más simple del modelo animal, un solo fenotipo, efectos genéticos aditivos directos e individuos de una misma raza. Luego se discuten extensiones que incluyen varios fenotipos, efectos maternos, efectos de ambiente permanente y de camada, modelos multirraciales y selección genómica. Se incluyen ejemplos numéricos implementados en el programa libre R-project.

Dado que se cubren solamente modelos lineales, el texto se enfoca en fenotipos para los cuales estos sean adecuados, por ejemplo, variables continuas o algunas variables discretas transformadas, como se hace con el recuento de células somáticas de la leche. Así, variables medidas en escalas ordinales o nominales no se consideran.

Ahora bien, en nuestra experiencia, una de las mayores dificultades que se encuentran al abordar los modelos de evaluación genética es la falta de formación en teoría básica de la probabilidad, álgebra matricial y teoría de modelos lineales. Además, en el caso particular de la selección genómica, donde se cuenta con una familia de modelos bayesianos de uso frecuente conocida como el *alfabeto bayesiano*, la falta de conocimiento en estadística bayesiana representa otra barrera importante para acceder al conocimiento que se obtiene a partir de la implementación de estos modelos en diferentes especies de interés zootécnico. Por esta razón, buscando generar un texto autocontenido, estos temas se incluyen como capítulos o complementos. En el caso de la teoría de la probabilidad, se presentan fundamentos teóricos y se ilustran con ejemplos en genética. Sumado a esto, en pro de la simplicidad, en algunos casos se sacrifica el rigor matemático y se incluyen definiciones que no son del todo formales o que son consecuencias de la definición formal.

Los autores

DEDICATORIAS

Carlos Solarte dedica este libro a su esposa Margoth y la familia Grande y agradece a la Universidad de Nariño por la comisión de Año Sabático para escribir este libro.

Carlos A. Martínez dedica este libro a la base canónica que genera el espacio vectorial de su vida: su familia; en especial a su compañera de vida Carmen Helena, sus padres Mery y Nazario, sus amigas caninas Ío, Roma y Artemisa, y a la memoria de dos seres muy especiales: su hermano German y la de su amigo Perseo, el mejor perro que ha conocido, un campeón dentro y fuera de las pistas.

La figura que aparece en la carátula hace referencia al símbolo *Gnosa*, elaborada por Mario Cerón-Muñoz (Anexo 16.1).

$$\begin{bmatrix} \textbf{G}\text{nosa libera tu mente,} \\ \textbf{n}\text{aturaleza y espíritu.} \\ \textbf{O}\text{mnipotente siente,} \\ \textbf{s}\text{abiduría de almas,} \\ \textbf{a}\text{mor eterno, } \Sigma o\varphi\iota\alpha \end{bmatrix}$$

PRIMERA PARTE

MEJORAMIENTO GENÉTICO ANIMAL

1 CAPÍTULO UNO

Introducción al mejoramiento genético animal y los modelos de evaluación genética

Carlos Eugenio Solarte Portilla
Universidad de Nariño

Partamos de una de las definiciones más difundida y aceptada sobre mejoramiento genético animal (MGA), expresada por Montaldo y Barria [1] en los siguientes términos: "El Mejoramiento Genético Animal consiste en aplicar principios biológicos, económicos y matemáticos, con el fin de encontrar estrategias óptimas para aprovechar la variación genética existente en una especie de animales en particular para maximizar su mérito. Esto involucra tanto la variación genética entre los individuos de una raza, como la variación entre razas y cruzas".

De la anterior definición conviene destacar que, la complejidad del MGA estriba en la necesidad de conocer y aplicar principios de naturaleza biológica y matemática, con el propósito de realizar evaluaciones objetivas que permitan escoger, con la mayor precisión y confiabilidad posibles, los futuros reproductores, que se espera mejoren una o varias características identificadas como importantes en los objetivos de producción.

No puede perderse de vista el hecho clave relacionado con la variación genética, puesto que esto implica establecer, también con el mayor grado de confianza posible, si las diferencias observadas entre los individuos de una población animal se deben a factores hereditarios que pueden transmitirse de una generación a otra o si por el contrario, la influencia genética es baja. Conocer el tipo de acción génica y cuantificarla, resulta una labor de importancia superlativa cuando se pretende establecer un programa de mejora genética, independientemente de la especie animal con la que se trabaje.

Otro elemento relevante, es la mención a la existencia de variación en una raza o en más de una raza y sus cruces. Esto es importante por cuanto existe una idea bastante generalizada, pero equivocada, que reduce el MGA al simple hecho de cruzar animales de distintas razas.

Finalmente, cuando se menciona que se maximiza el mérito de la especie, se hace referencia al proceso de difusión intensiva del material genético selecto, de modo especial, con el uso de tecnologías de la reproducción asistida, como la inseminación artificial, la multiovulación, el sexaje de semen y embriones, y la transferencia de estos.

Quizá sea necesario agregar que, en la definición de Montaldo y Barria [1],anteriormente mencionada, no se indica expresamente el uso de las herramientas moleculares, probablemente porque esta posibilidad tecnológica se empezó a utilizar con mucha fuerza a finales del siglo XX y principios del XXI, aunque cuando se indica que en el Mejoramiento Genético Animal se combinan principios biológicos, el uso de dichas tecnologías se encontraría implícito en la definición.

Tal vez resulte obvio, pero no puede dejar de mencionarse, que cuando se considere la posibilidad de iniciar un programa de mejora genética, será altamente conveniente, útil, necesario y prácticamente imprescindible, aunar esfuerzos entre las comunidades académicas, las comunidades de productores y las entidades estatales responsables de fortalecer el sector pecuario, con el propósito de canalizar de la mejor manera posible las potencialidades y fortalezas en cada una de ellas. Con ese trabajo articulado se facilita de gran manera la definición de los objetivos del programa; se vuelve mucho más eficiente la recolección, análisis y sistematización de la información de campo, y por supuesto, se favorece el estricto control de la producción, lo que conduce a la realización de evaluaciones genéticas más precisas y confiables y por ende, a una mejor identificación de los animales genéticamente superiores para su posterior uso intensivo.

Adicionalmente y no menos importante, se debe considerar que, en dependencia de los ciclos productivos y vitales de cada especie, los resultados esperados por efecto de la selección de animales genéticamente superiores son observables en el mediano y largo plazo, incluso en especies de corto intervalo generacional. Cuanto más largo es dicho intervalo, mayor tiempo se requerirá para observar y cuantificar los impactos.

En lo referente a la forma en que ha evolucionado el MGA, San Primitivo [2] indicó que el desarrollo de esta área del conocimiento, a lo largo de la segunda mitad del siglo XX, se ha basado en los avances conceptuales de la genética cuantitativa, las posibilidades de aplicación de las metodologías estadísticas e informáticas; los sistemas cada vez mejores para identificar individualmente a los animales, los controles genealógicos y el uso de marcadores moleculares. Todos estos factores han contribuido a incrementar la precisión de la estimación de los valores genéticos y por ende aumentar la respuesta a la selección. Sin embargo, aunque es indiscutible el desarrollo conceptual y metodológico del MGA, persisten serias limitaciones para aplicarlo rutinariamente, en una gran cantidad de especies animales y en varios países.

También se puede asegurar que, tanto la evolución teórica de los métodos y procedimientos, como los mayores avances en los sistemas de producción animal por efecto del MGA, son mucho más evidentes e importantes en algunos países, especialmente de Europa y Norteamérica. En los países en vía de desarrollo, una de las mayores limitantes sigue siendo la enorme dificultad para articular los esfuerzos entre el Estado, la industria y la academia, además de la pretensión equivocada que busca obtener grandes resultados, en poco tiempo y con baja inversión; aunque en varios países de Latinoamérica se evidencian claros avances, especialmente por la formación de profesionales al más alto nivel, y por el interés y compromiso cada vez mayor por parte de los criadores para vincularse con universidades y centros de investigación para estructurar programas de mejora genética, en diversas especies animales.

Para la mayoría de países latinoamericanos, el reto debe ser la continuidad de los procesos que se han iniciado en algunas regiones y con especies de interés local, regional o nacional, insistiendo en la obligatoriedad de buscar mejores mecanismos de articulación entre todos los actores que tienen influencia directa en los procesos de producción animal. Solo así será posible hacer un uso óptimo de las ventajas que ofrece el MGA, para beneficiar a grandes, pequeños y medianos productores.

Complementariamente a lo antes expresado, es conveniente mencionar otros puntos de interés e importancia para diseñar y ejecutar evaluaciones genéticas como soporte a los programas de mejora animal.

Una vez identificados los animales de mayor mérito, a través de los procesos de evaluación genética, es imprescindible programar los apareamientos y la difusión intensiva del material genético seleccionado. En la programación de apareamientos se debe mantener la consanguinidad por debajo de un umbral cuidadosamente establecido. Una herramienta muy útil para lograr este propósito, es el concepto de óptima contribución se basa en implementar métodos de optimización con restricciones para determinar el aporte de cada individuo a la siguiente generación buscando maximizar el progreso genético y controlar el nivel de consanguinidad.

Como puede verse, son varios los desafíos que deben asumirse cuando se diseña un programa de mejoramiento genético, en el cual se parte de la definición de unos objetivos de selección, en dependencia de las particularidades y necesidades de cada entorno y sistema de producción. Realizada dicha definición, resulta de crucial

importancia estudiar e identificar el tipo de herencia de los rasgos a mejorar, determinar la variabilidad genética existente en la población objeto de selección y calcular los parámetros genéticos de interés, los cuales permiten definir si es más conveniente la selección dentro de razas o el cruzamiento, para luego continuar con las evaluaciones genéticas que permiten establecer el valor genético de todos los animales, incluidos en las bases de datos de producción y de genealogías, y seleccionar aquellos que se utilizarán como reproductores de la siguiente generación. Una vez llevado a cabo este proceso, se procederá a establecer los mecanismos para difundir ese material genético selecto y este ciclo se repetirá en cada generación seleccionada, para finalmente,

información fenotípica convencional, ya que los marcadores de ADN se convierten, hasta ahora, en un muy buen complemento para a la información fenotípica y genealógica.

De lo indicado en los párrafos anteriores, queda claro que existe una enorme brecha en el desarrollo de los programas de MGA entre los países desarrollados y los países en vía de desarrollo, y para cerrarla, los programas de mejora genética deben estructurarse con el fin de contribuir a cubrir las necesidades particulares de sus sistemas de producción, para no continuar con la dependencia tecnológica en cuanto al uso de material genético evaluado en condiciones ambientales totalmente diferentes, aunque no hay problema en utilizar individuos foráneos, siempre y cuando hayan sido evaluados a través de progenies que se desempeñen bajo las condiciones en las que se está desarrollando el programa de mejoramiento genético.

En el ámbito global, se prevé que en los próximos años se continuará con un crecimiento mucho más notable de los programas de MGA, debido fundamentalmente a los avances teóricos para utilizar información molecular y aquella proveniente de otros tipos de datos *ómicos*, junto con los modelos convencionales desarrollados inicialmente por el profesor Charles Roy Henderson, los cuales se han aplicado, gracias a los desarrollos que permitieron aumentar la capacidad de cómputo y sin los cuales hubiese sido imposible alcanzar los logros que hoy se aprecian en todo el mundo.

También es necesario mencionar que, si bien el mejoramiento genético, a lo largo de la historia, ha procurado alcanzar mayores niveles de producción en todas las especies animales donde ha intervenido, a futuro tiene una enorme responsabilidad con la preservación del medio ambiente y el equilibrio de los ecosistemas, de tal manera que, entre los objetivos de mejora, debe ser una obligación considerar esos aspectos y no limitarse a producir más, en menor tiempo y a menor costo.

Para finalizar, debe destacarse que los extraordinarios avances, en cuanto al conocimiento en detalle de los genomas, será determinante para tener impactos aún mayores por efecto del mejoramiento genético, ya que actualmente se puede identificar tempranamente, incluso en estado embrionario, los animales portadores de alelos que transmiten caracteres deseables, de un modo rápido, seguro y confiable. Igualmente, se espera que la información molecular permita entender con mucha claridad fenómenos genéticos, como por ejemplo, la interacción entre el genotipo-ambiente y la heterosis, y con ello, desarrollar mejores procedimientos de evaluación genética.

2 CAPÍTULO DOS

Historia de tres desarrollos estadísticos relevantes en el mejoramiento genético

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá.

En este capítulo se discuten tres tipos de predictores de gran importancia en la historia del mejoramiento genético. Antes de hacer referencia expresa sobre estos predictores, es necesario responder una pregunta, que al ser resuelta evita incurrir en errores y confusiones. Dicha pregunta es la siguiente: ¿El valor genético se estima o se predice?; en estadística, generalmente se habla de predecir efectos aleatorios y estimar efectos fijos. Cuando se dice que un efecto es aleatorio, este se trata como una variable aleatoria y por consiguiente, se le asigna una distribución de probabilidad.

El reto aquí es que se está trabajando con variables aleatorias no observables, pues no se conoce el verdadero valor genético de un animal y no hay forma de medirlo directamente; por otro lado, variables como el peso de un vacuno a los 24 meses, son observables, pues podemos medirlas, y una vez se obtiene el registro (por ejemplo 395.5 kg), se dice que este es una realización o valor realizado de la variable aleatoria. Si bien los valores genéticos no son observables, el problema de inferirlos puede verse como la estimación del valor realizado de una variable aleatoria, es decir, estimar cuál fue el valor que la naturaleza eligió para dicha variable. Por ejemplo, supongamos que los valores genéticos de una población de bovinos para el peso al destete son variables aleatorias que siguen una distribución normal con media cero y varianza cuatro, entonces, a partir de dicha distribución, la naturaleza muestrea los valores genéticos de cada uno de los individuos de la población. Pese a no ser observables se emplean los datos disponibles, como por ejemplo los fenotipos, para estimar estos valores realizados. Por otro lado, cuando se considera un escenario futuro, porque la realización de la variable aleatoria no ha ocurrido, se tiene un problema de predicción.

En virtud de las ideas antes expuestas, Henderson [3] comentó que el problema estadístico asociado a la evaluación genética puede verse como predicción o como estimación. Por ejemplo, si se tiene interés en el valor genético de la progenie que podría resultar del apareamiento de un macho y una hembra dados, se tiene un problema de predicción pues el animal no ha nacido. Por otra parte, si tenemos un individuo ya existente, caso en el que su valor genético ya fue determinado, pero es desconocido, se tiene un problema de estimación. Así, en la literatura se encuentran los dos términos, valor genético predicho y valor genético estimado.

Los métodos estadísticos empleados para encontrar el mérito genético de los individuos de una población para uno o más fenotipos de interés experimentaron una gran evolución desde mediados del siglo pasado [4]. Estos han venido cambiando en función de los tipos de registros que se emplean, las condiciones zootécnicas, la estructura de los datos, la naturaleza de la variable respuesta y de la necesidad de encontrar predicciones con propiedades estadísticas y biológicas deseables. A continuación, se describen tres tipos de predictores, donde el tercero es el más usado en la actualidad.

2.1. Mejor predictor (MP)

Consideremos los registros de un fenotipo de interés observados en n individuos y arreglados en un vector columna denotado como y. Estos registros son tratados como variables aleatorias observables, es decir, que existe una forma de medirlas directamente, por consiguiente, se dice que y es un vector aleatorio observable n-dimensional. La idea es emplear estos datos para predecir los valores genéticos, estos corresponden a variables aleatorias no observables y se arreglan en un vector denotado como u de dimensión px1. La idea es predecir u empleando una función de los registros $h(y)$, esto es, $\widehat{u} = h(y)$. En este caso, el principio que se emplea para encontrar dicha función es minimizar el error cuadrático medio de predicción (ECMP), de allí el término "mejor", el cual se define como:

$$E\left[(\widehat{u} - u)^T(\widehat{u} - u)\right]$$

Después de un proceso matemático, en el que se minimiza esta función con respecto a $\widehat{u}$, se encuentra que el mejor predictor es la esperanza condicional de u dado y, como lo indicó Hénderson [5] explícitamente:

$$\widehat{u} = MP(u) = E\left[u \mid y\right] = \int_{\omega} u f(u \mid y) du$$

Donde ω es el conjunto soporte de la distribución de u dado y.

Por construcción, el MP minimiza el error cuadrático medio; además, es insesgado y minimiza la varianza del predictor $Var\left[\widehat{u}\right]$ [6]. También minimiza la varianza del error de predicción, es decir, $Var\left[\widehat{u}-u\right]$ y maximiza la correlación entre el predictor $\widehat{u}$ y el predictando u, la cual se emplea para definir la confiabilidad de la predicción del valor genético. Sin embargo, este predictor presenta una limitante importante en la práctica, puesto que se requiere conocer la distribución de u dado y, razón que limita su implementación. Comúnmente se asume distribución normal multivariada de u y y, lo cual implica que la distribución de u dado y también es normal multivariada; y bajo este escenario:

$$MP(u) = \mu_u + C^T V^{-1}(y - \mu_y)$$

Donde μ_u y μ_y son los vectores de medias de u y y, $V = Var\left[y\right]$ y C es la matriz de covarianza $Cov\left[u, y\right]$. Sin embargo, no existe garantía de que esa sea la verdadera distribución y se deben conocer los vectores de medias y las matrices V y C.

2.2. Mejor predictor lineal (MPL)

El segundo tipo de predictor que estudiaremos en esta sección pertenece a la familia de predictores lineales, esto es, aquellos en los cuales la función $h(y)$ es lineal en los datos, explícitamente, el MPL tiene la forma $\widehat{u} = a + By$. El supuesto de distribución sobre y y u es:

$$\begin{bmatrix} u \\ y \end{bmatrix} \sim \left[\begin{bmatrix} \mu_u \\ \mu_y \end{bmatrix}, \begin{bmatrix} V & C \\ C^T & G \end{bmatrix}\right]$$

Donde a es un vector de dimensión px1 y B es una matriz de dimensión pxn. En este problema, el vector a y la matriz B tienen que elegirse de manera tal que se minimice el ECMP, que aquí tiene la forma:

$$E\left[(\widehat{u}-u)^T(\widehat{u}-u)\right] = E\left[(a+By-u)^T(a+By-u)\right]$$

Tras manipular esta expresión y al usar resultados del álgebra matricial se obtiene que, el MPL de u es $\widehat{u} = \mu_u + C^T V^{-1}(y-\mu_y)$. Vale la pena destacar que, bajo el supuesto de normalidad multivariada, el MP y el MPL son iguales. Además, el MPL también es insesgado y en la clase de los predictores lineales, este maximiza la correlación entre el predictor y el predictando. Otra ventaja del MPL es que no se necesita conocer la distribución de u dado y, esto debido a la restricción de linealidad; sin embargo, el precio que se paga por ello es que, en general, la varianza del error de predicción del MPL es mayor a la del MP [5], excepto en el caso de normalidad en el que son iguales. No obstante, para computar el MPL se precisa conocer los verdaderos vectores de medias y matrices de covarianza de y y u, cosa que no sucede en la práctica, siendo esta la principal limitante del MPL.

2.3. Mejor Predictor Lineal Insesgado (MPLI)

En mejoramiento genético animal y en otros campos, es común asumir que el valor esperado de un registro es una función lineal de un conjunto de parámetros, esto es:

$$E\left[y\right] = X\beta$$

Recordemos que, en el ámbito de la evaluación genética, el vector u contiene los valores genéticos de los individuos y posiblemente otros efectos aleatorios no genéticos (como se verá más adelante). Por ejemplo, el modelo más simple es aquel en el que los únicos efectos aleatorios que se consideran son los efectos genéticos aditivos directos.

Bajo el modelo que se acaba de presentar, para conocer el vector de medias verdaderas de los registros, se debe conocer el vector β, cosa que, como se mencionó antes, no ocurre en la práctica. Así, en este contexto, el MPL requeriría que se conozcan con total certeza estos parámetros y aquellos de los que dependen las matrices de covarianza V y G. Ahora bien, en la clase de los predictores lineales, el MPLI resuelve el problema de conocer β con total certeza. Al ser un predictor lineal, su forma sigue siendo $a + By$, pero en esta ocasión el vector a y la matriz B no solo deben satisfacer el criterio de mínimo ECMP, sino también la propiedad de insesgamiento, esto es:

$$E\left[\widehat{u}\right] = E\left[u\right]$$

Es decir, la esperanza del predictor es igual a la esperanza del predictando. Empleando una técnica proveniente del cálculo multivariado conocida como los multiplicadores de Lagrange se minimiza el ECMP, sujeto a las restricciones inducidas por la condición de insesgamiento, resultando en el MPLI de u:

$$\widehat{u} = C^T V^{-1}(y - X\beta^0)$$

Donde β^0 es una solución de las ecuaciones normales correspondientes a mínimos cuadrados generalizados. En esta expresión vale la pena resaltar lo siguiente:

1: El término $y - X\beta^0$ indica que se emplean registros corregidos por aquellos efectos asociados al vector β, empleando el estimador β^0.

2: Se requiere la inversa de la matriz de covarianzas de los fenotipos (V^{-1}), que corresponde a la matriz de precisión de los mismos.

3: Si no existe covarianza entre los registros fenotípicos y los valores genéticos aditivos, es decir $C = 0$, el MPLI de u es el vector nulo.

Otro aspecto de gran relevancia que se debe resaltar es que, el MPLI no requiere que se tenga una distribución normal multivariada de y y u, este se deriva sin acudir

a dicha suposición, aunque bajo normalidad multivariada el MPLI presenta algunas propiedades extras (Henderson [5]).

Ahora bien, una de las formas más frecuentes de modelar un fenotipo continuo en función de efectos genéticos y ambientales es mediante un modelo lineal mixto (estudiado en la sección 3.5), que en el contexto del problema de predicción de valores genéticos es como sigue:

$$y = X\beta + Zu + e,$$

Donde y y u ya se definieron antes; β es un vector de efectos fijos; e es un vector de errores aleatorios y X y Z son las matrices que relacionan los registros con los elementos de β y u, respectivamente. Recordemos que los supuestos probabilísticos del modelo lineal mixto con distribución normal multivariante ($\mathcal{NMV}$) son:

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim \mathcal{NMV}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right)$$

Lo cual induce:

$$y \sim \mathcal{NMV}\left(X\beta, ZGZ^T + R\right)$$

Por lo tanto, en este caso el MPLI no requiere que se conozcan los efectos fijos y, además, los registros fenotípicos se corrigen por estos efectos empleando el estimador β^0. Vale la pena recordar la relación entre el modelo lineal mixto y el MPLI estudiada en la sección 3.5, al solucionar las ecuaciones de modelos mixtos se encuentra el MPLI de u. Ahora bien, computar el MPLI de u requiere invertir la matriz de covarianzas de los efectos aleatorios, la cual tiene dimensión pxp, siendo p el número de efectos aleatorios del modelo, además del cómputo necesario para resolver las ecuaciones de modelos mixtos.

Por otro lado, si se desea computar directamente el MPLI empleando $C^T V^{-1}(y - X\beta^0)$, se requiere invertir la matriz de covarianzas de los registros fenotípicos que tiene dimensión nxn. ¿Qué es más ventajoso?; en evaluación genética animal, es muy frecuente que el número de animales a evaluar exceda el número de datos, esto es, el número de efectos aleatorios es mayor al tamaño de muestra; entonces tendría un mayor costo computacional invertir G que invertir V. Sin embargo, una contribución muy relevante desde el punto de vista computacional fue hecha por Henderson [7], quien desarrolló las denominadas reglas de Henderson, las cuales permiten construir G^{-1} directamente a partir del pedigrí, esto representa un enorme ahorro computacional, ya que no se precisa invertir la matriz. Este desarrollo, sumado a la existencia de algoritmos computacionalmente eficientes para resolver grandes conjuntos de ecuaciones permitieron que la solución de las ecuaciones de modelos mixtos sea la vía más utilizada para encontrar los MPLI de los valores genéticos de los individuos objeto de análisis.

3 CAPÍTULO TRES

CONCEPTOS FUNDAMENTALES DE LA TEORÍA DE MODELOS LINEALES

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá.

3.1. Generalidades

Este capítulo pretende brindar elementos básicos de la teoría de modelos lineales (efectos fijos y aleatorios), convirtiéndose en un primer paso para comunicar al lector los conocimientos necesarios para aplicarlos en evaluaciones genéticas. Se presentan algunas definiciones básicas, varias formas de clasificar los modelos lineales, estimación de parámetros, algunas propiedades de los estimadores y el concepto de estimabilidad de funciones lineales de los parámetros, al igual que la discusión del enfoque frecuentista y bayesiano. Vale la pena destacar que, por el ámbito de este texto, no se toca el tema de prueba de hipótesis, ni de estimación por intervalo, en consecuencia, el capítulo se enfoca en estimación puntual. Para entrar en materia, definimos un modelo estadístico como la clase de modelos matemáticos que buscan describir el proceso que generó los datos y que tienen al menos dos componentes:

Componente sistemático: una función matemática que relaciona la esperanza (u otro parámetro de localización) de la variable respuesta, con las variables explicativas o independientes.

Componente estocástico: modela la variación aleatoria de la variable respuesta mediante una distribución de probabilidad.

Por lo tanto, siempre que se defina un modelo estadístico, se debe definir la función que muestra la relación entre la esperanza de la variable respuesta y las variables explicativas y los supuestos de distribución que se hacen.

Un modelo estadístico de tipo lineal se define como aquel en el cual la esperanza de los registros ($E\left[y\right]$) se expresa como una combinación lineal de los parámetros del modelo. Los siguientes son ejemplos de modelos lineales:

$$E\left[y\right] = \beta_0 + \beta_1 X_1$$

$$E\left[y\right] = \beta_0 + \beta_1 X_1^2 + \beta_2 X_1 X_2^3$$

$$E\left[y\right] = \beta_0 + \beta_1 X_1^2 + \beta_2 \exp^{X_1 + X_2^2}$$

$$E\left[y\right] = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_1 X_2^3$$

En donde los β_j, con $j = 0, 1, 2$, son los parámetros desconocidos de cada modelo y las X_l con $l = 1, 2$, son las variables explicativas.

En los ejemplos anteriores, las relaciones *entre los parámetros del modelo y los registros* son de tipo lineal. De estos ejemplos también se puede ver que los modelos lineales pueden describir tendencias no lineales. Es decir, la forma de la función resultante no es lineal. Por ejemplo, un modelo lineal correspondiente a un polinomio de orden 3 podría ser útil para describir un fenómeno en el que la variable respuesta y la variable explicativa tienen una relación no lineal (cúbica en este caso); sin embargo, la relación entre la variable respuesta y los parámetros sigue siendo lineal. Como ejemplo considere el siguiente modelo ajustado para explicar el consumo de alimento en cerdas lactantes tomado de Martínez et al [8]:

$$\widehat{E\left[y\right]} = 2.27 + 0.31x - 0.01x^2$$

En donde $\widehat{E\left[y\right]}$ es la media estimada de consumo y x es el tiempo en lactancia expresado en días $x \in \left[0, 21\right]$. El concepto de estimación, no es importante por ahora, más adelante se presentan diferentes métodos de estimación y se definen los registros ajustados. La anterior ecuación muestra una tendencia curvilínea según la cual las cerdas consumen alimento hasta alcanzar un pico después del cual el consumo decrece.

A continuación, se presentan ejemplos de modelos estadísticos no lineales:

$$E\left[y\right] = \beta_0 x^{\beta_1}$$

$$E\left[y\right] = \beta_0 + \beta_1 x + \exp^{\beta_2 x - \beta_3}$$

$$E\left[y\right] = \beta_0 \exp^{\beta_1 t - \beta_2 z + \beta_3 zt}$$

Donde los β_j, con $j = 0, 1, 2, 3$ son los parámetros desconocidos de cada modelo y x, z y t son las variables explicativas. En cada caso se observa que la relación entre la esperanza de los registros y los parámetros es de tipo no lineal.

Los modelos estadísticos no estiman los datos observados de manera perfecta; siempre existirá una variación de los registros alrededor de su valor esperado(variación conocida como error aleatorio). De acuerdo con Searle [9], esta diferencia se debe a errores en la medición de los registros y a deficiencias del modelo. Un ejemplo de errores en la medición se presenta cuando se toman pesos de animales, como por ejemplo los bovinos. En estos casos, las básculas empleadas no registran el peso de manera precisa. Así, el peso real puede ser 200.2434 kg, pero el valor que se registrará será 200.2 kg. Un ejemplo de error en la especificación del modelo sería ajustar una función lineal para describir el crecimiento de un animal cuando en realidad el fenómeno es de tipo curvilíneo. En general, el error incluye factores que afectan la característica pero que no se tienen en cuenta en el modelo. Siguiendo la definición del error este se escribe así:

$$e = y - E\left[y\right]$$

Por lo tanto:

$$E\left[e\right] = E\left[y - E\left[y\right]\right] = E\left[y\right] - E\left[y\right] = 0$$

Estos desvíos de las observaciones de su esperanza son de naturaleza aleatoria. Se asume que su varianza existe y se denomina varianza del error.

Por ahora nos enfocaremos en un modelo lineal con efectos fijos, los modelos lineales mixtos serán estudiados más adelante. Cada registro es modelado como una función de p variables explicativas conocidas y un error de naturaleza aleatoria. Dicho error se asume distribuido normalmente con esperanza nula y en el caso más simple del modelo lineal general se hace el supuesto de homocedasticidad, es decir, que los errores de todos los registros tienen la misma varianza. Las variables explicativas se denotarán como $X_1, X_2, \ldots, X_p$. Si se tienen n registros, el i-ésimo puede escribirse como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i$$

Donde x_{ij} y β_j con $j = 1, 2, \cdots, p$ son respectivamente, el valor observado de la variable X_j para la i-ésima observación y el parámetro desconocido (efecto) asociado a esta variable, el parámetro β_0 corresponde al intercepto, y e_i es el error aleatorio asociado al registro para el cual se asume:

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \forall i = 1, 2, \cdots, n$$

Esto implica que:

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right)$$

Si se consideran las n observaciones se tendrá lo siguiente:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + e_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + e_n \end{aligned}$$

Téngase en cuenta que, en este sistema de ecuaciones lineales, los parámetros son incógnitas. Este sistema se puede escribir explícitamente en lenguaje matricial así:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \sum_{j=1}^{p} X_{1j}\beta_j + e_1 \\ \beta_0 + \sum_{j=1}^{p} X_{2j}\beta_j + e_2 \\ \vdots \\ \beta_0 + \sum_{j=1}^{p} X_{nj}\beta_j + e_n \end{bmatrix}$$

La notación general[1] de este modelo es:

$$y = X\beta + e$$

En donde y es un vector columna aleatorio de orden nx1 que contiene los registros, β es un vector de orden $(p + 1)$x1, cuyas entradas son parámetros desconocidos correspondientes a los efectos fijos, e es un vector aleatorio de errores de orden nx1 y X es una matriz de orden nx$(p + 1)$, que relaciona los registros con los elementos del vector de parámetros β la cual es conocida como la *matriz de diseño*.

3.2. Clasificación de los modelos lineales

Existen varias maneras de clasificar los modelos lineales. Una de ellas es de acuerdo con la naturaleza de los efectos incluidos, los cuales pueden ser fijos o aleatorios, dando lugar a modelos de efectos fijos, efectos aleatorios y efectos mixtos. Otra, es según la naturaleza de las variables explicativas, las cuales pueden ser de tipo cuantitativo o cualitativo. Una tercera forma de clasificar los modelos lineales, es según el rango de la matriz de diseño (ver sección 8), en este caso, se tienen modelos de rango completo y modelos de rango incompleto.

[1] En el álgebra matricial, se suelen usar letras minúsculas para denotar vectores y letras mayúsculas para denotar las matrices. Por otro lado, en la teoría de la probabilidad, se emplean letras mayúsculas para denotar variables aleatorias y letras minúsculas para denotar valores particulares de las mismas. Así, al denotar vectores aleatorios, se tiene la duda sobre el uso de letras mayúsculas o minúsculas. La elección varía con el autor, en este caso, se seguirán empleando letras minúsculas para denotar vectores, aún si son aleatorios.

3.2.1. Clasificación de los modelos lineales de acuerdo a la naturaleza de los efectos

En este caso, los modelos pueden ser de efectos fijos o simplemente fijos, cuando la totalidad de los efectos contemplados son de tipo fijo, de efectos aleatorios o aleatorios cuando los efectos en el modelo son en su totalidad de tipo aleatorio; finalmente, tenemos modelos de efectos mixtos o simplemente mixtos, cuando se tienen tanto efectos fijos como aleatorios. En este punto, es preciso definir efectos fijos y aleatorios y discutir algunas de sus implicaciones.

Efectos fijos: cuando se tiene un modelo en el cual los efectos de los diferentes factores estudiados son tratados como constantes desconocidas se dice que tales efectos son fijos. Para variables cualitativas, un efecto se declara fijo cuando los niveles que se tienen en el estudio son todos los posibles niveles de tal efecto o cuando la inferencia se realizará solo para los niveles que se tienen en el estudio Littell et al [10]. Bajo el enfoque frecuentista, cuando un efecto es fijo, no presenta componentes de varianza asociados pues son constantes.

Efectos aleatorios: cuando los niveles de un factor en el modelo corresponden a una muestra aleatoria que está representando una población de niveles se dice que los efectos de dicho factor son aleatorios. Es decir, los niveles que se tienen en el modelo son una muestra aleatoria de una población de niveles. En el caso de los efectos aleatorios, estos tienen un componente de varianza asociado pues corresponden a una variable aleatoria. El supuesto usual es que los efectos aleatorios siguen una distribución normal o Gaussiana con vector de medias nulo y matriz de covarianzas desconocida.

Cuando se va a plantear un modelo lineal, la decisión sobre declarar un efecto fijo o aleatorio no siempre es fácil y puede ser objeto de discusión. Siempre se deben tener en cuenta los objetivos de la investigación y el uso que se dará a los resultados obtenidos de la misma. Consideremos el siguiente ejemplo. Si en cierta investigación se quiere analizar el efecto de la semana del año, en un año particular, sobre una variable de interés, la pregunta natural es si la semana del año es un efecto fijo o aleatorio. Si se tienen registros en todas las semanas del año, el efecto de semana sería fijo puesto que se tienen todos los posibles niveles de dicho factor. En adición, no existiría necesidad de extrapolar pues en este caso se tuvo información para todas las semanas del año. De otro lado, si se tiene información para un grupo particular de semanas como por ejemplo las 24 primeras del año y los investigadores sólo están interesados en concluir para esa ventana de tiempo, el efecto de la semana también sería fijo. En un tercer escenario, si las semanas incluidas en el modelo son una muestra aleatoria de las 52 semanas del año y se desea concluir para la totalidad del año, el efecto de semana sería aleatorio.

Ahora bien, en ciertos casos, la naturaleza de la variable facilita la decisión acerca de cómo modelarla. En genética cuantitativa, al analizar poblaciones multirraciales, el efecto de la raza o grupo racial de un animal es tratado como fijo, puesto que las conclusiones solo se realizarán para los grupos raciales considerados. De otro lado, los efectos genéticos como; por ejemplo, los genéticos aditivos directos, son de naturaleza aleatoria debido a que el genotipo de un animal está compuesto por una muestra aleatoria de la mitad de los genes de su padre y la mitad de los genes de su madre. En la teoría de la genética cuantitativa el genotipo es una variable aleatoria y al ser el valor genético una función del genotipo, este también es una variable aleatoria.

3.2.2. Clasificación de los modelos lineales de acuerdo con la naturaleza de las variables explicativas

Aquí interesa diferenciar los modelos lineales[2] en los cuales las variables explicativas son de tipo continuo, aquellos con variables explicativas que son de cualitativos (o clasificación) y aquellos que incluyen los dos tipos de variables.

Modelos de regresión: estos modelos contemplan solamente variables explicativas continuas. Como se estudió en la sección 9, estas son variables que toman valores en conjuntos de dimensión infinita y no son numerables como por ejemplo la altura a la cruz de un animal, el contenido de proteína de una dieta o el peso a una edad determinada. Debe tenerse en cuenta que, en la práctica, los modelos de regresión se emplean cuando se tienen variables explicativas, medidas en una escala cuantitativa; pero el término cuantitativo no implica continuidad, ya que variables como por ejemplo los conteos son de tipo cuantitativo, pero no son variables continuas. Un ejemplo sería estudiar el efecto del tamaño de camada en cuyes sobre el peso medio de los animales al nacimiento; otros ejemplos serían el conteo de células somáticas en la leche de búfalas y el número de huevos puestos por año en un galpón. Estas variables tomarían valores en los enteros positivos, que a su vez son un subconjunto de los números reales. Recordemos que, en este caso, las variables son discretas ya que toman valores en un conjunto finito. En un modelo de regresión las variables explicativas son llamadas regresores.

Modelos con variables explicativas de tipo cualitativo (ANOVA): en este caso, las variables independientes son de clasificación, es decir, son factores cuyos niveles corresponden a categorías en las cuales se clasifican los registros o, en otras palabras, en vez de tener observaciones de variables cuantitativas para cada registro, las observaciones están clasificadas en categorías definidas por niveles o combinaciones de niveles de las variables explicativas. Un ejemplo sería el peso a los 24 meses de

[2]Aunque la siguiente clasificación aplica para modelos con efectos mixtos, la notación se presenta para el modelo lineal general.

bovinos de carne en función del manejo discriminado en tres categorías: estabulación completa, semi-estabulación y pastoreo. En este caso, la variable explicativa en su forma de observación original no puede tomar valores numéricos. Una pregunta natural es ¿cómo se construye la matriz de diseño en este caso, dado que las variables independientes no toman valores numéricos?, lo que se hace, es llevar a cabo un procedimiento conocido como regresión en variables ficticias (dummy en inglés). Lo que se busca, es estimar los efectos de cada uno de esos niveles. Este proceso se detalla más adelante.

Modelos de análisis de covarianza: en este caso, se consideran modelos en los cuales se tienen tanto factores de clasificación como variables continuas entre las variables explicativas. Así, estos modelos pueden verse como combinaciones de modelos de regresión y de clasificación.

3.2.3. Clasificación según el rango de la matriz de diseño

Cuando la matriz de diseño es de rango columna completo, se dice que el modelo es de rango completo, mientras que cuando no lo es, el modelo es denominado de rango incompleto.

Normalmente, los modelos de regresión son asociados con los modelos de rango completo debido a que en la gran mayoría de casos lo son. Sin embargo, hay situaciones en las que existen restricciones sobre las variables explicativas que hacen que se tenga un modelo de regresión con dependencias lineales entre las columnas de la matriz de diseño. Este fenómeno se presenta en modelos de evaluación genética multirracial, en los cuales es frecuente la dependencia lineal de las columnas de las matrices de diseño de efectos genéticos fijos [11].

Veamos el EJEMPLO 3.2.3, relacionado con vacunos de carne. Se quiere expresar el área de ojo del lomo en función de la composición racial. Se tienen cuatro razas y la variable explicativa es la fracción esperada de cada raza en el individuo. En este caso se tiene un modelo de regresión. En adición, se tiene la variable heterocigosidad, la cual hace referencia a la probabilidad de alelos de diferentes razas en un locus tomado al azar. La TABLA NRO. 3.1 muestra los animales, sus composiciones raciales y la heterocigosidad.

Así, la matriz de diseño (X) tendrá seis columnas (una para el intercepto, una para porcentaje de heterocigosidad y cuatro para las razas).

El orden de las columnas es: Intercepto, heterocigosis, Raza A, Raza B, Raza C y Raza D. En las filas, se encuentran representados los animales en el mismo orden que aparecen en la TABLA NRO. 3.1. Así, se tiene la siguiente matriz de diseño:

TABLA 3.1: Información de animales puros y cruzados

Animal	Grupo genético	Heterocigosidad	Raza_A	Raza_B	Raza_C	Raza_D
1	A	0	1	0	0	0
2	AB	1	0.5	0.5	0	0
3	AB	1	0.5	0.5	0	0
4	AC	1	0.5	0	0.5	0
5	AD	1	0.5	0	0	0.5
6	AD	1	0.5	0	0	0.5
7	AD	1	0.5	0	0	0.5
8	A	0	1	0	0	0
9	A	0	1	0	0	0

Nota: la tabla contiene información de animales puros y cruzados de las razas A, B, C y D con la heterocigosidad y las proporciones raciales.
Fuente: elaboración propia (2024).

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0.5 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0.5 \\ 1 & 1 & 0.5 & 0 & 0 & 0.5 \\ 1 & 1 & 0.5 & 0 & 0 & 0.5 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Nótese que la suma de las columnas 3, 4, 5 y 6 es igual a la columna 1, es decir, la columna asociada al intercepto. Por lo tanto, la matriz X no es de rango columna completo. Otra dependencia lineal que se tiene en dicha matriz es la siguiente: $2(X_{.4} + X_{.5} + X_{.6}) = X_{.2}$, siendo $X_{.j}$ con $j = 1, 2, .., 6$ la j-ésima columna de la matriz X. La restricción que se tiene aquí es que la suma de las composiciones raciales es 1; por lo tanto, el modelo se puede parametrizar en términos de tres razas ya que la composición de la cuarta es uno menos la suma de las fracciones de las otras tres, esto aliviaría los problemas de dependencia lineal, pero el ejemplo ilustra que el mero hecho de tener variables cuantitativas no garantiza que la matriz diseño sea de rango columna completo. Además, como se verá más adelante en el EJEMPLO 3.4.2, hay parametrizaciones de interés que requieren que se mantengan todas las razas.

Por otro lado, también es posible tener modelos con variables explicativas de tipo cualitativo en el que la matriz diseño es de rango columna completo, un ejemplo es el denominado modelo anova a una vía bajo la parametrización conocida como modelo de medias de celda. Así, el hecho de que un modelo lineal tenga variables explicativas de tipo cualitativo no necesariamente implica que se tiene un modelo de rango incompleto.

3.3. Modelo lineal cuando se tienen variables explicativas de tipo cualitativo

Considérese el caso de un solo factor con m niveles. Así, para el registro i del nivel j denotado como y_{ij} con $i = 1, 2, \ldots, n_j$ y $j = 1, 2, \ldots, m$ siendo n_j el número de registros en el nivel j, el modelo es:

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + e_{ij}$$

Donde β_j con $j = 1, 2, \ldots, m$ es el efecto del j-ésimo nivel del factor y β_0 es un término común a todas las observaciones, conocido como el intercepto. Se tiene que:

$$\sum_{j=1}^{m} n_j = n$$

Cuando $n_1 = n_2 = \cdots = n_m$, el modelo se dice balanceado. Teniendo en cuenta que cada observación pertenecerá a un nivel de cada factor, para el i-ésimo registro, la variable explicativa x_{ij} valdrá uno si ese registro pertenece al nivel k y cero en otro caso. Las variables x_{ij} se conocen como variables ficticias. En este caso, la forma matricial del modelo es:

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ \vdots \\ y_{1m} \\ y_{2m} \\ \vdots \\ y_{n_m m} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ \vdots \\ e_{n_1 1} \\ \vdots \\ e_{1m} \\ e_{2m} \\ \vdots \\ e_{n_m m} \end{bmatrix}$$

Así, existirán dos unos en cada fila de la matriz de diseño, uno para el intercepto y otro para el nivel al cual pertenece la observación. El número de parámetros es $m + 1$.

El procedimiento es el mismo para el caso general, en el cual se tienen t factores, cada uno con m_j niveles ($j = 1, 2, \ldots, t$); en cada uno de los cuales se tienen n_{kj} registros con $k = 1, 2, \ldots, m_j$. En este caso, si no se consideran interacciones y se ajusta un intercepto, en cada fila habrán $t + 1$ unos, uno asociado con el intercepto y uno por cada factor (pues cada registro pertenece a un único nivel de cada factor). El número de parámetros será:

$$1 + \sum_{j=1}^{t} m_j$$

Nótese que, en el caso de un solo factor presentado previamente, la columna del intercepto es la suma de las demás columnas, esto es, existe una columna que es

combinación lineal de las demás, lo cual implica que se tiene un conjunto linealmente dependiente y por lo tanto, la matriz de diseño no es de rango completo. En el caso general, la suma de las columnas de los niveles de cada factor será igual a la columna del intercepto. Además, cuando hay desbalanceo y celdas vacías, pueden existir dependencias lineales entre columnas de diferentes factores.

3.4. Modelo lineal general

En la sección 3.1 se definió como:

$$y = X\beta + e$$

Para simplificar la notación, en adelante la dimensión del vector β será p. Así, en la anterior expresión, y es un vector columna aleatorio de orden nx1 que contiene los registros, con valor esperado $E\left[y\right] = X\beta$; β es un vector de orden px1 en el que sus entradas son parámetros desconocidos correspondientes a efectos fijos, e es un vector aleatorio de errores de orden nx1 y X es la matriz de diseño de dimensión nxp. Este vector se define como el desvío de los registros de su valor esperado y es de naturaleza aleatoria. Por lo tanto, $e = y - X\beta$, por consiguiente $E\left[e\right] = E\left[y - X\beta\right] = X\beta - X\beta = 0$.

Denotamos por V la matriz de covarianzas del vector de registros y por R la del vector de errores, entonces: $e \sim (0, R)$.

Por lo tanto:

$$\begin{gathered} Var\left[y\right] = V = Var\left[X\beta + e\right] \\ = Var\left[X\beta\right] + Var\left[e\right] + 2 * Cov\left[X\beta, e^T\right] \\ = 0 + R + 2 * 0 \\ = R \\ Cov\left[y, e^T\right] = Cov\left[X\beta + e, e^T\right] = Cov\left[X\beta, e^T\right] + Cov\left[e, e^T\right] \\ = 0 + Var\left[e\right] \\ = R \end{gathered}$$

En el procedimiento anterior los ceros son matrices nulas de dimensión nxn.

Así, $y \sim (X\beta, R)$. Cabe anotar que hasta el momento no se está asumiendo una distribución particular, la notación empleada muestra que los registros son una variable aleatoria con vector de medias $X\beta$ y matriz de covarianzas $V = R$. En el modelo lineal se asume una distribución normal multivariada del vector de errores, esto es, $e \sim \mathcal{NMV}(0, R)$. Entonces, $y \sim \mathcal{NMV}(X\beta, R)$. El valor de la covarianza sigue siendo el mismo, así que se puede escribir:

$$\begin{bmatrix} y \\ e \end{bmatrix} \sim \mathcal{NMV}\left[\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} R & R \\ R & R \end{bmatrix}\right]$$

Bajo el supuesto de errores homocedásticos e independientes $R = \sigma^2 I$, donde σ^2 es la varianza del error, e I es la matriz de identidad. En este caso:

$$\begin{bmatrix} y \\ e \end{bmatrix} \sim \mathcal{NMV}\left[\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} I_n & I_n \\ I_n & I_n \end{bmatrix}\right]$$

3.4.1. Estimación de parámetros de localización

En el modelo lineal general tenemos dos grupos de parámetros desconocidos, aquellos asociados a la matriz de covarianza V y aquellos asociados al vector de medias $X\beta$, bajo el supuesto de errores homocedásticos y no correlacionados, es decir, $V = \sigma^2 I$; el único parámetro asociado a V es σ^2, en tanto que para el vector de medias el parámetro correspondiente es β. En este caso, se dice que σ^2 es un parámetro de dispersión, mientras que β es un parámetro de localización.

Estimación por mínimos cuadrados: se quiere minimizar la suma de cuadrados de las desviaciones de los registros respecto a su valor esperado. Así, la idea detrás del método de mínimos cuadrados es simple: encontrar el estimador de β que haga que la suma de cuadrados de los desvíos de los valores observados respecto a los esperados (S) sea mínima. En primer lugar, se presenta el método de mínimos cuadrados ordinarios (MCO).

Mínimos cuadrados ordinarios: bajo los mínimos cuadrados ordinarios, la función S está definida como:

$$S = e^T e = (y - X\beta)^T (y - X\beta) = y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta$$

Nótese que $y^T X\beta$ es un escalar y por lo tanto, este es igual a su transpuesta, es decir: $y^T X\beta = (y^T X\beta)^T = \beta^T X^T y$, de aquí se sigue que:

$$S = y^T y - 2\beta^T X^T y - \beta^T X^T X\beta$$

Ahora, empleando diferenciación matricial, se deriva la expresión anterior respecto al vector de parámetros β y se obtiene:

$$\begin{aligned} \frac{\partial S}{\partial \beta} &= \frac{\partial}{\partial \beta}(y^T y - 2\beta^T X^T y - \beta^T X^T X\beta) \\ &= 2\frac{\partial}{\partial \beta}(\beta^T X^T y) - \frac{\partial}{\partial \beta}(\beta^T X^T X\beta) \\ &= 2X^T y - 2X^T X\beta \end{aligned}$$

Esta expresión se iguala a cero para obtener el valor crítico de β, el cual corresponde a un mínimo ya que se puede probar que la función S es convexa o cóncava hacia arriba.

$$2X^T y - 2X^T X\widehat{\beta} = 0 \Rightarrow X^T X\widehat{\beta} = X^T y$$

El anterior conjunto de ecuaciones se conoce como las *ecuaciones normales*. En este punto, es importante hacer una diferenciación entre los modelos de rango completo y los de rango incompleto.

Téngase en cuenta que $r(X) = r(X^TX)$, donde $r(\cdot)$ representa el rango de una matriz. Por lo tanto, si el modelo es de rango completo, X^TX es de rango completo y en consecuencia es no singular. En este caso, la solución de las ecuaciones normales es única y se obtiene así:

$$(X^TX)^{-1}(X^TX)\widehat{\beta} = (X^TX)^{-1}X^Ty$$

$$\Rightarrow \widehat{\beta} = (X^TX)^{-1}X^Ty$$

A esta solución única se le conoce como estimador de mínimos cuadrados ordinarios (EMCO). De otro lado, en el modelo de rango incompleto, la teoría de matrices establece que el sistema será consistente (soluble) si y solo si:

$$r(X^TX \vdots X^Ty) = r(X^TX)$$

Efectivamente, este sistema de ecuaciones es soluble porque:

a) Al aumentar el número de columnas de una matriz el rango no puede disminuir, este se mantendrá o aumentará, de aquí: $r(X^TX \vdots X^Ty) \geq r(X^TX)$

b) $r(X^TX \vdots X^Ty) = r(X^T(X \vdots y)) \leq r(X^T) = r(X^TX)$.

De *a)* y *b)* se sigue la igualdad y por ende, el sistema es consistente. Ahora bien, dado que este sistema es consistente y la matriz de coeficientes no es de rango completo, existen infinitas soluciones. Haciendo uso de otro teorema de la teoría de matrices, se puede encontrar una expresión para una solución particular. El sistema de ecuaciones normales será consistente, si y solo si:

$$(X^TX)(X^TX)^-X^Ty = X^Ty$$

En donde $(X^TX)^-$ es una inversa generalizada de (X^TX) De aquí, una solución particular, es:

$$\beta^0 = (X^TX)^-X^Ty$$

Nótese que para el modelo de rango completo, el estimador del vector de parámetros se denota como $\widehat{\beta}$, mientras para el caso del modelo de rango incompleto, una solución de las ecuaciones normales se denota como β^0. Esta diferenciación se hace para recordar que en el primer caso la solución es única, mientras en el segundo, existen infinitas soluciones (una por cada inversa generalizada que se elija). Como se estudiará más adelante, en el caso del modelo de rango incompleto, existe el concepto de funciones estimables, las cuales son invariantes a la elección de la inversa generalizada

de X^TX. Ahora bien, debe tenerse en cuenta que, en la derivación de los estimadores de mínimos cuadrados, no se hizo ningún supuesto de distribución sobre los errores.

Debido a la existencia de infinitas soluciones cuando el modelo no es de rango completo, una pregunta natural es si cualquier solución de las ecuaciones normales minimiza la suma de cuadrados (S), la respuesta a esta pregunta se encuentra en un teorema de la teoría de modelos lineales que establece que, cualquier solución de las ecuaciones normales, minimiza la suma de cuadrados S.

Mínimos cuadrados generalizados (MCG): para derivar los estimadores de mínimos cuadrados generalizados (EMCG), se tiene en cuenta una estructura más general de la matriz de covarianzas de e denotada como $Var\left[e\right] = R$, la matriz R es una matriz de dimensión nxn conocida y definida positiva. La estructura de R dependerá de la situación que se esté analizando. Un ejemplo es cuando se modelan errores correlacionados. Nótese, que los MCO corresponden al caso particular en el cual $R = \sigma^2 I$.

En este caso, se debe minimizar la siguiente función [9]:

$$(y - X\beta)^T R^{-1}(y - X\beta)$$

Expandiendo esta expresión:

$$(y - X\beta)^T R^{-1}(y - X\beta) = y^T R^{-1} y - y^T R^{-1} X\beta - \beta^T X^T R^{-1} y + \beta^T X^T R^{-1} X\beta$$

Como en el caso de los MCO: $y^T R^{-1} X\beta = \beta^T X^T R^{-1}$ y, así:

$$(y - X\beta)^T R^{-1}(y - X\beta) = y^T R^{-1} y - 2\beta^T X^T R^{-1} y + \beta^T X^T R^{-1} X\beta$$

Diferenciando respecto a β:

$$\frac{\partial}{\partial \beta}\left[(y - X\beta)^T R^{-1T}(y - X\beta)\right] = -2X^T R^{-1} y + 2X^T R^{-1} X\beta$$

Igualando a cero y escribiendo la ecuación resultante en términos de $\widehat{\beta}$, se tiene:

$$-2X^T R^{-1} y + 2X^T R^{-1} X\widehat{\beta} = 0$$

$$\Rightarrow X^T R^{-1} X\widehat{\beta} = X^T R^{-1} y$$

Estas son las ecuaciones normales de los MCG.

Para el caso del modelo de rango completo:

$$EMCG(\beta) = \widehat{\beta}_{MCG} = (X^T R^{-1} X)^{-1} X^T R^{-1} y$$

Análogo al caso de los MCO, para el modelo de rango incompleto, una solución particular del sistema de ecuaciones normales es:

$$EMCG(\beta) = \beta^0_{MCG} = (X^T R^{-1} X)^- X^T R^{-1} y$$

Máxima verosimilitud: en los métodos presentados hasta ahora no se han hecho supuestos de distribución del vector e. Bajo el método de máxima verosimilitud, se debe realizar un supuesto sobre la distribución de dicho vector para poder escribir matemáticamente la función de verosimilitud; el supuesto usual es asumir normalidad. La función de verosimilitud tiene la misma forma de la función de densidad conjunta de las observaciones, dados los parámetros; la diferencia radica en la forma de ver estas funciones. En probabilidad, la densidad conjunta es vista como una función de las variables aleatorias $Y_1, Y_2, \ldots, Y_n$, pues, se asume, que los parámetros son conocidos, mientras que, en estadística, la función de verosimilitud es vista como una función de los parámetros, pues lo que se tiene en problemas de la vida real son realizaciones de las variables $Y_1, Y_2, \ldots, Y_n$ (los registros) y se desconocen los parámetros. Si se consideran el vector de parámetros θ, el vector de registros y la densidad de y dado θ, se denota como $f(y \mid \theta)$, mientras que la función de verosimilitud se denota como $L(\theta \mid y)$, así:

$$L(\theta \mid y) = f(y \mid \theta)$$

Nota: se emplea la notación θ para el vector de parámetros porque este contiene además del vector β los componentes de varianza. En lo que sigue, se detalla el método para la estimación del vector β bajo normalidad multivariada de e.

Si los registros son independientes, esta función se puede construir como el producto de las densidades marginales de los mismos. Una vez se tiene esta función, se procede a maximizarla en términos de los parámetros, de allí el nombre del método. Esta idea se debe a Sir Ronald A Fisher. En adelante, se denotará la función de verosimilitud como L.

Bajo normalidad, se tiene que $y \sim \mathcal{NMV}(X\beta, V)$ y la función de verosimilitud se escribe así:

$$L = (2\pi)^{\left(-\frac{1}{2}n\right)} \mid V \mid^{\frac{-1}{2}} \exp\left[\frac{-1}{2}(y - X\beta)^T V^{-1}(y - X\beta)\right]$$

Donde n es el número de registros. Maximizar esta ecuación es equivalente a maximizar su logaritmo natural, pero el proceso de diferenciación matricial y de resolver las ecuaciones resultantes de igualar la primera derivada a cero en términos de β, es más simple. Usando el logaritmo natural de la función se elimina el exponencial haciendo el proceso de maximización más sencillo.

$$\ln(L) = \frac{-1}{2}(n \ln(2\pi) + \ln(\mid V \mid) + (y - X\beta)^T V^{-1}(y - X\beta))$$

Ahora, aplicando la diferenciación matricial:

$$2\frac{\partial \ln(L)}{\partial \beta} = \frac{\partial}{\partial \beta}\left[-(y - X\beta)^T V^{-1}(y - X\beta)\right]$$

$$= \frac{-\partial}{\partial \beta}\left[y^T V^{-1} y - y^T V^{-1} X\beta - \beta^T X^T V^{-1} y + \beta^T X^T V^{-1} X\beta\right]$$

$$= 2\frac{\partial}{\partial \beta}(y^T V^{-1} X\beta) - \frac{\partial}{\partial \beta}(\beta^T X^T V^{-1} X\beta)$$

$$= 2X^T V^{-1} y - 2X^T V^{-1} X\beta$$

Igualando esta derivada a cero:

$$\frac{\partial \ln(L)}{\partial \beta} = 0$$

$$\Rightarrow X^T V^{-1} X\beta = X^T V^{-1} y$$

Como $V = R$, este es el mismo set de ecuaciones obtenido con el método de los mínimos cuadrados generalizados. Por lo tanto, bajo normalidad, el estimador de máxima verosimilitud (EMV) es equivalente al EMCG.

Así:

$$\widehat{\beta}_{MV} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

Es el EMV para el caso de modelos de rango completo, en tanto que

$$\beta^0_{MV} = (X^T V^{-1} X)^{-} X^T V^{-1} y$$

Es una solución particular al sistema de ecuaciones para modelos de rango incompleto.

Mejor estimación lineal insesgada (MELI): Cuando se quiere estimar una combinación lineal de los parámetros $k^T\beta$, donde k es un vector de dimensión px1 que contiene los coeficientes de la combinación, es deseable que el estimador sea insesgado y tenga mínima varianza (mejor). Si el estimador tiene una forma lineal, es decir, es una combinación lineal de los registros ($t^T y$), tenemos el denominado Mejor Estimador Lineal Insesgado (MELI o BLUE por sus siglas en inglés).

En primer lugar, si se desea que el estimador sea insesgado, se debe satisfacer lo siguiente:

$$E\left[t^T y\right] = k^T\beta$$

$$\Rightarrow t^T E\left[y\right] = k^T\beta$$

$$\Rightarrow t^T X\beta = k^T\beta$$

Si se quiere que la anterior condición se tenga para todo β, entonces:

$$t^T X = k^T \Leftrightarrow t^T X - k^T = 0$$

Siendo esta la condición de insesgamiento. Ahora, para que el estimador sea "mejor", es decir, tenga mínima varianza en la clase de los estimadores lineales, se debe minimizar:

$$Var\left[t^T y\right] = t^T V t$$

Sujeto a la restricción: $t^T X = k^T$. Para ello, se emplea una técnica conocida como multiplicadores de Lagrange (los siguientes cálculos pueden ser omitidos por el lector que no esté interesado en este tipo de detalles). Si se denota por 2δ el vector de multiplicadores de Lagrange, la función que se debe minimizar es:

$$l = t^T V t - 2\delta^T (X^T t - k)$$

Esta expresión se minimiza respecto a t obteniéndose:

$$\frac{\partial l}{\partial t} = 2Vt - 2X\delta$$

$$\frac{\partial l}{\partial t} = 0 \Leftrightarrow 2Vt - 2X\delta = 0$$

$$\Leftrightarrow Vt = X\delta$$

En este punto, es donde puede verse la conveniencia de elegir el vector de multiplicadores de Lagrange como 2δ, pues el escalar 2 se cancela en los dos lados de la ecuación resultante de derivar respecto a t. Ahora bien, como V es no singular, se puede escribir la última expresión como sigue:

$$t = V^{-1} X\delta$$

Si la ecuación $l = t^T V t - 2\delta^T (X^T t - k)$ se deriva respecto a δ, se corrobora que efectivamente se cumple la condición de insesgamiento al igualar la ecuación resultante a cero, pues:

$$\frac{\partial l}{\partial \delta} = -2X^T t + 2k$$

$$\frac{\partial l}{\partial \delta} = 0 \Leftrightarrow X^T t = k$$

$$\Leftrightarrow t^T X = k^T$$

Reemplazando $t^T = \delta^T X^T V^{-1}$ en esta igualdad se tiene:

$$k^T = t^T X = \delta^T X^T V^{-1} X$$

Este es un sistema lineal de ecuaciones, si el modelo es de rango completo la solución tiene la forma:

$$\delta^T = k^T (X^T V^{-1} X)^{-1}$$

Con este valor de δ^T, se obtiene la siguiente expresión para t^T:

$$t^T = k^T(X^TV^{-1}X)^{-1}X^TV^{-1}$$

Por lo tanto, el MELI de $k^T\beta$ es:

$$k^T(X^TV^{-1}X)^{-1}X^TV^{-1}y$$

En este caso la solución para δ^T es única. Antes de discutir el caso del modelo de rango incompleto, donde esta solución no es única, se presentan algunas propiedades del MELI.

a) El MELI garantiza que, en la clase de los estimadores lineales, este tiene la menor varianza. A continuación, se calcula dicha varianza:

$$Var\left[MELI(k^T\beta)\right] = Var\left[k^T(X^TV^{-1}X)^{-1}X^TV^{-1}y\right]$$
$$= k^T(X^TV^{-1}X)^{-1}X^TV^{-1}VV^{-1}X(X^TV^{-1}X)^{-1}k$$
$$= k^T(X^TV^{-1}X)^{-1}X^TV^{-1}X(X^TV^{-1}X)^{-1}k$$
$$= k^T(X^TV^{-1}X)^{-1}k$$

b) El MELI es único

c) Finalmente, si se tiene otra combinación lineal de interés, $c^T\beta$, la covarianza entre los MELI de $k^T\beta$ y $c^T\beta$ será:

$$Cov\left[MELI(k^T\beta), MELI(c^T\beta)^T\right]$$
$$= k^T(X^TV^{-1}X)^{-1}X^TV^{-1}Cov\left[y, y^T\right]V^{-1}X(X^TV^{-1}X)^{-1}c$$
$$= k^T(X^TV^{-1}X)^{-1}X^TV^{-1}VV^{-1}X(X^TV^{-1}X)^{-1}c$$
$$= k^T(X^TV^{-1}X)^{-1}c$$

Ahora bien, si se quiere estimar un conjunto de combinaciones lineales de los parámetros, lo único que se debe hacer es reemplazar el vector k^T por una matriz K^T, cuyas filas contienen los coeficientes de cada una de las combinaciones lineales de interés. Además, no se define un vector de multiplicadores de Lagrange, sino una matriz de multiplicadores de Lagrange. Tras seguir un proceso similar al presentado antes, se obtiene:

$$MELI(K^T\beta) = K^T(X^TV^{-1}X)^{-1}X^TV^{-1}y$$
$$Var[MELI(K^T\beta)] = K^T(X^TV^{-1}X)^{-1}K$$

Un caso de interés es cuando se quiere estimar β, esto es, $K = I$, entonces, el MELI de β es:

$$MELI(\beta) = (X^TV^{-1}X)^{-1}X^TV^{-1}y$$

Entonces, cuando el modelo es de rango completo, el MELI de β es igual el EMCG, y bajo normalidad, es también igual al EMV. Si se asumen errores no correlacionados y homecedásticos:

$$MELI(\beta) = EMV(\beta) = EMCG(\beta) = EMCO(\beta)$$

Por lo tanto, estos estimadores comparten sus propiedades.

Cuando el modelo no es de rango completo, se tienen infinitas soluciones para t^T. En caso tal, se tiene interés en saber si existen funciones $k^T\beta$ cuyo estimador permanezca invariante a la escogencia de la inversa generalizada de $X^TV^{-1}X$. Dicho problema nos lleva a la siguiente sección, en la cual se estudiarán este tipo de funciones y al final de la cual se establece el MELI de las mismas mediante el teorema de Gauss-Markov.

3.4.2. Funciones estimables

En muchos estudios no se tiene interés en los parámetros de localización sino en funciones lineales (combinaciones lineales) de los mismos. Un ejemplo sería las diferencias entre pares medias. De otro lado, en el caso de los modelos de rango incompleto, existen infinitas soluciones a las ecuaciones normales y cualquiera de ellas minimiza la suma de cuadrados S.

Lo que hace variar la solución es la elección de la inversa generalizada de la matriz de coeficientes de las ecuaciones normales. En este caso, el interés se centra en determinar si existen combinaciones lineales de los parámetros que sean invariantes a la elección de la inversa generalizada, es decir, que sin importar que inversa generalizada se use, su valor estimado es siempre el mismo. La pregunta que surge es: ¿Qué tipo de funciones lineales de los parámetros satisfacen esta propiedad?. La respuesta es que efectivamente existe este grupo de funciones; estas se definen a continuación.

Definición: un conjunto de funciones lineales de los parámetros de localización de un modelo lineal se dice estimable si estas son iguales a funciones lineales de la esperanza del vector de registros [9, 12]. La anterior definición se escribe matricialmente como:

$$K^T\beta = T^TE[y] = T^TX\beta$$

Se requiere que la anterior condición se cumpla para todo β, lo cual implica que:

$$K^T = T^TX$$

No se exige la unicidad de T^T, solamente su existencia. Como puede verse, se tiene la misma condición de insesgamiento del MELI. Ahora bien, no siempre será sencillo

verificar si existe una matriz T que satisfaga esta condición. En Searle [9] y Henderson [3] se pueden consultar algunos métodos para chequear estimabilidad. No obstante, el siguiente teorema resulta útil para examinar la estimabilidad si se dispone de una inversa generalizada de la matriz X^TX. La demostración del teorema se omite.

Teorema 3.1. En el modelo lineal $y = X\beta + e$ con $E[y] = X\beta, Var[e] = \sigma^2 I$, un conjunto de funciones lineales de los parámetros contenidos en la matriz K^T son estimables, si y solo si, $K^T(X^TX)^-X^TX = K^T$. En dónde $(X^TX)^-$ es una inversa generalizada de la matriz X^TX.

Este teorema brinda una estrategia para comprobar estimabilidad y fue empleada por Martínez et al [8] al estimar efectos aditivos de raza y de heterosis en un grupo de bovinos cruzados. Esta aplicación se describe en más detalle en el EJEMPLO 3.4.2.

Propiedades de las funciones estimables: ahora se presentan ciertas características de las funciones estimables que las hacen atractivas.

a) El valor esperado de las observaciones es estimable.

b) Combinaciones lineales de funciones estimables son estimables.

c) Invarianza a la solución de las ecuaciones normales. Esta es una importante y atractiva propiedad de las funciones estimables. Como se discutía en la derivación del MELI, este es único para el caso de modelos de rango completo, pero existe la necesidad de establecer si existen funciones lineales de los parámetros que permanezcan invariantes a la escogencia de la inversa generalizada de X^TX. Las funciones estimables satisfacen tal condición, hecho que las hace atractivas, pues sin importar que inversa generalizada se elija para resolver las ecuaciones normales, si $K^T\beta$ es estimable entonces $K^T\beta^0$ siempre tendrá el mismo valor.

d) Cuando el modelo es de rango completo, cualquier combinación lineal de los parámetros es estimable.

e) Combinaciones lineales de $X\beta$ y de $X^TX\beta$ son estimables.

f) Funciones lineales de $E[\beta^0]$ son estimables.

g) El número máximo de funciones estimables linealmente independientes es igual al rango de la matriz de diseño.

Los anteriores resultados se presentaron para el modelo lineal general con errores homocedásticos. Para el caso del modelo lineal con una estructura de covarianzas del error más general: $Var[e] = R$, con R simétrica y definida positiva, se tiene que $K^T\beta$ será estimable, si y solo si:

$$K^T(X^TR^{-1}X)^-X^TR^{-1}X = K^T$$

A continuación, se presenta un teorema de importancia en la teoría de modelos lineales, que establece como calcular el MELI de funciones estimables.

Teorema 3.2. *Gauss-Markov*: para el modelo lineal $y = X\beta + e$ con momentos: $E[y] = X\beta$; $Var[y] = Var[e] = \sigma^2 I$, el mejor estimador lineal insesgado (MELI) de una función paramétrica estimable $k^T\beta$ es $k^T\beta^0$, donde β^0 es cualquier solución de las ecuaciones normales: $X^T X\beta^0 = X^T y$.

Ahora veamos el EJEMPLO 3.4.2 de una población multirracial, donde se muestra una aplicación del concepto de estimabilidad en modelos lineales en genética cuantitativa. Se basa en la estructura de una población bovina multirracial estudiada en Martínez et al [8]. Se tenía interés en explorar la estimabilidad de los efectos de grupo racial definida como desvíos de la solución de cada raza con respecto a la solución de una raza o grupo racial de referencia. En la población estudiada se tenían animales F1 y Brahman gris puros, producto del apareamiento de toros de 9 razas con hembras Brahman gris. La importancia de una estimación de los efectos de grupo racial radica en que estos se emplean en el cómputo de los valores genéticos de cada individuo y en adición son útiles para la comparación de razas en términos de su desempeño medio, el capítulo 6.2 versa sobre evaluación genética multirracial.

En las poblaciones multirraciales son muy comunes los problemas de dependencia lineal [11]; por esta razón, se debe explorar la estimabilidad de efectos genéticos fijos cuando se realizan análisis que involucren los efectos genéticos aditivos de grupo racial y los efectos de heterosis, que por lo general se consideran efectos fijos. El efecto de heterosis suele estimarse empleando el porcentaje de heterocigosis de los individuos como variables de regresión. Este porcentaje no es más que la probabilidad de que un individuo tenga alelos de diferentes razas en un locus tomado al azar y es una función lineal de la composición racial esperada del animal.

En este ejemplo, se va a verificar la estimabilidad de funciones paramétricas, las cuales no son más que combinaciones lineales del vector de soluciones de efectos genéticos de grupo racial y heterosis. Para ello, se presenta un ejemplo con pocos datos en el cual solo se consideran efectos genéticos fijos. Las razas de los toros fueron: Brahman gris, Brahman rojo, Guzerat, Blanco Orejinegro, Braunvieh, Limousin, Normando, Romosinuano y Simmental. La estructura de estos datos presenta un interesante problema de estimabilidad; este se focaliza en la manera de definir los efectos genéticos fijos.

En el estudio de Martínez et al [8] se indica la estimabilidad de los efectos de heterosis y genéticos aditivos directos de raza, bajo diferentes definiciones de grupo racial y del porcentaje de heterocigosis empleando un modelo animal para un solo carácter que incluyó los valores genéticos aditivos como efectos aleatorios, los efectos aditivos de grupo racial y de heterosis como efectos genéticos fijos y otros efectos fijos de tipo ambiental. De acuerdo con sus resultados, las razas Brahman gris y rojo debían tratarse como un solo grupo denominado Brahman y en este caso, los desvíos de los promedios de las demás razas respecto al grupo Brahman eran estimables junto con la heterosis individual; por lo tanto, esta será la definición de grupos raciales que se

usará en el EJEMPLO 3.4.2. Sin embargo, por cuestiones de espacio, solo se considera un animal de cada raza, esto es, 9 animales (TABLA NRO. 3.2)

TABLA 3.2: Información de animales puros y cruzados de 8 razas

Raza	*Int*	*H*	Bra	Guz	BlO	Brv	Lim	Nor	Rom	Sim
Bra	1	0	1	0	0	0	0	0	0	0
Bra	1	1	1	0	0	0	0	0	0	0
Guz	1	1	0.5	0.5	0	0	0	0	0	0
BlO	1	1	0.5	0	0.5	0	0	0	0	0
Brv	1	1	0.5	0	0	0.5	0	0	0	0
Lim	1	1	0.5	0	0	0	0.5	0	0	0
Nor	1	1	0.5	0	0	0	0	0.5	0	0
Rom	1	1	0.5	0	0	0	0	0	0.5	0
Sim	1	1	0.5	0	0	0	0	0	0	0.5

Nota: Int=intercepto, *H*=Heterocigosis, Bra=Brahman, Guz=Guzerat, BlO=Blanco Orejinegro, Brv=Braunvieh Lim=Limousin, Nor=Normando, Rom=Romosinuano y Sim=Simmental.
Fuente: elaboración propia (2024).

La matriz de diseño tiene diez columnas (una para el intercepto, una para heterocigosis y ocho para los grupos raciales) y nueve filas. El orden de las columnas es: intercepto, heterocigosis, y fracciones raciales de Brahman, Guzerat, Blanco Orejinegro, Braunvieh, Limousin, Normando, Romosinuano y Simmental. En las filas, se encuentran representados animales de cada una de las razas en el siguiente orden: Brahman gris, Brahman rojo, Guzerat, Blanco Orejinegro, Braunvieh, Limousin, Normando, Romosinuano, Simmental.

Empleando dicha información, se tiene la siguiente matriz de diseño:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

En esta matriz existe colinealidad ya que la suma de las columnas 3, 4, 5, 6, 7, 8, 9 y 10 es igual a la columna uno, es decir la columna asociada al intercepto y por ende, la matriz X no es de rango completo; esto implica que la matriz X^TX es singular, pues debe recordarse que: $r\left[X\right] = r\left[X^TX\right]$. Aquí la matriz tiene 9 filas y 10 columnas, lo

que implica automáticamente que no es de rango completo, pero cabe aclarar que este grado de dependencia lineal se mantendrá para un tamaño de muestra arbitrario. El modelo que se ajusta es aquel con errores homocedásticos. La matriz $X^{T'}X$ es:

$$X^TX = \begin{bmatrix} 9 & 8 & 5.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 8 & 8 & 4.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 5.5 & 4.5 & 3.75 & 0.25 & 0.25 & 0.25 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.25 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 \end{bmatrix}$$

La inversa generalizada que se empleará es la de Moore-Penrose:

$$(X^TX)^- = \begin{bmatrix} 3.56 & -2.44 & -2.67 & 0.89 & 0.89 & 0.89 & 0.89 & 0.89 & 0.89 & 0.89 \\ -2.44 & 2 & 1.44 & -0.56 & -0.56 & -0.56 & -0.56 & -0.56 & -0.56 & -0.56 \\ -2.67 & 1.44 & 2.78 & -0.78 & -0.78 & -0.78 & -0.78 & -0.78 & -0.78 & -0.78 \\ 0.89 & -0.56 & -0.78 & 3.67 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & 3.67 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & -0.33 & 3.67 & -0.33 & -0.33 & -0.33 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & -0.33 & -0.33 & 3.67 & -0.33 & -0.33 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & -0.33 & -0.33 & -0.33 & 3.67 & -0.33 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 & 3.67 & -0.33 \\ 0.89 & -0.56 & -0.78 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 & 3.67 \end{bmatrix}$$

Se debe tener en cuenta el error de redondeo cuando se realizan los cómputos, ya que los programas de análisis estadístico no empelan fraccionarios, sino decimales. Por cuestión de espacio, los números se aproximaron a dos cifras decimales.

Se chequeará la estimabilidad de un conjunto de 8 funciones lineales paramétricas: Heterosis, y los desvíos de cada una de las razas respecto al grupo Brahman. Los coeficientes de estas combinaciones lineales se incluyen en la matriz K^T así:

$$K^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La primera columna de ceros en K^T muestra que las funciones de interés no incluyen el intercepto, sin embargo, esta debe incluirse para que la matriz K^T sea conformable para el producto con la matriz $(X^TX)^-$.

Ahora, de acuerdo con el Teorema 3.1, se computa el siguiente producto para verificar la estimabilidad de este conjunto de funciones:

$$K^T(X^TX)^-(X^TX) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = K^T$$

Por lo tanto, el conjunto de funciones lineales de los parámetros contenidas en K^T es estimable. Esto nos dice que bajo la estructura racial de este grupo de animales, los coeficientes asociados a los efectos de cada raza no son estimables, pero los desvíos de estos coeficientes con respecto al grupo Brahman (que incluye al Brahman gris y al rojo) son estimables junto con la heterosis individual.

3.4.3. Distribución de los estimadores de los parámetros de localización

Ahora se discutirán las propiedades de distribución de los estimadores de los parámetros de localización del modelo lineal general, bajo el supuesto de normalidad multivariada de los errores. Como se estudió previamente, asumiendo $e \sim \mathcal{NMV}(0, R)$ se tiene que $y \sim \mathcal{NMV}(X\beta, R)$ y $Cov\left[y, e^T\right] = R$.

Ahora bien, si se conoce σ^2, cualquier solución de las ecuaciones normales β^0 sigue una distribución normal puesto que β^0 es una combinación lineal de los registros, lo cual implica que también sigue una distribución normal con media y varianza:

$$E\left[\beta^0\right] = E\left[(X^TX)^-X^Ty\right] = (X^TX)^-X^TX\beta = H\beta$$

$$Var\left[\beta^0\right] = Var\left[(X^TX)^-X^Ty\right] = \sigma^2(X^TX)^-X^TX((X^TX)^-)^T$$

Donde $H = (X^TX)^-X^TX$.

Por lo tanto: $\beta^0 \sim \mathcal{NMV}(H\beta, \sigma^2(X^TX)^-X^TX((X^TX)^-)^T)$.

En el caso del modelo de rango completo, puesto que X^TX es invertible, se tiene que:

$$E\left[\widehat{\beta}\right] = E\left[(X^TX)^{-1}X^Ty\right] = (X^TX)^{-1}X^TX\beta = \beta$$

$$Var\left[\widehat{\beta}\right] = Var\left[(X^TX)^{-1}X^Ty\right] = \sigma^2(X^TX)^{-1}X^TX((X^TX)^{-1})^T$$

$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$$

$$= \sigma^2(X^TX)^{-1}$$

Recordemos que en este caso β es estimable, mientras que en el modelo de rango incompleto no lo es. Es posible probar la existencia de una inversa generalizada que permite que la varianza de β^0 se pueda expresar de manera análoga al caso del modelo de rango completo:

$$Var\left[\beta^0\right] = \sigma^2(X^TX)^-$$

Hasta aquí, se tiene que cualquier solución de las ecuaciones normales sigue una distribución normal multivariada y se han derivado los parámetros de dicha distribución (vector de medias y matriz de covarianzas).

Recordemos que, en el caso del modelo de rango incompleto, no se tiene interés en el vector β sino en funciones estimables del mismo. Para este modelo se sigue que:

$$Var\left[K^T\beta^0\right] = \sigma^2K^T(X^TX)^-K$$

Por lo tanto, los elementos de la diagonal de la matriz de covarianzas de las funciones estimables, se emplean para obtener los errores estándar de los estimadores, ya que estos son las raíces cuadradas de las varianzas de los estimadores que se encuentran en la diagonal principal de $\sigma^2K^T(X^TX)^-K$.

Como σ^2 es desconocido, $Var\left[K^T\beta^0\right]$ se estima reemplazando σ^2 por alguno de sus estimadores. Bajo el supuesto de normalidad multivariada, se tiene que el estimador de máxima verosimilitud o máximo-verosímil de la varianza del error es:

$$\widehat{\sigma}^2_{MV} = \frac{y^T(I - X(X^TX)^-X^Ty)}{n}$$

El numerador de esta expresión corresponde a la suma de cuadrados del residual. Existe otro estimador que es insesgado y tiene la forma:

$$\widehat{\sigma}^2 = \frac{y^T(I - X(X^TX)^-X^T)y}{n - r(X)}$$

Esta última expresión corresponde al cuadrado medio del error que se reporta en el análisis de varianza de un modelo lineal. Como en la vida real se emplea un estimador

de la varianza del error, no se tiene una distribución normal sino una distribución t de Student, si $\lambda^T\beta$ es estimable, entonces:

$$\frac{\lambda^T\beta^0 - \lambda^T\beta}{\widehat{\sigma}^2\lambda^T(X^TX)^-\lambda} \sim \mathrm{t}(n - r(X))$$

Esto es, la variable aleatoria $\frac{\lambda^T\beta^0-\lambda^T\beta}{\widehat{\sigma}^2\lambda^T(X^TX)^-\lambda}$ sigue una distribución t de Student con $n - r(X)$ grados de libertad. El resultado también aplica en el caso del modelo de rango completo. Recordemos que bajo este escenario cualquier $\lambda^T\beta$ es estimable, simplemente se reemplazaría β^0 por $\widehat{\beta}$ y $(X^TX)^-$ por $(X^TX)^{-1}$.

3.4.4. Formulación bayesiana del modelo lineal

Consideramos la verosimilitud normal multivariada con vector de medias $X\beta$ y matriz de covarianzas $\sigma^2 I_n$ y presentamos tres distribuciones a priori diferentes.

A priori impropia: la distribución a priori es de la forma:

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Entonces:

$$\pi(\beta, \sigma^2 \mid y, X) \propto f(y \mid X, \beta, \sigma^2)\pi(\beta, \sigma^2).$$

La distribución posterior es propia si $n > p$, $p = r(X)$, es decir, el modelo es de rango completo, por lo tanto, nos enfocamos en este caso. Una vez se realiza este producto y se llevan a cabo algunas simplificaciones algebraicas, se obtiene:

$$\pi(\beta, \sigma^2 \mid y, X) \propto (\sigma^2)^{-\frac{n+2}{2}} \exp\left(\frac{1}{2\sigma^2}(\|y - \widehat{y}\|_2^2 + (\beta - \widehat{\beta})^T X^T X(\beta - \widehat{\beta}))\right)$$

Donde: $\widehat{\beta} = (X^TX)^{-1}X^Ty$, $\widehat{y} = X\widehat{\beta}$, $\|y - \widehat{y}\|_2^2 = (y - \widehat{y})^T(y - \widehat{y})$, el cuadrado de la norma Euclídea del vector de residuales $y - \widehat{y}$, que corresponde a la suma de cuadrados de los residuales. Aquí tenemos el núcleo de la densidad posterior conjunta de σ^2 y β, al integrar con respecto a β (marginalizar) podemos reconocer la expresión resultante como el núcleo de una densidad Gamma inversa, explícitamente:

$$\sigma^2 \mid y, X \sim \mathrm{GammaInv}\left(\frac{n - p}{2}, \frac{\|y - \widehat{y}\|_2^2}{2}\right)$$

La distribución Gamma inversa $\left(\frac{n-p}{2}, \frac{\|y-\hat{y}\|_2^2}{2}\right)$ es equivalente a una distribución chi-cuadrada escalada invertida con $n-p$ grados de libertad y parámetro de escala $S^2 = \frac{\|y-\hat{y}\|_2^2}{n-p}$. Al proceder de manera similar, es decir, integrando $\pi(\beta, \sigma^2 \mid y, X)$ con respecto a σ^2 (marginalizando), se puede encontrar que:

$$\beta \mid y, X \sim \mathrm{t}_p(n-r, \widehat{\beta}, S^2(X^TX)^{-1})$$

Esto es, una distribución t de Student multivariada con dimensión p, $n-p$ grados de libertad, parámetro de localización $\widehat{\beta}$ y matriz de escala $S^2(X^TX)^{-1}$. Así, siempre y cuando se cumpla con la condición $n > p$, el uso de esta a priori impropia nos lleva a una situación en la que las posteriores son tratables matemáticamente, y por lo tanto, se puede hacer inferencia exacta, porque los momentos de estas distribuciones y las condiciones bajo las que existen son conocidos. Por ejemplo, la media posterior de esta distribución, que sirve como estimador puntual del vector de parámetros de localización β existe si $n-p > 1$, pero debido a que tanto n como p son números naturales, esta condición equivale a $n \geq p+2$, esto es, el tamaño de muestra excede al número de parámetros de localización por dos o más unidades. Bajo esta condición, se tiene que:

$$E\left[\beta \mid y, X\right] = \widehat{\beta}.$$

Pero este vector también corresponde a la moda y a la mediana posteriores, y así, en virtud de lo expuesto en el Capítulo 10, se estaría minimizando el riesgo posterior respecto a las funciones de pérdida de error cuadrático medio, error absoluto medio y $0-1$; como también, se tiene un estimador máximo a posteriori (MAP).

Este ejemplo nos muestra algo muy interesante, si usamos $\widehat{\beta}$ como estimador puntual, tenemos un caso de equivalencia entre el estimador máximo-verosímil y un estimador bayesiano; sumado a esto, como el modelo es de rango completo se sigue que β es estimable y, en consecuencia, su media posterior es también MELI y EMCO. La matriz de covarianzas de β se puede inferir mediante la matriz de covarianzas posterior, la cual existe si $n-p > 2$, que equivale a $n \geq p+3$, es decir, el tamaño de muestra excede al número de parámetros de localización por tres o más unidades, si eso se satisface, entonces:

$$Var\left[\beta \mid y, X\right] = \frac{n-p}{n-p-2} S^2(X^TX)^{-1}$$

$$= \frac{\|y-\hat{y}\|_2^2}{(n-p-2)}(X^TX)^{-1}$$

Nótese que tenemos un estimador de la matriz de covarianza que de cierta manera es similar al utilizado en el caso frecuentista, en el que se usa un estimador de la varianza del error que al ser multiplicado por $(X^TX)^{-1}$ nos daba la matriz de covarianzas de $\widehat{\beta}$. No obstante, en el caso frecuentista no tiene sentido hablar de $Var\left[\beta\right]$ pues este vector es una constante (p-dimensional) desconocida. Recordemos que S^2 es el cuadrado medio del error proveniente del análisis de varianza, y que se presentó en la formulación frecuentista del modelo lineal general, así, la semejanza es más evidente.

Finalmente, la inferencia sobre el único parámetro de dispersión, σ^2, se puede llevar a cabo empleando alguna "medida" de tendencia central y otra de dispersión de su distribución posterior. En particular, la media posterior existe si $\frac{n-p}{2} > 1 \Leftrightarrow n \geq p+3$ y en caso tal:

$$E\left[\sigma^2 \mid y, X\right] = \frac{\frac{\|y-\widehat{y}\|_2^2}{2}}{\frac{n-p}{2}-1}$$

$$= \frac{\|y-\widehat{y}\|_2^2}{n-p-2}$$

Por otro lado, la moda posterior (estimador MAP) siempre existe y tiene la forma:

$$Mo\left[\sigma^2 \mid y, X\right] = \frac{\frac{\|y-\widehat{y}\|_2^2}{2}}{\frac{n-p}{2}+1}$$

$$= \frac{\|y-\widehat{y}\|_2^2}{n-p+2}$$

En tanto que, la varianza posterior existe si $\frac{n-p}{2} > 2 \Leftrightarrow n \geq p+5$ y tiene la forma:

$$Var\left[\sigma^2 \mid y, X\right] = \frac{\frac{(\|y-\widehat{y}\|_2^2)^2}{2}}{\left(\frac{n-p}{2}-1\right)^2\left(\frac{n-p}{2}-2\right)}$$

$$= \frac{2\|y-\widehat{y}\|_2^4}{(n-p-2)^2(n-p-4)}$$

Cabe anotar que la matriz de covarianzas posterior de β es igual al producto de la media posterior de la varianza del error y $(X^TX)^{-1}$.

A priori semi-conjugadas: ahora empleamos una distribución a priori diferente bajo la misma verosimilitud. Estas distribuciones a priori se conocen como semi-conjugadas, no todos los autores coinciden en el uso que se le da al término, pero en este texto, una distribución a priori es semi-conjungada para una verosimilitud dada si la distribución condicional completa del parámetro (en lugar de la posterior) pertenece a la misma familia.

$$\beta \sim \mathrm{N_p}(\beta_0, \Sigma_0)$$

$$\frac{1}{\sigma^2} := \gamma \sim \mathrm{Gamma}(\frac{\upsilon_0}{2}, \frac{\upsilon_0 \sigma_0^2}{2})$$

Donde β_0 es el vector p-dimensional que corresponde a la media a priori de β, Σ_0 es una matriz definida positiva que corresponde a la matriz de covarianzas a priori de β, $\frac{1}{\sigma^2}$ corresponde a la precisión, $\frac{\upsilon_0}{2}$ y $\frac{\upsilon_0 \sigma_0^2}{2}$ son los parámetros de forma y de tasa de la distribución Gamma, $\upsilon_0 > 0$, $\sigma_0^2 > 0$. Asignar una distribución Gamma a la precisión es equivalente a asignar una Gamma inversa a la varianza del error. La razón para incluir el denominador 2 en ambos parámetros es la simplificación de las operaciones algebraicas que se llevan a cabo para encontrar las condicionales completas, que bajo este modelo tienen la forma:

$$\beta \mid y, X, \sigma^2 \sim \mathrm{N_p}\left(\left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{X^T y}{\sigma^2}\right), \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}\right)$$

$$\sigma^2 \mid y, X, \beta \sim \mathrm{GammaInv}\left(\frac{(\upsilon_0 + n)}{2}, \frac{(\upsilon_0 \sigma_0^2 + SCE(\beta))}{2}\right)$$

Donde $SCE(\beta) = (y - X\beta)^T(y - X\beta)$. Nótese que en este modelo no se requiere la inversa de $X^T X$, sino de $\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}$, la cual, en vista del Teorema 8.1, siempre existe puesto que: Σ_0 es definida positiva y, por lo tanto, su inversa también lo es, así, este modelo puede ser empleado para el caso de rango incompleto.

Como las condicionales completas son conocidas, se puede implementar un muestreador de Gibbs (Sección 10.10.1) en el cual, la iteración k sería de la siguiente forma:

a) Calcule $V^{(k)} = \left(\Sigma_0^{-1} + \frac{X^T X}{(\sigma^2)^{(k-1)}}\right)^{-1}$, $m^{(k)} = V^{(k)}\left(\Sigma_0^{-1}\beta_0 + \frac{X^T y}{(\sigma^2)^{(k-1)}}\right)$

b) Muestree $\beta^{(k)}$ de una distribución $\mathrm{N_p}(m^{(k)}, V^{(k)})$

c) Calcule $SCE(\beta^{(k)}) = (y - X\beta^{(k)})^T(y - X\beta^{(k)})$

d) Muestree $(\sigma^2)^{(k)}$ de una distribución GammaInv $\left(\frac{\upsilon_0+n}{2}, \frac{(\upsilon_0\sigma_0^2+SCE(\beta^{(k)})}{2}\right)$.

Donde $(\sigma^2)^{(k-1)}$ es el valor de la varianza del error muestreado en la iteración $k-1$.

No siempre es fácil determinar los valores de los hiperparámetros, que en el ejemplo anterior son: $\beta_0, \Sigma_0, \upsilon_0$ y σ_0^2. Una alternativa para expresar la incertidumbre que se tenga sobre los mismos es emplear un modelo jerárquico y asignar distribuciones a cada uno de estos; por otro lado, se pueden considerar distribuciones a priori como la que se estudia a continuación.

A priori débilmente informativas: estudiaremos la propuesta de Kass y Wasserman [13], la cual está fundamentada en la idea de utilizar una distribución a priori que contenga la información aportada por un solo registro. Estos autores recomiendan la siguiente distribución a priori para el modelo de rango completo:

$$\beta|\sigma^2 \sim \mathrm{N_p}\left(\widehat{\beta}, \frac{X^T X}{n\sigma^2}\right)$$

$$\frac{1}{\sigma^2} := \gamma \sim \mathrm{Gamma}\left(\frac{1}{2}, \frac{\widehat{\sigma}^2}{2}\right)$$

Donde $\widehat{\beta}$ es el EMCO de β y $\widehat{\sigma}^2$ el estimador insesgado de la varianza del error. Una crítica que puede recibir este tipo de distribución a priori es que requiere conocer los datos para construirla, por lo tanto, podría decirse que no es una a priori en sentido estricto. Nótese que las condicionales completas y, por ende, el muestreador de Gibbs sería el mismo que se presentó en la sección anterior al ver que en este caso:

$$\beta_0 = \widehat{\beta}$$

$$\Sigma_0 = \frac{X^T X}{n\sigma^2}$$

$$\upsilon_0 = 1$$

$$\sigma_0^2 = \hat{\sigma}^2$$

Para cerrar el planteamiento bayesiano del modelo lineal, vale la pena mencionar que existen distribuciones a priori que también dependen de los datos, pero que tienen algunas ventajas a nivel computacional. Una de estas es la a priori g de Zellner (Zellner, 1986), que permite obtener muestras directas de la posterior conjunta de β y σ^2 y por consiguiente se usan métodos de Monte Carlo y no métodos MCMC, lo cual implica un ahorro en esfuerzo computacional. Además, esta distribución a priori se obtiene bajo el principio de invarianza a la escala de los regresores, lo cual se convierte en otra de sus propiedades.

3.5. Modelos lineales mixtos

En esta sección se presenta la extensión del modelo lineal general que considera un conjunto de parámetros que son variables aleatorias y por ende se les asigna una distribución de probabilidad, se trata del modelo lineal mixto, que tiene múltiples aplicaciones y resulta de particular interés en mejoramiento genético, debido a que el modelo animal, que en realidad es una familia de modelos, corresponde a un modelo lineal mixto.

En notación matricial el modelo lineal mixto es como sigue:

$$y = X\beta + Zu + e$$

Donde y es el vector aleatorio de dimensión nx1 que contiene los registros, β es el vector de efectos fijos de dimensión kx1, u es el vector de efectos aleatorios de dimensión px1, e es el vector aleatorio de errores de dimensión nx1, X de dimensión nxk y Z de dimensión nxp son matrices relacionan a los vectores β y u con el vector de registros, respectivamente. Los supuestos de distribución son:

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim \mathcal{NMV}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right]$$

Lo cual implica que:

$$y \mid u \sim \mathcal{NMV}(X\beta + Zu, R)$$

$$y \sim \mathcal{NMV}(X\beta, ZGZ^T + R)$$

Esta notación nos dice que u y e siguen una distribución normal multivariado con vectores de medias nulos y matrices de covarianza G y R, que dados los efectos aleatorios, el vector y se distribuye normal multivariado con media $X\beta + Zu$ y matriz de covarianzas R, mientras que marginalmente y sigue una distribución normal multivariada con esperanza $X\beta$ y matriz de covarianzas $V = ZGZ^T + R$.

Recordemos que $f(y,u) = f(y|u)f(u)$, la regla del producto estudiada en el capítulo 9. Henderson maximizó $f(y,u)$ con respecto a u y β para obtener un sistema de ecuaciones lineales conocido como las ecuaciones de modelos mixtos (EMM), cuya solución permite estimar β y predecir u. Estas ecuaciones tienen la siguiente forma:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \widehat{u} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix}$$

Es un sistema de ecuaciones lineales en los que la matriz de coeficientes es:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{bmatrix}$$

La cual es simétrica. El vector de coeficientes es:

$$\begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix}$$

El vector de incógnitas es:

$$\begin{bmatrix} \beta^0 \\ \widehat{u} \end{bmatrix}$$

El sistema es soluble y tendrá solución única cuando la matriz de coeficientes es de rango completo e infinitas soluciones en el caso contrario. Hasta el momento se ha empleado una notación general en la que las matrices de covarianza de efectos aleatorios y errores pueden tener cualquier estructura.

Cuando se encontraron las EMM se descubrió una forma de estimar efectos fijos y predecir efectos aleatorios en un modelo lineal mixto, se sabía que la solución de efectos fijos era una solución a las ecuaciones normales MCG, es decir:

$$\beta^0 = (X^TV^{-1}X)^-X^TV^{-1}y$$

Bajo normalidad, también es el estimador de máxima verosimilitud, pero se desconocían las propiedades del predictor. Un resultado muy importante fue el encontrado por Henderson [14], quién mostró que la solución de las EMM para $\widehat{u}$ corresponde al MPLI(u), que tiene la forma:

$$MPLI(u) = GZ^TV^{-1}(y - X\beta^0)$$

Esta equivalencia puede hacerse usando el hecho de que bajo los supuestos presentados anteriormente:

$$V^{-1} = R^{-1} - R^{-1}Z(Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1}$$

Esta es la identidad de Woodbury que fue introducida a la estadística por Henderson (Ghosh, comunicación personal), convirtiéndose así en otra de sus grandes contribuciones.

Ahora se extiende el concepto de función estimable al caso del modelo lineal mixto. Un conjunto de combinaciones lineales del vector de parámetros $\begin{bmatrix} \beta \\ u \end{bmatrix}$ de la forma:

$$L^T \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} L_1^T & L_2^T \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix}$$

$$= L_1^T \beta + L_2^T u$$

Se dice predecible si $L_1^T \beta$ es estimable. Aquí, L_1^T es una matriz de dimension lxk que contiene los coeficientes de cada una de las l funciones lineales asociados a los efectos fijos, mientras que L_2^T es una matriz de dimensión lxp que contiene los coeficientes de cada una de las l funciones lineales asociados a los efectos aleatorios.

En vista de lo expuesto en el capítulo 2, en este caso también se puede hablar de función estimable puesto que también se puede hablar de estimar efectos aleatorios. La razón para que la estimabilidad/predictibilidad de $L^T \begin{bmatrix} \beta \\ u \end{bmatrix}$ dependa de la estimabilidad de $L_1^T \beta$ es que la submatriz $Z^T R^{-1} Z + G^{-1}$ de la matriz de coeficientes de las EMM, es de rango completo debido a que:

a) Existe G^{-1} y, por consiguiente, G y su inversa son matrices simétricas definidas positivas

b) $Z^T R^{-1} Z$ es una matriz simétrica real (Teorema 8.1)

Por lo tanto, cualquier combinación lineal de u es siempre estimable/predecible.

En el modelo lineal mixto, el vector u es aleatorio y se asume que su matriz de covarianzas es G; el vector $u - \hat{u}$ se conoce como el vector de errores de predicción, que contiene las diferencias entre predictando y predictor. Los elementos de la diagonal principal de la matriz de covarianzas de este vector, es decir $Var\left[u - \hat{u}\right]$, contienen las varianzas del error de predicción. Por otro lado, como el vector β es una constante multidimensional desconocida, su matriz de covarianza es nula al igual que su matriz de covarianza con β^0 y por ende $Var\left[\beta^0 - \beta\right] = Var\left[\beta^0\right]$, así, la matriz de covarianza

del error de estimación es igual a la matriz de covarianza del estimador. Consideremos la siguiente notación para los bloques de la inversa generalizada del lado izquierdo de las EMM:

$$\begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} := \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^-$$

Es decir, C^{11} es el bloque correspondiente a $X^T R^{-1} X$ en la inversa generalizada, C^{12} el correspondiente a $X^T R^{-1} Z$, C^{21} el correspondiente a $Z^T R^{-1} X$ y C^{22} el correspondiente a $Z^T R^{-1} Z + G^{-1}$. Pese a que la matriz original es simétrica, cuando esta es singular, no existe garantía de que se use una inversa generalizada simétrica.

Análogo al caso del modelo lineal general, tenemos el siguiente resultado:

$$Var \begin{bmatrix} \beta^0 - \beta \\ u - \hat{u} \end{bmatrix} = Var \begin{bmatrix} \beta^0 \\ u - \hat{u} \end{bmatrix} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

Cuando la matriz del lado izquierdo de las EMM no es de rango columna completo, se tiene interés sobre la matriz de covarianzas de un conjunto de funciones predecibles, en este caso:

$$Var \left[\begin{bmatrix} L_1^T & L_2^T \end{bmatrix} \begin{bmatrix} \beta^0 \\ u - \hat{u} \end{bmatrix} \right] = \begin{bmatrix} L_1^T & L_2^T \end{bmatrix} \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$$

Ahora bien, los parámetros de dispersión (todos aquellos asociados a las matrices de covarianza), son desconocidos; por lo tanto, una vez se estiman se obtienen $\hat{R}^{-1}$ y $\hat{G}^{-1}$ y la matriz de coeficientes de las EMM se construye así:

$$\hat{C} = \begin{pmatrix} X^T \hat{R}^{-1} X & X^T \hat{R}^{-1} Z \\ Z^T \hat{R}^{-1} X & Z^T \hat{R}^{-1} Z + \hat{G}^{-1} \end{pmatrix}$$

Existen dos métodos frecuentistas bastante usados en la estimación de componentes de varianza y covarianza del modelo lineal mixto, máxima verosimilitud y máxima verosimilitud restringida. A continuación, se presentan algunas generalidades sobre los mismos.

En el método de máxima verosimilitud se procede de forma similar a lo que se hizo en el caso de los parámetros de localización del modelo lineal general. Se parte del supuesto de normalidad multivariada de los datos, pero nos enfocamos solamente en la parte asociada a la matriz de covarianzas de los datos V. La función objetivo, es decir, la que se quiere optimizar se puede escribir así [10]:

$$l(G, R) = -\frac{1}{2} \ln(V) - \frac{1}{2} \delta^T V^{-1} \delta - \frac{n}{2} \ln(2\pi)$$

Por otro lado, la máxima verosimilitud restringida o residual [15], también conocida como REML por sus siglas en inglés, es un método en el que se busca una invarianza a

los efectos fijos del modelo y usa una función de los datos que podría verse como una transformación de los mismos. En este caso, la función objetivo es [10]:

$$l_{Res}(G,R) = -\frac{1}{2}\ln(V) - \frac{1}{2}\ln \mid X^T V^{-1} X \mid -\frac{1}{2}\delta^T V^{-1}\delta - \frac{n-r}{2}\ln(2\pi)$$

Donde $\delta = y - X(X^T V^{-1} X)^{-} X^1 V^{-1} y$; $r = r(X)$. La dependencia sobre G y R viene dada por $V = ZGZ^T + R$. Conceptualmente, REML tiene la ventaja de tener en cuenta la pérdida de grados de libertad por la estimación del vector de efectos fijos β. En la práctica, este es el método más empleado para estimar componentes de varianza en modelos lineales mixtos. En general, las funciones objetivo $l(G,R)$ y $l_{Res}(G,R)$ no pueden optimizarse de forma exacta; por lo tanto, se usan métodos como el de Newton-Raphson para encontrar soluciones numéricas.

Como se indicó en el capítulo 2, cuando se trabaja con el MPLI, se asumen que los componentes de varianza son conocidos, sin embargo, no lo son y se emplean sus valores estimados para construir las EMM; por tanto, se habla de MPLI empírico (MPLIE), o EBLUP por sus siglas en inglés (empirical BLUP). Sin embargo, no es frecuente que en la literatura se haga esta aclaración, en las evaluaciones genéticas de rutina y en los trabajos científicos se está usando el MPLIE, pero se sigue hablando de MPLI, lo cual no representa un problema mayor. No obstante, vale la pena aclarar que el MPLIE no tiene todas las propiedades del MPLI, por ejemplo, ya no es lineal en los datos, pero bajo ciertas condiciones sigue siendo insesgado [16, 17].

Por último, vale la pena resaltar que el MPLI puede verse como un estimador bayesiano. Por ejemplo, este puede derivarse siguiendo una aproximación conocida como bayes empírico, en el cual se usa la distribución de los datos dados los hiperparámetros para estimar estos últimos. Además, las EMM también se pueden obtener desde una perspectiva bayesiana, al maximizar la distribución posterior de u y β bajo distribuciones a priori normales multivariadas e independientes para u y e, y una a priori difusa para β.

En el ámbito bayesiano puede resultar un poco confuso hablar de modelos lineales mixtos, debido a que todos los parámetros se tratan como variables aleatorias, así, las propiedades probabilísticas de los efectos fijos (no tienen varianza) que se estudiaron en la primera parte de este capítulo no aplican en un modelo lineal bayesiano. Existen dos nociones que permiten separar los parámetros del modelo en dos grupos: “fijos” y aleatorios. Una de estas clasifica como efectos “fijos” a aquellos a los cuales se asignan distribuciones a priori impropias y como aleatorios a aquellos a los que se asignan distribuciones a priori propias. La segunda se da en el contexto de variables explicativas cualitativas que definen grupos, se denominan efectos “fijos” a aquellos que son constantes a través de grupos, mientras que aquellos específicos para cada grupo son aleatorios.

La especificación del modelo es muy similar a la del modelo lineal general, pero los parámetros de localización se dividen en “fijos” y aleatorios, así, podemos considerar tres grupos de parámetros, efectos “fijos”, efectos aleatorios y parámetros de dispersión.

3.6. Ejercicios en R-project

EJEMPLO 3.2.3 de 9 animales puros y cruzados de las razas A, B, C y D:

```
BD=data.frame(matrix(ncol=7,byrow=TRUE, c(
1,"A ",0,1.0,0.0,0.0,0.0,
2,"AB",1,0.5,0.5,0.0,0.0,
3,"AB",1,0.5,0.5,0.0,0.0,
4,"AC",1,0.5,0.0,0.5,0.0,
5,"AD",1,0.5,0.0,0.0,0.5,
6,"AD",1,0.5,0.0,0.0,0.5,
7,"AD",1,0.5,0.0,0.0,0.5,
8,"A ",0,1.0,0.0,0.0,0.0,
9,"A ",0,1.0,0.0,0.0,0.0)))
colnames(BD)=c("id","GG","Hetero",
               "Raza_A","Raza_B","Raza_C","Raza_D")
```

Montaje de matrices:

```
X=matrix(ncol=6,as.numeric(cbind(rep(1,9),
          BD$Hetero,
          BD$Raza_A,BD$Raza_B,BD$Raza_C,BD$Raza_D)))
colnames(X)=c("Media","Hetero","Raza_A",
              "Raza_B","Raza_C","Raza_D")
```

```
X
```

```
##       Media Hetero Raza_A Raza_B Raza_C Raza_D
##  [1,]     1      0    1.0    0.0    0.0    0.0
##  [2,]     1      1    0.5    0.5    0.0    0.0
##  [3,]     1      1    0.5    0.5    0.0    0.0
##  [4,]     1      1    0.5    0.0    0.5    0.0
##  [5,]     1      1    0.5    0.0    0.0    0.5
##  [6,]     1      1    0.5    0.0    0.0    0.5
##  [7,]     1      1    0.5    0.0    0.0    0.5
##  [8,]     1      0    1.0    0.0    0.0    0.0
##  [9,]     1      0    1.0    0.0    0.0    0.0
```

```
matrix(ncol=1,apply(X[,3:6],1,sum))
```

```
##       [,1]
##  [1,]    1
##  [2,]    1
```

```
##   [3,]    1
##   [4,]    1
##   [5,]    1
##   [6,]    1
##   [7,]    1
##   [8,]    1
##   [9,]    1
```

```
matrix(ncol=1,2*apply(X[,4:6],1,sum))
```

```
##        [,1]
##   [1,]    0
##   [2,]    1
##   [3,]    1
##   [4,]    1
##   [5,]    1
##   [6,]    1
##   [7,]    1
##   [8,]    0
##   [9,]    0
```

```
matrix(ncol=1,X[,2])
```

```
##        [,1]
##   [1,]    0
##   [2,]    1
##   [3,]    1
##   [4,]    1
##   [5,]    1
##   [6,]    1
##   [7,]    1
##   [8,]    0
##   [9,]    0
```

```
XpX=t(X)%*%X
XpX
```

```
##         Media Hetero Raza_A Raza_B Raza_C Raza_D
## Media     9.0    6.0   6.00    1.0   0.50   1.50
## Hetero    6.0    6.0   3.00    1.0   0.50   1.50
## Raza_A    6.0    3.0   4.50    0.5   0.25   0.75
## Raza_B    1.0    1.0   0.50    0.5   0.00   0.00
## Raza_C    0.5    0.5   0.25    0.0   0.25   0.00
## Raza_D    1.5    1.5   0.75    0.0   0.00   0.75
```

EJEMPLO 3.4.2 de 9 animales puros y cruzados de varias razas vacunas:

```
BD=data.frame(matrix(ncol=12,byrow=TRUE, c(
1,"Bra",1.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
2,"Bra",1.0,1.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
3,"Guz",1.0,1.0,0.5,0.5,0.0,0.0,0.0,0.0,0.0,0.0,
4,"BlO",1.0,1.0,0.5,0.0,0.5,0.0,0.0,0.0,0.0,0.0,
5,"Brv",1.0,1.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0,0.0,
6,"Lim",1.0,1.0,0.5,0.0,0.0,0.0,0.5,0.0,0.0,0.0,
7,"Nor",1.0,1.0,0.5,0.0,0.0,0.0,0.0,0.5,0.0,0.0,
8,"Rom",1.0,1.0,0.5,0.0,0.0,0.0,0.0,0.0,0.5,0.0,
9,"Sim",1.0,1.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.5
  )))
colnames(BD)=c(
"caso","GG", "Int", "H", "Bra", "Guz", "BlO",
"Bra", "Lim", "Nor", "Rom","Sim")
```

Para el montaje de matrices utilizaremos la librería ≪MASS≫ [18] para calcular la inversa:

```
X=as.matrix(BD[,3:12])
X=matrix(ncol=10,c(as.numeric(X)))
colnames(X)=c(
"Int", "H", "Bra", "Guz",
"BlO", "Bra", "Lim",
"Nor", "Rom","Sim")
X
```

```
##       Int H Bra Guz BlO Bra Lim Nor Rom Sim
##  [1,]   1 0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
##  [2,]   1 1 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
##  [3,]   1 1 0.5 0.5 0.0 0.0 0.0 0.0 0.0 0.0
##  [4,]   1 1 0.5 0.0 0.5 0.0 0.0 0.0 0.0 0.0
##  [5,]   1 1 0.5 0.0 0.0 0.5 0.0 0.0 0.0 0.0
##  [6,]   1 1 0.5 0.0 0.0 0.0 0.5 0.0 0.0 0.0
##  [7,]   1 1 0.5 0.0 0.0 0.0 0.0 0.5 0.0 0.0
##  [8,]   1 1 0.5 0.0 0.0 0.0 0.0 0.0 0.5 0.0
##  [9,]   1 1 0.5 0.0 0.0 0.0 0.0 0.0 0.0 0.5
```

```
matrix(ncol=1,apply(X[,3:10],1,sum))
```

```
##       [,1]
##  [1,]    1
##  [2,]    1
##  [3,]    1
##  [4,]    1
##  [5,]    1
##  [6,]    1
##  [7,]    1
##  [8,]    1
##  [9,]    1
```

```
XpX=t(X)%*%X
```

```
XpX
```

```
##     Int   H  Bra  Guz  BlO  Bra  Lim  Nor  Rom  Sim
## Int 9.0 8.0 5.50 0.50 0.50 0.50 0.50 0.50 0.50 0.50
## H   8.0 8.0 4.50 0.50 0.50 0.50 0.50 0.50 0.50 0.50
## Bra 5.5 4.5 3.75 0.25 0.25 0.25 0.25 0.25 0.25 0.25
## Guz 0.5 0.5 0.25 0.25 0.00 0.00 0.00 0.00 0.00 0.00
## BlO 0.5 0.5 0.25 0.00 0.25 0.00 0.00 0.00 0.00 0.00
## Bra 0.5 0.5 0.25 0.00 0.00 0.25 0.00 0.00 0.00 0.00
## Lim 0.5 0.5 0.25 0.00 0.00 0.00 0.25 0.00 0.00 0.00
## Nor 0.5 0.5 0.25 0.00 0.00 0.00 0.00 0.25 0.00 0.00
## Rom 0.5 0.5 0.25 0.00 0.00 0.00 0.00 0.00 0.25 0.00
## Sim 0.5 0.5 0.25 0.00 0.00 0.00 0.00 0.00 0.00 0.25
```

```
library(MASS)
XpXin=(round(ginv(XpX),1))
```

```
XpXin
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]  3.6 -2.4 -2.7  0.9  0.9  0.9  0.9  0.9  0.9   0.9
##  [2,] -2.4  2.0  1.4 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6  -0.6
##  [3,] -2.7  1.4  2.8 -0.8 -0.8 -0.8 -0.8 -0.8 -0.8  -0.8
##  [4,]  0.9 -0.6 -0.8  3.7 -0.3 -0.3 -0.3 -0.3 -0.3  -0.3
##  [5,]  0.9 -0.6 -0.8 -0.3  3.7 -0.3 -0.3 -0.3 -0.3  -0.3
##  [6,]  0.9 -0.6 -0.8 -0.3 -0.3  3.7 -0.3 -0.3 -0.3  -0.3
```

```
##  [7,]  0.9 -0.6 -0.8 -0.3 -0.3 -0.3  3.7 -0.3 -0.3  -0.3
##  [8,]  0.9 -0.6 -0.8 -0.3 -0.3 -0.3 -0.3  3.7 -0.3  -0.3
##  [9,]  0.9 -0.6 -0.8 -0.3 -0.3 -0.3 -0.3 -0.3  3.7  -0.3
## [10,]  0.9 -0.6 -0.8 -0.3 -0.3 -0.3 -0.3 -0.3 -0.3   3.7
```

Estimabilidad de un conjunto de 8 funciones lineales paramétricas:

```
Kp=matrix(ncol=10, byrow=TRUE, c(
0,1, 0,0,0,0,0,0,0,0,
0,0,-1,1,0,0,0,0,0,0,
0,0,-1,0,1,0,0,0,0,0,
0,0,-1,0,0,1,0,0,0,0,
0,0,-1,0,0,0,1,0,0,0,
0,0,-1,0,0,0,0,1,0,0,
0,0,-1,0,0,0,0,0,1,0,
0,0,-1,0,0,0,0,0,0,1))
```

```
Kp
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    1    0    0    0    0    0    0    0     0
## [2,]    0    0   -1    1    0    0    0    0    0     0
## [3,]    0    0   -1    0    1    0    0    0    0     0
## [4,]    0    0   -1    0    0    1    0    0    0     0
## [5,]    0    0   -1    0    0    0    1    0    0     0
## [6,]    0    0   -1    0    0    0    0    1    0     0
## [7,]    0    0   -1    0    0    0    0    0    1     0
## [8,]    0    0   -1    0    0    0    0    0    0     1
```

```
t(Kp)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
##  [1,]    0    0    0    0    0    0    0    0
##  [2,]    1    0    0    0    0    0    0    0
##  [3,]    0   -1   -1   -1   -1   -1   -1   -1
##  [4,]    0    1    0    0    0    0    0    0
##  [5,]    0    0    1    0    0    0    0    0
##  [6,]    0    0    0    1    0    0    0    0
##  [7,]    0    0    0    0    1    0    0    0
##  [8,]    0    0    0    0    0    1    0    0
##  [9,]    0    0    0    0    0    0    1    0
## [10,]    0    0    0    0    0    0    0    1
```

```
cheq=round(Kp%*%ginv(XpX)%*%(XpX),1)
cheq


##      Int H Bra Guz BlO Bra Lim Nor Rom Sim
## [1,]   0 1   0   0   0   0   0   0   0   0
## [2,]   0 0  -1   1   0   0   0   0   0   0
## [3,]   0 0  -1   0   1   0   0   0   0   0
## [4,]   0 0  -1   0   0   1   0   0   0   0
## [5,]   0 0  -1   0   0   0   1   0   0   0
## [6,]   0 0  -1   0   0   0   0   1   0   0
## [7,]   0 0  -1   0   0   0   0   0   1   0
## [8,]   0 0  -1   0   0   0   0   0   0   1
```

4 CAPÍTULO CUATRO

Modelo animal

Carlos Eugenio Solarte Portilla
Universidad de Nariño

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

4.1. Generalidades

Para explicar los aspectos teóricos más importantes, lo mismo que los procedimientos y operaciones con matrices correspondientes al modelo animal, se puede consultar un abundante material bibliográfico, aunque son de especial importancia las publicaciones de Henderson [14, 19], Kennedy [20] y Mrode [21].

El término modelo animal hace alusión a una familia de modelos estadísticos empleados para predecir efectos genéticos para uno o más fenotipos de interés, bajo diferentes tipos de acción génica (aditiva, no aditiva, directa, materna), estructuras de datos (una sola observación o varias observaciones por animal) y teniendo en cuenta diferentes tipos de efectos ambientales dependiendo del manejo zootécnico que se tenga. También se emplean para estimar parámetros genéticos, y predecir registros futuros.

Estos modelos pueden agruparse en una misma familia porque son modelos lineales mixtos (ver sección 3.5) en los cuales los efectos genéticos aleatorios tienen matrices de covarianza que se construyen a partir de principios de genética cuantitativa; por ejemplo, la matriz de covarianzas de los efectos genéticos aditivos directos emplea la relación genética aditiva entre individuos, que corresponde al numerador del coeficiente de parentesco de Wright y se construye a partir del pedigrí.

Antes de entrar en materia, es importante aclarar que a través del texto se tratarán variables para las cuales es apropiado emplear un modelo lineal mixto, como aquellas variables continuas con distribución normal o transformaciones de variables discretas como lo es el logaritmo del recuento de células somáticas de la leche.

En el capítulo 3.5 estudiamos el modelo lineal mixto y algunas de sus propiedades, aquí nos enfocamos en el caso más simple del modelo animal, al cual nos referimos como el modelo básico o modelo animal básico. Este modelo se emplea en las evaluaciones genéticas para características que se miden una sola vez durante toda la vida productiva de los animales. El modelo puede incluir efectos fijos de tipo clasificatorio (ej: finca, sexo, año, nivel de alimentación, región, sistema de producción) o covariables (ej: edad, talla) y en lo referente a efectos genéticos, solo considera efectos aditivos directos, además, estos son los únicos efectos aleatorios.

En este modelo, la matriz covarianzas de los efectos genéticos aditivos directos es proporcional a la matriz de relaciones genéticas aditivas o matriz de parentesco. El modelo en representación matricial es:

$$y = X\beta + Za + e$$

Donde: y es un vector de dimensión nx1 que contiene n registros de la variable a evaluar.

β es el vector de los efectos fijos de dimensión px1.

a es el vector de dimensión qx1 que contiene los valores genéticos aditivos directos de cada animal, donde q corresponde al número de individuos en la matriz de parentesco.

e es el vector de dimensión nx1, que contiene errores aleatorios.

X es una matriz de orden nxp, que relaciona los registros con los efectos fijos.

Z es una matriz de orden nxq, que relaciona los registros con los individuos.

Tanto X como Z se denominan matrices de diseño.

En este modelo se asume que los vectores aleatorios a y e siguen distribuciones normales multivariadas independientes, con valores esperados $E(a) = E(e) = 0$ y matrices de covarianza $Var[a] = G = \sigma_a^2 A$ y $Var[e] = \sigma_e^2 I_n$.

Esto se puede escribir matricialmente así:

$$\begin{bmatrix} a \\ e \end{bmatrix} \sim \mathcal{NMV}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 A & 0 \\ 0 & \sigma_e^2 I_n \end{bmatrix}\right]$$

Donde $\mathcal{NMV}$ indica la distribución normal multivariada. Bajo estos supuestos probabilísticos tenemos la siguiente distribución del vector de registros

$$y \sim \mathcal{NMV}(X\beta, \sigma_a^2 ZAZ^T + \sigma_e^2 I_n)$$

Una propiedad importante del modelo animal es la descomposición del MPLI del valor genético aditivo en tres componentes. Recordemos que en este caso el MPLI del vector a tiene la forma:

$$\widehat{a} = GZ^T V^{-1}(y - X\beta^0)$$

Entonces, la solución para el i-ésimo individuo es el producto punto entre la i-ésima fila de la matriz $GZ^T V^{-1}$ y el vector de registros corregidos por efectos fijos $y - X\beta^0$. A su vez, vale notar que la i-ésima fila de $GZ^T V^{-1}$ es un vector resultante del producto de la i-ésima fila de G y la matriz $Z^T V^{-1}$, y $G = \sigma_a^2 A$. Por ende, la i-ésima fila de esta matriz contiene las relaciones genéticas aditivas del i-ésimo animal con todos los demás individuos evaluados y $1 + F_i$ en la posición i, multiplicados por la varianza genética aditiva, donde F_i es el coeficiente de endogámia. Manipulando estas expresiones se puede encontrar que el MPLI del valor genético aditivo directo del individuo i, denotado como $\widehat{a}_i$ tiene la forma:

$$\widehat{a}_i = w_{1i}\widehat{PP}_i + w_{2i}\widehat{PA}_i + w_{3i}\widehat{CP}_i$$

Donde $\widehat{PP}_i$ es el promedio de los MPLI de los padres del individuo i, $\widehat{PA}_i$ son los registros del animal ajustados por efectos fijos, $\widehat{CP}_i$ es la contribución de la progenie, que corresponde a una combinación lineal de los valores genéticos aditivos directos de la progenie corregidos por la mitad del valor genético de su otro parental, esto es, la HPT o DEP del otro parental. Las formas funcionales de estos pesos y de $\widehat{CP}_i$ pueden consultarse en VanRaden y Wiggans [22]. Los pesos en esta combinación lineal satisfacen la siguiente propiedad:

$$w_1 + w_2 + w_3 = 1,\ w_i \geq 0, i = 1, 2, 3$$

Esto le confiere la propiedad de ser una combinación convexa.

La anterior descomposición del valor genético aditivo predicho cobra importancia a la hora de comprender su origen, pues nos muestra las tres piezas de información de las que este se compone: la proveniente de los padres, la que se obtiene de los registros propios del individuo y la que procede de la progenie tras corregir por la HPT o DEP del otro progenitor. Así, en ausencia de una o dos de estas fuentes, el valor genético predicho dependerá de las o la restante. Por ello, cuando se tienen animales sin progenie ni registros propios, como suele ocurrir en el caso de individuos muy jóvenes, el valor genético aditivo se predice como el promedio de los padres y, por consiguiente, hermanos completos tendrán la misma predicción.

En el capítulo 7 se presenta la selección genómica, la cual permite tener en cuenta información del genoma de cada individuo y en este escenario, genera predicciones diferentes para hermanos completos (a pesar de que no tengan registros propios ni progenie). Además, esta forma de expresar el valor genético aditivo predicho muestra que, animales sin registros pueden ser evaluados, siempre y cuando sus padres se conozcan o tengan progenies. Esto puede ocurrir porque no se observó el fenotipo del animal, este es muy joven y aún no es posible medirlo, o porque biológicamente no se puede medir; un ejemplo de la última situación, es la producción de leche en machos.

El hecho de que los únicos ancestros que aparecen sean los padres y que no haya aporte de los colaterales, se debe a un argumento de independencia condicional, este dice que, dados los valores genéticos de los padres, los demás ancestros y los colaterales no aportan información al valor genético del animal i, pues ninguno de ellos brinda información de manera directa, sino que lo hacen mediante los padres de i, un hecho que se advierte fácilmente mirando el pedigrí. En el argot de una clase de modelos estadísticos que representan suposiciones de independencia probabilística mediante grafos y se conocen como modelos gráficos, esta característica se conoce como propiedad de Markov dirigida, resulta que, un pedigrí es un tipo de grafo conocido como grafo dirigido acíclico.

Como se vio en el capítulo 3, las soluciones para a y β se obtienen a partir de las EMM, que bajo las formas particulares de R y G asumidas en el modelo animal básico, tienen la siguiente forma:

$$\begin{bmatrix} X^T\sigma_e^{-2}I_nX & X^T\sigma_e^{-2}I_nZ \\ Z^T\sigma_e^{-2}I_nX & Z^T\sigma_e^{-2}I_nZ + \sigma_a^{-2}A^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T\sigma_e^{-2}I_ny \\ Z^T\sigma_e^{-2}I_ny \end{bmatrix} \tag{4.1}$$

$$\Rightarrow \sigma_e^{-2} \begin{bmatrix} X^TI_nX & X^TI_nZ \\ Z^TI_nX & Z^TI_nZ + \frac{\sigma_e^2}{\sigma_a^2}A^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{a} \end{bmatrix} = \sigma_e^{-2} \begin{bmatrix} X^TI_ny \\ Z^TI_ny \end{bmatrix}$$

El inverso multiplicativo de la varianza del error se cancela dando origen al siguiente sistema de ecuaciones:

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix} \tag{4.2}$$

Donde: A = matriz de parentesco o matriz de relaciones genéticas aditivas (Numerator Relationship Matrix, NRM, en inglés) y $\alpha = \frac{\sigma_e^2}{\sigma_a^2}$

Al recordar que la heredabilidad en sentido estricto tiene la forma:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

Se advierte que α puede escribirse como función de la heredabilidad, en efecto:

$$\frac{1}{h^2} = \frac{\sigma_a^2 + \sigma_e^2}{\sigma_a^2}$$

$$= 1 + \frac{\sigma_e^2}{\sigma_a^2}$$

$$= 1 + \alpha$$

$$\Rightarrow \alpha = \frac{1}{h^2} - 1$$

$$= \frac{1 - h^2}{h^2}$$

Esta identidad resulta de gran utilidad por la siguiente razón. Nótese que α toma valores en el intervalo $[0, \infty)$ y no está definido si la heredabilidad es 0. Cuando la heredabilidad es 1, α vale 0, a medida que la heredabilidad se aproxima a 0, α tiende a ∞.

En el primer escenario, $\alpha A^{-1} = 0$; por lo tanto, el pedigrí no tiene aporte alguno al predecir los valores genéticos aditivos. ¿Por qué ocurre esto? En primer lugar, la matriz A^{-1} contiene la información que aporta el pedigrí, en segundo lugar, si un fenotipo tiene heredabilidad en sentido estricto de 1, esto quiere decir que el fenotipo es indicador inequívoco del valor genético aditivo y que las diferencias fenotípicas (de tipo cuantitativo) que exhiben los progenitores, se heredan completamente por la progenie; por lo tanto, la información de parentesco que contiene el pedigrí no se necesita. Si bien este escenario es prácticamente hipotético porque en la práctica ningún fenotipo tiene heredabilidad de 1, resulta de interés porque nos muestra la conexión entre el álgebra y los razonamientos basados en los principios de la genética cuantitativa. En resumen, si un rasgo tiene heredabilidad de 1, el valor fenotípico es lo único que se necesita para seleccionar los individuos.

El otro extremo también es útil para entender el papel de α. A medida que la heredabilidad se acerca a 0, la información proveniente del pedigrí pesa más, lo cual tiene sentido porque cuando se tienen rasgos con poca heredabilidad, la información de parecido entre parientes se hace más relevante a la hora de predecir los valores genéticos de los individuos evaluados.

Por otro lado, para seguir comprendiendo la predicción del valor genético aditivo directo, a continuación, se presenta el EJEMPLO 4.1, tomado de Kennedy et al [23]. La TABLA NRO. 4.1 presenta el pedigrí y los registros para un fenotipo hipotético de ocho individuos.

TABLA 4.1: Información de pedigrí y registros del EJEMPLO 4.1

	Animal	Padre	Madre	Sexo	Fenotipo
1	1			1	10.00
2	2			2	9.00
3	3			1	8.00
4	4			2	7.00
5	5	1	2	1	9.00
6	6	1	2	2	10.00
7	7	3	4	1	8.00
8	8	5	6	2	11.00

Fuente: Datos tomados de Kennedy et al [23].

Se ajusta el siguiente modelo:

$$y = \mu 1_8 + Zu + e$$

Donde 1_8 es un vector de dimensión 8x1 cuyos elementos son todos iguales a 1, μ es un escalar correspondiente a la media general, y los demás elementos son tal y como se definieron antes. Por lo tanto, en este caso el único *efecto fijo* del modelo es la media general, que en sentido estricto no es un efecto, aquí, μ corresponde a β en el modelo general presentado arriba y 1_8 corresponde a X.

La solución para la media general y los valores genéticos aditivos directos tiene la forma:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \\ \widehat{a}_6 \\ \widehat{a}_7 \\ \widehat{a}_8 \end{bmatrix} = \begin{bmatrix} 8.7059 \\ 0.8676 \\ 0.3676 \\ -0.3676 \\ -0.8676 \\ 0.6716 \\ 1.0049 \\ -0.6471 \\ 1.3235 \end{bmatrix}$$

Más adelante se presenta el desarrollo completo del modelo animal para comprender la construcción de las EMM y el procedimiento para llegar a su solución, por ahora, dado que este ejemplo numérico busca mostrar algunas propiedades del modelo animal básico, centramos la discusión en las soluciones para los individuos 1 y 8 expresadas de forma algebraica.

Para el animal 1 tenemos:

$$\widehat{a}_1 = \frac{\left[(10 - \widehat{\mu}) + (\widehat{a}_5 - 0.5\widehat{a}_2) + (\widehat{a}_6 - 0.5\widehat{a}_2\right]}{3}$$

Así, vemos que la predicción del valor genético aditivo para uno de los machos fundadores, cuyos padres no se conocen, es una combinación del registro corregido por la media general, el único *efecto* fijo y la suma de los valores genéticos predichos de sus dos progenies, los individuos 5 y 6, corregidos por la mitad del valor genético de la hembra 2, que en este caso es la madre de 5 y 6. Cada una de estas tres fuentes tiene pesos iguales a $\frac{1}{3}$. Al no conocerse los padres de este individuo, el promedio de sus valores genéticos no entra en la predicción.

Ahora nos enfocamos en la solución para el macho 8, cuyos padres son 5 y 6 y el cual no cuenta con progenies:

$$\widehat{a}_8 = \frac{\widehat{a}_5 + \widehat{a}_6}{2} + \frac{\left[11 - \widehat{\mu} - \frac{\widehat{a}_5 + \widehat{a}_6}{2}\right]}{3}$$

Su predicción está compuesta por el promedio de los padres y un término que corresponde al registro del animal ajustado por la media general y el promedio de los padres. Este segundo término es un estimador del desvío de segregación mendeliano o término de segregación Mendeliana que corresponde a la diferencia entre el valor genético del individuo y el promedio de los valores genéticos de sus padres.

Ahora bien, el promedio de los MPLI de los valores genéticos aditivos de la población base es:

$$\frac{\widehat{a}_1 + \widehat{a}_2 + \widehat{a}_3 + \widehat{a}_4}{4} = \frac{0.8676 + 0.3676 - 0.3676 - 0.8676}{4} = 0$$

Lo cual nos muestra otra propiedad del modelo, pues, este es un supuesto implícito del mismo. Por otro lado, el promedio de los valores genéticos aditivos predichos de las demás generaciones no es nulo, para ilustrarlo, consideremos el promedio de los descendientes directos de los individuos fundadores:

$$\frac{\widehat{a}_5 + \widehat{a}_6 + \widehat{a}_7}{3} = \frac{0.6716 + 1.0049 - 0.6471}{3} = 0.343$$

Este cambio puede atribuirse a las fuerzas evolutivas de deriva o selección.

También es importante destacar que, además de las propiedades estadísticas del MPLI estudiadas en la sección 2.3, este predictor maximiza la probabilidad de un ordenamiento correcto entre pares de individuos, es decir, cuando se comparan dos potenciales reproductores usando los MPLI de sus valores genéticos, la probabilidad de identificar al que tiene el mayor valor genético verdadero se maximiza, aunque esto ocurre en la clase de predictores que siguen una propiedad denominada invarianza a la traslación. El lector interesado en profundizar en este tema, puede consultar a Henderson [14].

Las predicciones de los valores genéticos se acompañan de un valor de confiabilidad. En el argot del mejoramiento genético animal se define la exactitud del valor genético predicho del i-ésimo individuo ($\widehat{a}_i$) como su coeficiente de correlación de Pearson con el valor genético verdadero (a_i; Henderson [3]), esto es:

$$r(a_i, \widehat{a}_i) = \frac{Cov\left[a_i, \widehat{a}_i\right]}{\sqrt[2]{Var\left[a_i\right] Var\left[\widehat{a}_i\right]}}$$

Es importante considerar que esta definición no concuerda con el concepto estadístico de exactitud, el cual se asocia al sesgo, a menor sesgo, mayor exactitud. En general, la correlación toma valores entre -1 y 1, pero bajo los supuestos probabilísticos del modelo lineal mixto, se tiene que $r(a_i, \widehat{a}_i)$ es no-negativa (típicamente positiva) debido a que:

$$Cov\left[a_i, \widehat{a}_i\right] = Var\left[\widehat{a}_i\right] \geq 0$$

Los detalles de la identidad $Cov\left[a_i, \widehat{a}_i\right] = Var\left[\widehat{a}_i\right]$ pueden consultarse en las notas de mejoramiento animal de Elzo [24].

La teoría general del modelo lineal mixto indica que esta correlación es de la forma:

$$r(a_i, \widehat{a}_i) = \sqrt[2]{1 - \frac{Var\left[a_i - \widehat{a}_i\right]}{Var\left[a_i\right]}}$$

$$= \sqrt[2]{1 - \frac{VEP_i}{Var\left[a_i\right]}}$$

Donde VEP_i es la varianza del error de predicción del i-ésimo animal, pues la variable aleatoria $a_i - \widehat{a}_i$ se conoce como el error de predicción. La VEP_i se obtiene de la posición correspondiente al valor genético predicho del i-ésimo individuo en la diagonal principal de la inversa generalizada de la matriz de coeficientes de las ecuaciones de modelos mixtos. Bajo la notación utilizada en la sección 3.4, la inversa generalizada de la matriz de coeficientes de las EMM se representó como:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

Por lo tanto, esta varianza es el i-ésimo elemento de la diagonal principal de C^{22}. Sin embargo, en el caso particular que aquí se estudia, en el que las EMM se simplifican dando lugar a la matriz de coeficientes:

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \alpha A^{-1} \end{bmatrix}$$

Empleamos la siguiente notación:

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \alpha A^{-1} \end{bmatrix}^{-} = \begin{bmatrix} C^{*11} & C^{*12} \\ C^{*21} & C^{*22} \end{bmatrix}$$

Sea d_i el i-ésimo elemento de la diagonal principal de C^{*22}, en este caso $VEP_i = \sigma_e^2 d_i$. A la raíz cuadrada de VEP_i se le denomina error estándar de predicción.

Ahora bien, bajo el modelo animal considerado, se tiene que:

$$Var\left[a_i\right] = G_{ii} = \sigma_a^2(1 + F_i)$$

Donde G_{ii} es la i-ésima entrada de la diagonal principal de la matriz de covarianzas de a (el vector de valores genéticos aditivos), σ_a^2 es la varianza genética aditiva del fenotipo que se está evaluando y F_i es el coeficiente de endogamia del i-ésimo individuo. Entonces:

$$r(a_i, \widehat{a}_i) = \sqrt[2]{1 - \frac{\sigma_e^2 d_i}{\sigma_a^2(1 + F_i)}}$$

Al cuadrado de esta correlación se le denomina confiabilidad y puede reescribirse de la siguiente forma:

$$r^2(a_i, \widehat{a}_i) = 1 - \frac{\sigma_e^2 d_i}{\sigma_a^2(1 + F_i)}$$

$$= 1 - \alpha \frac{d_i}{1 + F_i}$$

En el caso de poblaciones sin consanguinidad, $F_i = 0 \; \forall \; i = 1, 2, \ldots, n$ y por consiguiente esta expresión se reduce a:

$$r^2(a_i, \widehat{a}_i) = 1 - \alpha d_i$$

Cuando se ejecuta una evaluación genética es necesario realizar actividades previas como la organización de la genealogía, se debe renumerar los animales en procura de que los ancestros tengan numeración menor a sus descendientes. En la sección 11

presentamos una rutina para realizar este procedimiento con la ayuda de R-project [25]. También es necesario analizar los registros para evidenciar errores o problemas en las bases de datos.

Adicionalmente, es necesario confirmar que los grupos contemporáneos tengan una buena cantidad de datos y que estén conectados genéticamente, es decir, existan parientes que estén en distintos grupos contemporáneos. Para esto, es necesario el análisis de conectibidad. En la sección 13 presentamos algunos conceptos y ejercicios resueltos.

4.2. Ejemplo en cuyes

Para entender de mejor manera el procedimiento que permite encontrar las soluciones de $\widehat{a}$ y β^0, se presenta el EJEMPLO 4.2 para la evaluación genética de un grupo de cuyes, en la cual la característica a mejorar, por ser de interés económico y productivo, corresponde al peso vivo a las ocho semanas de edad. En el EJEMPLO 4.2 se indicará la construcción del sistema de ecuaciones del modelo mixto, junto con los códigos de ejecución para resolverlas con el programa R-project [25].

La estructura de la base de datos contiene la información genealógica (animal, padre y madre), el sexo (que se modela como efecto fijo) y el peso vivo a las ocho semanas de edad. Los primeros cuatro individuos son considerados la población base (no tienen información de padres), además, no tienen información del pesaje (TABLA NRO. 4.2).

TABLA 4.2: Información de pedigrí y peso a las ocho semanas (g) cuyes (*Cavia porcellus Rodentia: caviidae*)

Animal	Padre	Madre	Sexo	Peso
1			1	
2			2	
3			1	
4			2	
5	1	2	1	750.00
6	1	4	2	630.00
7	3	4	1	620.00
8	3	2	2	600.00

Fuente: elaboración propia (2024).

Las ecuaciones del modelo mixto correspondientes a la base de datos antes descrita, se construyen de la siguiente manera:

Las columnas de la matriz X corresponden a los efectos fijos y aparecen en el mismo orden que estos tienen en el vector β, en este caso, el sexo que tiene dos niveles, machos

en la primera columna y hembras en la segunda columna. En las filas aparecerán los animales con registro para la característica por la cual se va a seleccionar.

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

En la matriz Z, las filas representan los animales con registros y en las columnas todos los animales, es decir, los ejemplares con registro del peso vivo a las ocho semanas y los padres y madres que carecen del mismo, pero que aparecen en el archivo genealógico o de pedigrí.

$$Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El vector y corresponde a los pesajes de los animales.

$$y = \begin{bmatrix} 750 \\ 630 \\ 620 \\ 600 \end{bmatrix}$$

Para la construcción del sistema de ecuaciones del modelo mixto se efectúan las siguientes operaciones matriciales:

$$X^T X = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$X^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1370 \\ 1230 \end{bmatrix}$$

$$Z^T y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 750 \\ 630 \\ 620 \\ 600 \end{bmatrix}$$

Antes de construir la matriz A se requiere tener organizada la base de datos con información genealógica de los individuos en tres columnas correspondientes al animal (a), al padre (p) y a la madre (m). Para esta organización se requiere mucho cuidado y conocimiento en la identificación de animales, ya que es muy común que se presenten errores en la digitación, que generan falsos parentescos, como por ejemplo: la identificación de un individuo se cambie al convertirse en padre, el mismo nombre para dos individuos, mala digitación, entre otros.

Para construir la matriz A se deben aplicar los procedimientos indicados por Henderson [7], de la siguiente manera:

1) Ordenar la genealogía iniciando con los animales que no tienen padre y madre (considerada la generación cero o población base) y posteriormente los animales por el orden de generación que les corresponde, así ningún individuo tiene un índice mayor al de sus descendientes, como se indica en el EJEMPLO 4.2 de las tres primeras columnas de la TABLA NRO. 4.2.

2) Se construirá la matriz simétrica A de dimensión nxn, siendo n el número de individuos evaluados, la entrada $i-j$ de esta matriz es la relación genética aditiva entre los individuos i y j cuando $i \neq j$, mientras que cuando $i = j$, es decir, en la diagonal principal, la entrada es 1, más el coeficiente de endogamia del animal.

3) Esta matriz se construye de forma recursiva desde el primer individuo hasta el último.

4) Para ello, los elementos de a_{ii} donde $i = 1, 2, \ldots, n$ corresponden a 1 más la mitad de la relación genética aditiva de sus padres de i (m_i y p_i). Es decir: $a_{ii} = 1 + \frac{1}{2} a_{p_i m_i}$. La relación genética aditiva entre dos individuos es el numerador del coeficiente de parentesco de Wright, razón por la cual a esta matriz se le denomina "numerator relationship matrix" en la literatura en inglés, que en español sería "matriz de relaciones del numerador".

5) Los elementos a_{ij}, donde $i \neq j$ corresponden a la relación genética aditiva entre

los individuos i y j. El cálculo se puede hacer por filas o por columnas. Al hacerlo por filas la entrada a_{ij} está dada por mitad de la suma de la relación genética aditiva del individuo i con el padre y la madre de j, esto es:

$$a_{ij} = \frac{1}{2}\left(a_{ip_j} + a_{im_j}\right)$$

Aquí se puede ver la importancia de ordenar los individuos por generación y la razón por la cual es un método recursivo, cuando se llega a la posición i, j, ya se tienen las relaciones genéticas aditivas de i con los padres de j, puesto que estos tienen índices menores a j.

En el EJEMPLO 4.2 la matriz de parentesco se construye así:

Como el individuo 1 no tiene padre y madre conocidos, se tendría el primer elemento:

$$a_{11} = 1 + 0$$

Ahora continuamos con la fila uno:

$$a_{12} = a_{13} = a_{14} = \frac{1}{2}(0 + 0) = 0$$

Los padres del individuo 5 son 1 y 2, por lo tanto:

$$a_{15} = \frac{1}{2}(a_{11} + a_{12}) = \frac{1}{2}(1 + 0) = 1/2$$

Los padres de 6 son también 1 y 2, por consiguiente:

$$a_{16} = \frac{1}{2}(a_{11} + a_{12}) = 1/2$$

Los individuos 7 y 8 son hermanos completos, sus padres son 3 y 4, por lo tanto:

$$a_{17} = \frac{1}{2}(a_{13} + a_{14}) = \frac{1}{2}(0 + 0) = 0 = a_{18}$$

Como la matriz es simétrica, la primera columna es igual a la primera fila, así, al pasar a la segunda fila iniciamos en la posición 22:

$$a_{22} = 1 + 0 = 1$$

Siguiendo el mismo procedimiento usado en la fila 1, tenemos:

$$a_{23} = \frac{1}{2}(0 + 0) = 0 = a_{24}$$

$$a_{25} = \frac{1}{2}(a_{21} + a_{22}) = \frac{1}{2}(0 + 1) = 1/2 = a_{26}$$

$$a_{27} = \frac{1}{2}(a_{23} + a_{24}) = \frac{1}{2}(0 + 0) = 0 = a_{28}$$

La matriz de parentesco sería:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0.25 & 0 & 0.25 \\ 0.5 & 0 & 0 & 0.5 & 0.25 & 1 & 0.25 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0.25 & 1 & 0.25 \\ 0 & 0.5 & 0.5 & 0 & 0.25 & 0 & 0.25 & 1 \end{bmatrix}$$

La inversa es[1]:

$$A^{-1} = \begin{bmatrix} 2 & 0.5 & 0 & 0.5 & -1 & -1 & 0 & 0 \\ 0.5 & 2 & 0.5 & 0 & -1 & 0 & 0 & -1 \\ 0 & 0.5 & 2 & 0.5 & 0 & 0 & -1 & -1 \\ 0.5 & 0 & 0.5 & 2 & 0 & -1 & -1 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

La matriz A^{-1} se le multiplica el valor de $\widehat{\alpha}$. En el EJEMPLO 4.2 las estimaciones corresponden a las obtenidas por Solarte et al [26, 27], que encontraron los parámetros del peso a las ocho semanas en cuyes de: $\widehat{\sigma_a^2} = 352$, $\widehat{\sigma_e^2} = 660$, por lo tanto:

$$\widehat{\alpha} = \frac{\widehat{\sigma_e^2}}{\widehat{\sigma_a^2}} = \frac{660}{352} = 1.875$$

[1]Es importante aclarar que existe un método desarrollado por Henderson para construir directamente esta inversa a partir del pedigrí, esto evita el costo computacional de investir matrices. Sin embargo, en este texto, al tratarse de ejemplos pequeños que buscan ilustrar el cómputo de los valores genéticos, construimos la matriz A y la invertimos directamente, pero este no es el procedimiento que se sigue con grandes conjuntos de datos.

Por consiguiente,

$$Z^TZ + \alpha A^{-1} = \begin{bmatrix} 3.75 & 0.94 & 0 & 0.94 & -1.87 & -1.88 & 0 & 0 \\ 0.94 & 3.75 & 0.94 & 0 & -1.87 & 0 & 0 & -1.88 \\ 0 & 0.94 & 3.75 & 0.94 & 0 & 0 & -1.88 & -1.88 \\ 0.94 & 0 & 0.94 & 3.75 & 0 & -1.87 & -1.87 & 0 \\ -1.87 & -1.88 & 0 & 0 & 4.75 & 0 & 0 & 0 \\ -1.87 & 0 & 0 & -1.87 & 0 & 4.75 & 0 & 0 \\ 0 & 0 & -1.88 & -1.88 & 0 & 0 & 4.75 & 0 \\ 0 & -1.88 & -1.88 & 0 & 0 & 0 & 0 & 4.75 \end{bmatrix}$$

Recordemos la discusión sobre MPLI empírico que se hizo en la sección 3.5. El paso siguiente es construir el sistema de ecuaciones del modelo mixto, así:

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 3.75 & 0.94 & 0 & 0.94 & -1.87 & -1.88 & 0 & 0 \\ 0 & 0 & 0.94 & 3.75 & 0.94 & 0 & -1.87 & 0 & 0 & -1.88 \\ 0 & 0 & 0 & 0.94 & 3.75 & 0.94 & 0 & 0 & -1.88 & -1.88 \\ 0 & 0 & 0.94 & 0 & 0.94 & 3.75 & 0 & -1.87 & -1.87 & 0 \\ 1 & 0 & -1.87 & -1.88 & 0 & 0 & 4.75 & 0 & 0 & 0 \\ 0 & 1 & -1.87 & 0 & 0 & -1.87 & 0 & 4.75 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1.88 & -1.88 & 0 & 0 & 4.75 & 0 \\ 0 & 1 & 0 & -1.88 & -1.88 & 0 & 0 & 0 & 0 & 4.75 \end{bmatrix} * \begin{bmatrix} \widehat{s_1} \\ \widehat{s_2} \\ \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \widehat{a_6} \\ \widehat{a_7} \\ \widehat{a_8} \end{bmatrix} = \begin{bmatrix} 1370 \\ 1230 \\ 0 \\ 0 \\ 0 \\ 0 \\ 750 \\ 630 \\ 620 \\ 600 \end{bmatrix}$$

En este sistema de ecuaciones, el lado izquierdo (LHS), se invierte y se multiplica por el vector del lado derecho (RHS) y al realizar dichas operaciones matriciales, se obtienen las soluciones para los parámetros fijos del modelo, que en este caso corresponden a las medias de cada sexo; y el efecto aleatorio del animal. Cabe resaltar que en el sistema de ecuaciones también se incluyen los animales que no tienen registro (peso a las ocho semanas).

Se obtiene el siguiente resultado:

$$\begin{bmatrix} \widehat{s_1} \\ \widehat{s_2} \\ \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \widehat{a_6} \\ \widehat{a_7} \\ \widehat{a_8} \end{bmatrix} = \begin{bmatrix} 685 \\ 615 \\ 13.9 \\ 8.7 \\ -13.9 \\ -8.7 \\ 22.6 \\ 5.2 \\ -22.6 \\ -5.2 \end{bmatrix}$$

En la actualidad, existe una interesante discusión sobre el uso de los valores p, que se agudiza en el caso de datos observacionales, aunque no es material para discutirse en este texto, nos parece necesario mencionarlo en este punto.

Desglozando el vector solución, se tendrían las medias de sexo:

$$\begin{bmatrix} \widehat{s_1} \\ \widehat{s_2} \end{bmatrix} = \begin{bmatrix} 685 \\ 615 \end{bmatrix}$$

Los machos y las hembras tienen tienen un peso numéricamente diferente a las ocho semanas, aunque debe aclararse que, si resulta de interés inferir estas diferencias, se puede realizar la correspondiente prueba de hipótesis, con el fin de establecer si existen diferencias estadísticamente significativas entre las medias de los sexos.

Si en el modelo de evaluación se incluyesen más efectos fijos, en el vector solución aparecerán los resultados para cada nivel, de cada efecto considerado. Por lo general, la revisión de literatura previa a la evaluación genética orienta sobre los factores fijos importantes que se deberán tener en cuenta.

El resto de los valores que aparecen en el vector solución, para el EJEMPLO 4.2 en particular, corresponden al valor genético aditivo o valor de cría (Breeding value) de los animales. Si el propósito es incrementar la producción, teniendo en cuenta la característica a mejorar, cuanto más alto sea este valor, existirá mayor probabilidad de que la progenie resultante de utilizar este animal como reproductor, obtenga un rendimiento mayor, respecto a la media del rebaño. Lo denominaremos como $\widehat{VG}$

$$\widehat{VG} = \begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \widehat{a_6} \\ \widehat{a_7} \\ \widehat{a_8} \end{bmatrix} = \begin{bmatrix} 13.9 \\ 8.7 \\ -13.9 \\ -8.7 \\ 22.6 \\ 5.2 \\ -22.6 \\ -5.2 \end{bmatrix}$$

Debe tenerse en cuenta que la habilidad de transmisión predecible (PTA, por sus siglas en inglés) o diferencias esperadas en la progenie (DEP, pos sus siglas en inglés), se calcula dividiendo por dos el valor genético obtenido en el vector solución para cada animal y debe interpretarse de la siguiente manera: Si el animal se aparea al azar, su progenie tendrá un valor promedio del peso a las ocho semanas inferior o superior a la media de la población. Por ejemplo, el animal de más alto valor genético, que es el 5, tiene un valor genético de 22.6 gramos, por lo tanto, su habilidad de transmisión será de 11.3 g, lo que quiere decir que si este animal se aparea al azar en la población de

donde proviene, su progenie tendrá en promedio un peso a las ocho semanas superior en 11.3 g, respecto a la media de la población. El vector con las diferencias esperadas de progenie es:

$$\widehat{Dep} = \begin{bmatrix} 6.95 \\ 4.35 \\ -6.95 \\ -4.35 \\ 11.3 \\ 2.6 \\ -11.3 \\ -2.6 \end{bmatrix}$$

Con la información disponible se pueden calcular otros parámetros de suma importancia en la valoración genética de un individuo, como los que se indican a continuación: Ahora calculamos r^2 empleando el procedimiento que se usó antes, obtenemos el vector correspondiente a los elementos de la diagonal principal de C^{22}, denominaremos este vector como d, para el EJEMPLO 4.2 se tendría:

$$d = \begin{bmatrix} 0.49 \\ 0.49 \\ 0.49 \\ 0.49 \\ 0.44 \\ 0.44 \\ 0.44 \\ 0.44 \end{bmatrix}$$

Con el anterior vector, podemos realizar los cálculos de la correlación entre ese valor y el verdadero valor genético, la confiabilidad y la raiz cuadrada de la varianza del error de predicción (a la que denominaremos SEP), donde $\widehat{SEP} = \sqrt[2]{(d_i * \sigma_e^2)}$.

La confiabilidad sería:

$$r^2 = 1_8 - (d * \alpha) = \begin{bmatrix} 0.09 \\ 0.09 \\ 0.09 \\ 0.09 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix}$$

Ahora se toma la raiz cuadrada de cada elemento de este vector para obtener la exactitud:

$$r = \begin{bmatrix} 0.29 \\ 0.29 \\ 0.29 \\ 0.29 \\ 0.41 \\ 0.41 \\ 0.41 \\ 0.41 \end{bmatrix}$$

Así, empleando la fórmula descrita previamente, el vector de los SEP es:

$$\widehat{SEP} = \begin{bmatrix} 17.93 \\ 17.93 \\ 17.93 \\ 17.93 \\ 17.05 \\ 17.05 \\ 17.05 \\ 17.05 \end{bmatrix}$$

En la tabla TABLA NRO. 4.3 se tiene el cuadro detallado de las valoraciones genéticas de los animales. Ahora se pueden ordenar los animales por su mérito genético. Por ejemplo, el animal 5 tiene una *DEP* de 11.3 *g*, error estándar de la predicción de 17.05 *g* y confiabilidad de 0.17.

Es oportuno indicar que la confiabilidad se incrementa en dependencia de la cantidad y calidad de la información, tanto de los datos productivos, como del pedigrí.

En las aplicaciones reales, si se busca aumentar la media del fenotipo, se buscará seleccionar como reproductores a los animales de mayor valor genético y mayor confiabilidad. Similarmente, si se busca reducir la media del fenotipo, se seleccionan animales con menores valores genéticos y alta confiabilidad.

Al programar los apareamientos de los animales seleccionados por evaluación genética, es indispensable hacer un estudio de la consanguinidad que se espera en la siguiente generación y el tamaño efectivo. En la sección 12 podrá encontrar procedimientos para estos análisis y su importancia en la mejora genética de animales domésticos.

Para finalizar, a continuación, se presenta la correspondiente codificación para ejecutar los dos ejemplos de este capítulo en el programa R-Project [25]:

TABLA 4.3: Catálogo de cuyes

	DEP	Elemento diagonal de C^{22}	Confiabilidad	Exactitud	SEP
5	11.30	0.44	0.17	0.41	17.05
1	6.95	0.49	0.09	0.29	17.93
2	4.35	0.49	0.09	0.29	17.93
6	2.60	0.44	0.17	0.41	17.05
8	-2.60	0.44	0.17	0.41	17.05
4	-4.35	0.49	0.09	0.29	17.93
3	-6.95	0.49	0.09	0.29	17.93
7	-11.30	0.44	0.17	0.41	17.05

Nota: Catálogo de valoración genética de animales mediante modelo animal. DEP=diferencia esperada de la progenie, SEP=raiz cuadrada de la varianza del error de predicción.
Fuente: elaboración propia generada en R-project [25].

4.3. Ejercicios en R-project

Montaje de la genealogía del EJEMPLO 4.1, tomado de Kennedy et al [23] en R-project [25]:

```
GenoyFeno=data.frame(matrix(ncol=5,byrow=TRUE,c(
"1",NA,NA,1,10,
"2",NA,NA,2,9,
"3",NA,NA,1,8,
"4",NA,NA,2,7,
"5","1","2",1,9,
"6","1","2",2,10,
"7","3","4",1,8,
"8","5","6",2,11)))
colnames(GenoyFeno)=c("id","sire","dam","sex", "Fenotipo")
GenoyFeno$Fenotipo=as.numeric(GenoyFeno$Fenotipo)
```

```
GenoyFeno
```

```
##   id sire  dam sex Fenotipo
## 1  1 <NA> <NA>   1       10
## 2  2 <NA> <NA>   2        9
## 3  3 <NA> <NA>   1        8
## 4  4 <NA> <NA>   2        7
## 5  5    1    2   1        9
## 6  6    1    2   2       10
## 7  7    3    4   1        8
## 8  8    5    6   2       11
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id = GenoyFeno$id, dadid = GenoyFeno$sire,
                momid = GenoyFeno$dam,
                sex=as.numeric(GenoyFeno$sex))
```

El árbol genealógico está en la FIGURA NRO. 4.1:

```
plot(Genea1,
   mar = c(bottom=0, left=1, top=1, right=1), cex=1)
```

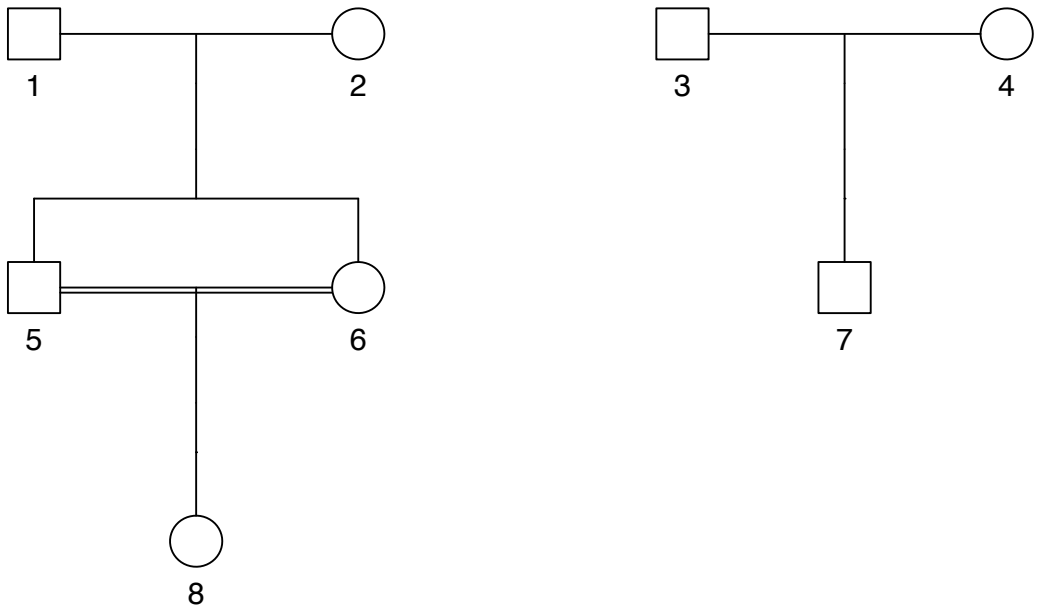

***Figura* 4.1:** Genealogía de ocho animales.
Fuente: elaboración propia generada en R-project [25].

Montaje de la matriz de parentesco (A):

```
A=2*kinship(GenoyFeno$id,GenoyFeno$sire,GenoyFeno$dam)
A
##     1   2   3   4    5    6   7    8
## 1 1.0 0.0 0.0 0.0 0.50 0.50 0.0 0.50
```

```
## 2 0.0 1.0 0.0 0.0 0.50 0.50 0.0 0.50
## 3 0.0 0.0 1.0 0.0 0.00 0.00 0.5 0.00
## 4 0.0 0.0 0.0 1.0 0.00 0.00 0.5 0.00
## 5 0.5 0.5 0.0 0.0 1.00 0.50 0.0 0.75
## 6 0.5 0.5 0.0 0.0 0.50 1.00 0.0 0.75
## 7 0.0 0.0 0.5 0.5 0.00 0.00 1.0 0.00
## 8 0.5 0.5 0.0 0.0 0.75 0.75 0.0 1.25
```

```
bitSize(Genea1)
```

```
## $bitSize
## [1] 4
##
## $nFounder
## [1] 4
##
## $nNonFounder
## [1] 4
```

Número de individuos:

```
n=nrow(GenoyFeno)
n
```

```
## [1] 8
```

Montaje del vector de media (m) la matriz Z y el vector y:

```
m=matrix(nrow=nrow(GenoyFeno), ncol=1, 1)
rownames(m)=GenoyFeno$id
colnames(m)=c("media")
m
```

```
##   media
## 1     1
## 2     1
## 3     1
## 4     1
## 5     1
## 6     1
## 7     1
## 8     1
```

```
library(MatrixModels)
Z=as.matrix(model.Matrix(~ as.factor(GenoyFeno$id) -1))
library(stringr)
colnames(Z)=word(colnames(Z), 2, sep = fixed(')'))
rownames(Z)=GenoyFeno$id
Z
```

```
##   1 2 3 4 5 6 7 8
## 1 1 0 0 0 0 0 0 0
## 2 0 1 0 0 0 0 0 0
## 3 0 0 1 0 0 0 0 0
## 4 0 0 0 1 0 0 0 0
## 5 0 0 0 0 1 0 0 0
## 6 0 0 0 0 0 1 0 0
## 7 0 0 0 0 0 0 1 0
## 8 0 0 0 0 0 0 0 1
```

```
y=as.matrix(GenoyFeno$Fenotipo)
colnames(y)=c("Fenotipo")
rownames(y)=GenoyFeno$id
y
```

```
##   Fenotipo
## 1       10
## 2        9
## 3        8
## 4        7
## 5        9
## 6       10
## 7        8
## 8       11
```

Montaje del sistema de ecuaciones:

```
mpm=t(m)%*%m
mpm
```

```
##       media
## media     8
```

```
mpZ=t(m)%*%Z
mpZ
```

```
##       1 2 3 4 5 6 7 8
## media 1 1 1 1 1 1 1 1
```

```
Zpm=t(Z)%*%m
```

```
ZpZ=t(Z)%*%Z
```

```
mpy=t(m)%*%y
mpy
```

```
##       Fenotipo
## media       72
```

```
Zpy=t(Z)%*%y
```

Inclusión del valor de α:

```
var_a=.5
var_e=.5
h2=var_a/(var_a+var_e)
h2
```

```
## [1] 0.5
```

```
alpha=var_e/var_a
alpha
```

```
## [1] 1
```

```
library(MASS)
alfaAinv=alpha%x%ginv(A)
ZpZmasalfaAinv=ZpZ+alfaAinv
```

Construcción del lado izquierdo del sistema:

```
Izq=rbind(
  cbind(mpm,mpZ),
  cbind(Zpm,ZpZmasalfaAinv))
Izqinv=solve(Izq)
```

Lado derecho del sistema:

```
Der=rbind(mpy,Zpy)
```

Solución al sistema:

```
Sol=round(Izqinv%*%Der,4)
rownames(Sol)=c(rownames(mpm),rownames(ZpZ))
```

La media:

```
solfijos=as.matrix(Sol[rownames(mpZ),])
solfijos
```

```
##       [,1]
## [1,]  8.7
```

Valores genéticos:

```
VG=as.matrix(Sol[rownames(ZpZ),])
VG
```

```
##     [,1]
## 1  0.87
## 2  0.37
## 3 -0.37
## 4 -0.87
## 5  0.67
## 6  1.00
## 7 -0.65
## 8  1.32
```

EJEMPLO4.2, con información de pesaje a las ocho semanas de cuyes, en R-project [25]:

```
Genealogia=data.frame(matrix(ncol=4,byrow=TRUE,c(
"1",NA,NA,1,
"2",NA,NA,2,
"3",NA,NA,1,
"4",NA,NA,2,
"5","1","2",1,
"6","1","4",2,
"7","3","4",1,
"8","3","2",2)))
colnames(Genealogia)=c("id","sire","dam","sex")
```

```
Genealogia
```

```
##   id sire  dam sex
## 1  1 <NA> <NA>   1
## 2  2 <NA> <NA>   2
## 3  3 <NA> <NA>   1
## 4  4 <NA> <NA>   2
## 5  5    1    2   1
## 6  6    1    4   2
## 7  7    3    4   1
## 8  8    3    2   2
```

```
library(kinship2)
Genea1=pedigree(id = Genealogia$id, dadid = Genealogia$sire,
                momid = Genealogia$dam,
                sex=as.numeric(Genealogia$sex))
```

Montaje de la matriz de parentesco:

```
A=2*kinship(Genealogia$id,Genealogia$sire,Genealogia$dam)
A
```

```
##     1   2   3   4    5    6    7    8
## 1 1.0 0.0 0.0 0.0 0.50 0.50 0.00 0.00
## 2 0.0 1.0 0.0 0.0 0.50 0.00 0.00 0.50
## 3 0.0 0.0 1.0 0.0 0.00 0.00 0.50 0.50
## 4 0.0 0.0 0.0 1.0 0.00 0.50 0.50 0.00
## 5 0.5 0.5 0.0 0.0 1.00 0.25 0.00 0.25
```

```
## 6 0.5 0.0 0.0 0.5 0.25 1.00 0.25 0.00
## 7 0.0 0.0 0.5 0.5 0.00 0.25 1.00 0.25
## 8 0.0 0.5 0.5 0.0 0.25 0.00 0.25 1.00
```

El árbol genealógico está en la FIGURA NRO. 4.2:

```
plot(Genea1,
   mar = c(bottom=0, left=1, top=6, right=1), cex=1)
```

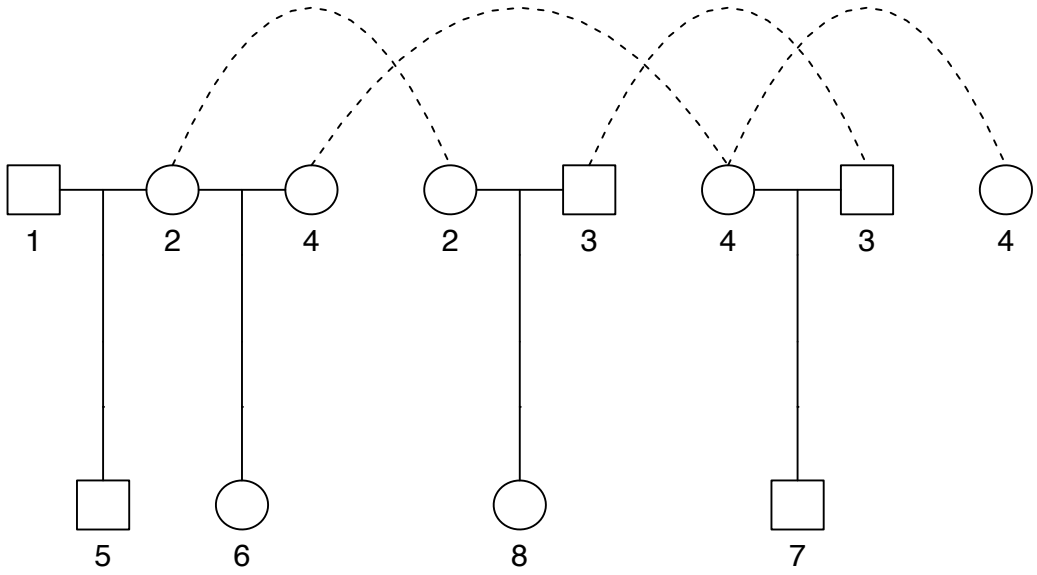

***Figura* 4.2:** Genealogía de ocho cuyes para el ejercicio de modelo animal.
Fuente: elaboración propia generada en R-project [25].

```
bitSize(Genea1)


## $bitSize
## [1] 4
##
## $nFounder
## [1] 4
##
## $nNonFounder
## [1] 4
```

Información de producciones:

```
Peso8s=data.frame(matrix(ncol=3,byrow=TRUE,c(
"5",1,750,
"6",2,630,
"7",1,620,
"8",2,600
  )))
colnames(Peso8s)=c("id","sexo","peso")
Peso8s$peso=as.numeric(Peso8s$peso)
Peso8s$sexo=as.numeric(Peso8s$sexo)
Peso8s
```

```
##   id sexo peso
## 1  5    1  750
## 2  6    2  630
## 3  7    1  620
## 4  8    2  600
```

Número de individuos:

```
n=nrow(Genealogia)
n
```

```
## [1] 8
```

Montaje de las matrices X y Z y el vector y:

```
library(MatrixModels)
X=as.matrix(model.Matrix(~ as.factor(Peso8s$sexo) -1))
rownames(X)=Peso8s$id
colnames(X)=c("s_1","s_2")
X
```

```
##   s_1 s_2
## 5   1   0
## 6   0   1
## 7   1   0
## 8   0   1
```

```
Peso8=as.matrix(model.Matrix(~ as.factor(Peso8s$id) -1))
library(stringr)
colnames(Peso8)=word(colnames(Peso8), 2, sep = fixed(')'))
rownames(Peso8)=Peso8s$id
round(Peso8,2)
```

```
##   5 6 7 8
## 5 1 0 0 0
## 6 0 1 0 0
## 7 0 0 1 0
## 8 0 0 0 1
```

```
Z=matrix(nrow=nrow(Peso8),ncol=n,0)
colnames(Z)=colnames(A)
rownames(Z)=colnames(Peso8)
Z[colnames(Peso8),colnames(Peso8)]=Peso8
Z
```

```
##   1 2 3 4 5 6 7 8
## 5 0 0 0 0 1 0 0 0
## 6 0 0 0 0 0 1 0 0
## 7 0 0 0 0 0 0 1 0
## 8 0 0 0 0 0 0 0 1
```

```
y=as.matrix(Peso8s$peso)
colnames(y)=c("P8")
rownames(y)=Peso8s$id
y
```

```
##    P8
## 5 750
## 6 630
## 7 620
## 8 600
```

Montaje del sistema de ecuaciones:

```
XpX=t(X)%*%X
```

```
XpZ=t(X)%*%Z
```

```
ZpX=t(Z)%*%X
```

```
ZpZ=t(Z)%*%Z
```

```
Xpy=t(X)%*%y
```

```
Zpy=t(Z)%*%y
```

Inclusión del valor de alfa:

```
var_a=352
var_e=660
h2=var_a/(var_a+var_e)
h2


## [1] 0.35


alpha=var_e/var_a
alpha


## [1] 1.9
```

```
library(MASS)
alfaAinv=alpha%x%ginv(A)
ZpZmasalfaAinv=ZpZ+alfaAinv
```

Solución del sistema, lado izquierdo de la matriz:

```
Izq=rbind(
  cbind(XpX,XpZ),
  cbind(ZpX,ZpZmasalfaAinv))
Izqinv=solve(Izq)
```

Lado derecho de la matriz:

```
Der=rbind(Xpy,Zpy)
```

Solución al sistema:

```
Sol=round(Izqinv%*%Der,1)
rownames(Sol)=c(rownames(XpX),rownames(ZpZ))
```

Efectos fijos:

```
solfijos=as.matrix(Sol[rownames(XpZ),])
solfijos
```

```
##      [,1]
## s_1   685
## s_2   615
```

Valores genéticos:

```
VG=as.matrix(Sol[rownames(ZpZ),])
VG
```

```
##     [,1]
## 1   13.9
## 2    8.7
## 3 -13.9
## 4  -8.7
## 5   22.6
## 6    5.2
## 7 -22.6
## 8  -5.2
```

Diferencia esperada de progenie:

```
Deps=VG/2
Deps
```

```
##     [,1]
## 1    7.0
## 2    4.3
```

```
## 3  -7.0
## 4  -4.3
## 5  11.3
## 6   2.6
## 7 -11.3
## 8  -2.6
```

Valores de la diagonal del lado izquierdo de las ecuaciones de modelo mixto relacionados con los animales:

```
diagonal=as.matrix(diag(Izqinv[rownames(ZpZ),colnames(ZpZ)]))
diagonal
```

```
##   [,1]
## 1 0.49
## 2 0.49
## 3 0.49
## 4 0.49
## 5 0.44
## 6 0.44
## 7 0.44
## 8 0.44
```

Confiabilidad:

```
Confiab_r2=as.matrix(1-diagonal*alpha)
Confiab_r2
```

```
##    [,1]
## 1 0.087
## 2 0.087
## 3 0.087
## 4 0.087
## 5 0.174
## 6 0.174
## 7 0.174
## 8 0.174
```

Exactitud:

```
Exact_r=sqrt(Confiab_r2)
Exact_r
```

```
##   [,1]
## 1 0.29
## 2 0.29
## 3 0.29
## 4 0.29
## 5 0.42
## 6 0.42
## 7 0.42
## 8 0.42
```

Los SEP son:

```
SEP=as.matrix(sqrt(diagonal*var_e))
SEP
```

```
##   [,1]
## 1   18
## 2   18
## 3   18
## 4   18
## 5   17
## 6   17
## 7   17
## 8   17
```

5 CAPÍTULO CINCO

MODELOS ANIMALES CON EFECTOS ADICIONALES

Carlos Eugenio Solarte Portilla
Universidad de Nariño

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

5.1. Modelo animal para medidas repetidas

Este tipo de modelos se utiliza para predecir valores genéticos cuando se evalúan características que se miden más de una vez durante la vida productiva, como por ejemplo, la producción de leche en varias lactancias de especies como los bovinos, ovinos y caprinos; la cantidad de lana de ovejas medida en distintas esquilas; el tamaño de camada en partos sucesivos en especies multíparas como conejos, cuyes y cerdos, entre otras muchas características. Una premisa importante es que se trata de la misma característica, es decir, que las observaciones en diferentes puntos en el tiempo son del mismo fenotipo, lo cual implica que la correlación genética entre estas sea 1.

En este tipo de modelos se tendrán tres componentes de varianza, dos que se estudiaron en el modelo anterior (σ_a^2 y σ_e^2), más la varianza debida a efectos ambientales que permanecen a través del tiempo, denominada de ambiente permanente (σ_p^2) y además, bajo el supuesto de independencia de los efectos genéticos aditivos directos, los de ambiente permanente y los errores. La varianza fenotípica (σ_f^2) será la sumatoria de las tres varianzas antes indicadas.

Los parámetros correspondientes a cocientes entre los componentes de varianza y la varianza fenotípica serían: la heredabilidad (h^2) y la repetibilidad (Rep). La Rep se define como la correlación entre los registros de un mismo individuo. Las ecuaciones serían:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2}$$

$$Rep = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2}$$

Es importante indicar que, bajo este modelo, se asume que los efectos genéticos y sus correspondientes interacciones, son iguales para cada una de las mediciones que se hagan de la característica a evaluar y que se pretende mejorar, en cambio los efectos temporales si pueden cambiar entre una medición y otra.

El número de mediciones puede extenderse, prácticamente de manera ilimitada y la base de datos fenotípica simplemente adquirirá mayor tamaño en función del número de mediciones, obviamente cuanto mayor sea el número de mediciones.

Las evaluaciones genéticas bajo este modelo permitirán predecir el valor genético y los efectos de ambiente permanente, que debe entenderse como la acción ambiental que altera el desempeño de los animales durante toda su etapa productiva, como por ejemplo, una enfermedad de la ubre que deja daño permanente o la pérdida de un pezón, independientemente de la causa de dicha pérdida o, en general, cualquier acción del ambiente con efecto a lo largo de la vida productiva de los animales, más los efectos genéticos no aditivos.

El modelo para estos casos, expresado en forma matricial, es el siguiente:

$$y = X\beta + Za + Sp + e$$

Donde y es el vector de observaciones, β es el vector de efectos fijos, a es el vector de valores genéticos aditivos directos, p es el vector de efectos aleatorios del ambiente permanente, e es el vector aleatorio de errores y X, Z y S, son matrices de incidencia que relacionan los registros con los efectos antes indicados, es decir, los efectos fijos del animal y ambientales permanentes, respectivamente.

Los supuestos probabilísticos para los valores genéticos aditivos directos y los errores son los mismos que se presentaron para el modelo animal básico, por otro lado, se asume que los efectos ambientales permanentes se distribuyen normal, independientemente, con vector de medias 0 y matriz de covarianza $\sigma_p^2 I$ y son independientes de los valores genéticos aditivos y de los errores.

La matriz de covarianza del error será igual a $\sigma_e^2 I = R$. La matriz de covarianza de los valores genéticos aditivos directos igual a $\sigma_a^2 A$ donde A es la matriz de parentesco.

La matriz de covarianza fenotípica será:

$$Var(y) = \sigma_a^2 ZAZ^T + \sigma_p^2 SS^T + \sigma_e^2 I$$

Las ecuaciones del modelo mixto (MME), para encontrar $\left[\widehat{\beta}, \widehat{a}, \widehat{p}\right]$ son las siguientes:

$$\begin{bmatrix} X^T X & X^T Z & X^T S \\ Z^T X & Z^T Z + \alpha_1 A^{-1} & Z^T S \\ S^T X & S^T Z & S^T S + \alpha_2 I \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{a} \\ \widehat{p} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \\ S^T y \end{bmatrix}$$

Donde: $\alpha_1 = \frac{\sigma_e^2}{\sigma_a^2}$ y $\alpha_2 = \frac{\sigma_e^2}{\sigma_p^2}$

A continuación, se presenta el EJEMPLO 5.1 con los datos de producción de proteína en bovinos para leche en el trópico alto de Nariño Colombia. En este ejemplo los valores de α_1 y α_2 se calcularán con los datos arbitrarios para la heredabilidad y la repetibilidad de 0.38 y 0.15, respectivamente; mientras que los registros de producción corresponden a los reportados por Solarte et al [31]. Es importante aclarar que los datos corresponden a lactancias completas, ajustadas a 305 días y equivalente adulto.

Para facilitar la lectura e interpretación de este modelo, se toman pocos datos, que corresponden a tres hembras con información de dos partos sucesivos. Estas hembras parieron entre los años 2019 y 2021 en dos hatos, por lo que se decidió formar grupos contemporáneos según el hato, el año y el semestre de parto, asi: Grupo 1) hato 1 en el primer semestre del 2019, Grupo 2) hato 2 en el segundo semestre del 2020 y Grupo 3) hato 2 en el primer semestre de 2021. La información productiva y genealógica se presenta en la TABLA NRO. 5.1.

A continuación, se indican las matrices para construir las ecuaciones de modelo mixto y las soluciones, que en este caso corresponderá al efecto fijo de grupo contemporáneo, el efecto aleatorio del animal (valor genético aditivo directo) y el efecto aleatorio del ambiente permanente.

Tabla 5.1: Información genelógica de seis animales y producción de proteína (kg) por lactancia de cuatro vacas

Animal	Padre	Madre	Sexo	Grupo Contemporáneo	Proteína
1			1		
2			2		
3			1		
4	1	2	2	1	145.00
4	1	2	2	2	153.00
5	3	4	2	1	128.00
5	3	4	2	3	131.00
6	1	2	2	2	172.00
6	1	2	2	3	172.00

Fuente: elaboración propia (2024).

La matriz X se construye de acuerdo con los niveles de los efectos fijos que se incluyan en el modelo, en este caso hay un solo efecto fijo que es el grupo contemporáneo con tres niveles. Así que en las filas aparecerán los animales con registro y en las columnas los niveles del efecto fijo, en este caso tres. Para el caso puntual del animal 4 se marca con 1 en el primer nivel del grupo contemporáneo, para la primera lactancia, con 1 en el segundo nivel para la segunda lactancia y no aparece con registros en el tercer nivel del efecto fijo[1].

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

La matriz Z contiene en las filas los animales que tienen mediciones, los cuales aparecen, cuantas veces tengan registro, en el Ejemplo 5.1 se tendrán 6 filas correspondientes a las vacas 4, 5 y 6 con sus dos lactancias cada una. En las columnas van todos los individuos que aparecen en la matriz de parentesco se tienen los animales 1, 2 y 3 que no tiene registros y los que tienen registro.

$$Z = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

[1] Al igual que en el capítulo anterior, el modelo no parametriza en términos de los efectos de cada grupo contemporáneo, sino de sus medias; sin embargo, para mantener la nomenclatura usual, se sigue mencionando como efectos fijos del modelo.

La matriz S relaciona los registros con el efecto ambiental permanente. Las columnas representan los animales con registros y las filas corresponden a los registros ordenados por animal:

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

El vector de observaciones y es:

$$y = \begin{bmatrix} 145 \\ 153 \\ 128 \\ 131 \\ 172 \\ 172 \end{bmatrix}$$

Las operaciones entre matrices son:

$$X^T X = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$X^T Z = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

$$X^T S = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$Z^T S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$S^T S = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 273 \\ 325 \\ 303 \end{bmatrix}$$

$$Z^T y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 298 \\ 259 \\ 344 \end{bmatrix}$$

$$S^T y = \begin{bmatrix} 298 \\ 259 \\ 344 \end{bmatrix}$$

La matriz A, cuya forma de construir ya fue explicada en el modelo básico del EJEMPLO 5.1 tiene la forma:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0.5 & 0.25 & 0.5 \\ 0 & 1 & 0 & 0.5 & 0.25 & 0.5 \\ 0 & 0 & 1 & 0 & 0.5 & 0 \\ 0.5 & 0.5 & 0 & 1 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.5 & 0.5 & 1 & 0.25 \\ 0.5 & 0.5 & 0 & 0.5 & 0.25 & 1 \end{bmatrix}$$

$$\alpha_1 = \frac{\sigma_e^2}{\sigma_a^2} = \frac{(1-r)}{h^2} = \frac{(1-0.38)}{0.15} = 4.13$$

$$Z^TZ + \alpha_1 A^{-1} = \begin{bmatrix} 8.27 & 4.13 & 0 & -4.13 & 0 & -4.13 \\ 4.13 & 8.27 & 0 & -4.13 & 0 & -4.13 \\ 0 & 0 & 6.2 & 2.07 & -4.13 & 0 \\ -4.13 & -4.13 & 2.07 & 12.33 & -4.13 & 0 \\ 0 & 0 & -4.13 & -4.13 & 10.27 & 0 \\ -4.13 & -4.13 & 0 & 0 & 0 & 10.27 \end{bmatrix}$$

$$\alpha_2 = \frac{\sigma_e^2}{\sigma_p^2} = \frac{1-r}{r-h^2} = \frac{1-0.38}{0.38-0.15} = 2.69$$

$$S^TS + I\alpha_2 = \begin{bmatrix} 4.7 & 0 & 0 \\ 0 & 4.7 & 0 \\ 0 & 0 & 4.7 \end{bmatrix}$$

Con la información suministrada hasta este punto es posible construir el sistema de ecuaciones del modelo mixto, cuya solución es:

$$\begin{bmatrix} \widehat{GC_1} \\ \widehat{GC_2} \\ \widehat{GC_3} \\ \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \widehat{a_6} \\ \widehat{p_4} \\ \widehat{p_5} \\ \widehat{p_6} \end{bmatrix} = \begin{bmatrix} 2 & 0 & . & . & 1 & 0 \\ 0 & 2 & . & . & 0 & 1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & 0 & . & . & 4.7 & 0 \\ 0 & 1 & . & . & 0 & 4.7 \end{bmatrix}^{-1} \begin{bmatrix} 273 \\ 325 \\ 303 \\ 0 \\ 0 \\ 0 \\ 298 \\ 259 \\ 344 \\ 298 \\ 259 \\ 344 \end{bmatrix}$$

La solución del sistema es:

$$\begin{bmatrix} \widehat{GC_1} \\ \widehat{GC_2} \\ \widehat{GC_3} \\ \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \widehat{a_6} \\ \widehat{p_4} \\ \widehat{p_5} \\ \widehat{p_6} \end{bmatrix} = \begin{bmatrix} 141 \\ 158.1 \\ 151.4 \\ 1 \\ 1 \\ -1.9 \\ -0.1 \\ -2.9 \\ 2.9 \\ -0.2 \\ -5.9 \\ 6.1 \end{bmatrix}$$

Los tres primeros elementos del vector corresponden a las medias de los tres grupos contemporáneos. En cuanto al valor genético, el animal 6 es el de mejor mérito con una cifra de 2.9 kg, por lo que su habilidad de transmisión predicha es de 1.45 kg, que debe interpretarse como la superioridad promedio de su descendencia, respecto a la media de la población, si este animal se aparea al azar en esa población. La misma interpretación es válida para el resto de los animales incluidos en la evaluación genética.

Las soluciones para los efectos permanentes representan, como ya se mencionó, tanto las influencias ambientales como los efectos genéticos no aditivos. Estas influencias pueden ser favorables o negativas. En el EJEMPLO 5.1, para los animales 4 y 5 la influencia es negativa y para el animal 6 es positiva.

Bajo este modelo, la suma del valor genético y del efecto ambiental permanente ($\widehat{a}_i + \widehat{p}_i$) produce el denominado *probable habilidad productora*, medida muy importante que representa un estimador del futuro desempeño del animal en el mismo hato.

El EJEMPLO 5.1 se desarrolla en R-Project [25] de la siguiente manera:

5.1.1. Desarrollo del EJEMPLO 5.1 en R-project

Producción de Prodeína por lactancia (kg) Montaje de la genealogía:

```
Genealogia=data.frame(matrix(ncol=4,byrow=TRUE,c(
"1",NA,NA,1,
"2",NA,NA,2,
"3",NA,NA,1,
"4","1","2",2,
"5","3","4",2,
"6","1","2",2)))
colnames(Genealogia)=c("id","sire","dam","sex")
```

```
Genealogia
```

```
##   id sire  dam sex
## 1  1 <NA> <NA>   1
## 2  2 <NA> <NA>   2
## 3  3 <NA> <NA>   1
## 4  4    1    2   2
## 5  5    3    4   2
## 6  6    1    2   2
```

Relaciones entre padres y madres:

```
table(Genealogia$sire,Genealogia$dam)
```

```
##
##     2 4
##   1 2 0
##   3 0 1
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id = Genealogia$id, dadid = Genealogia$sire,
                momid = Genealogia$dam,
                sex=as.numeric(Genealogia$sex))
```

Montaje de la matriz de parentesco:

```
A=2*kinship(Genealogia$id,Genealogia$sire,Genealogia$dam)
A
```

```
##      1    2   3   4    5    6
## 1 1.00 0.00 0.0 0.5 0.25 0.50
## 2 0.00 1.00 0.0 0.5 0.25 0.50
## 3 0.00 0.00 1.0 0.0 0.50 0.00
## 4 0.50 0.50 0.0 1.0 0.50 0.50
## 5 0.25 0.25 0.5 0.5 1.00 0.25
## 6 0.50 0.50 0.0 0.5 0.25 1.00
```

```
bitSize(Genea1)
```

```
## $bitSize
## [1] 3
##
## $nFounder
## [1] 3
##
## $nNonFounder
## [1] 3
```

El árbol genealógico está en la FIGURA NRO. 5.1:

```
plot(Genea1,mar=c(bottom=0, left=1, top=1, right=1))
```

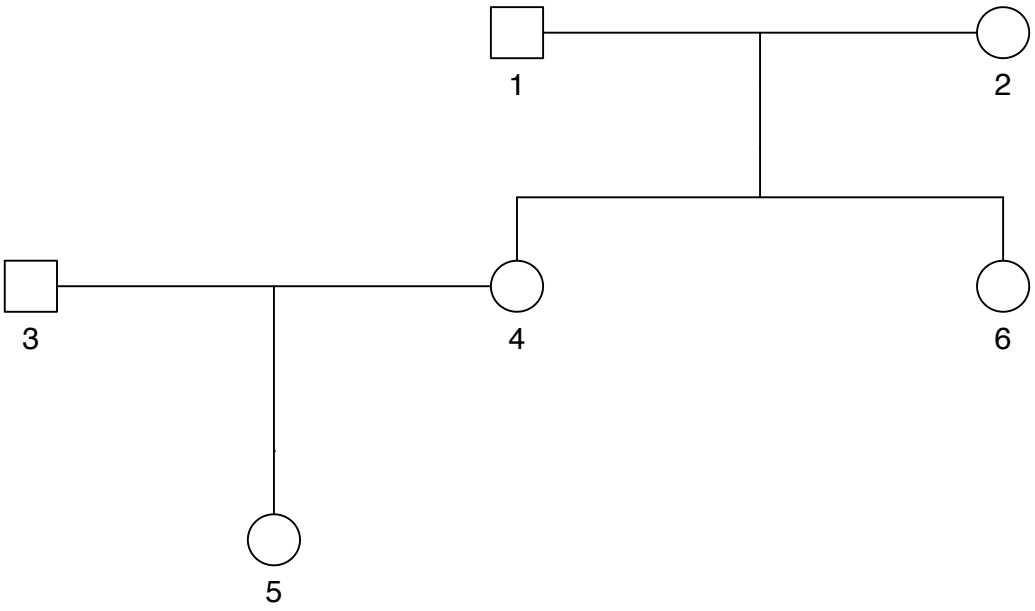

***Figura* 5.1:** Genealogía de los animales para el ejercicio de modelo animal con medidas repetidas.
Fuente: elaboración propia generada en R-project [25].

Información de producciones:

```
Prod=data.frame(matrix(ncol=3,byrow=TRUE,c(
  "4","1",145,
  "4","2",153,
  "5","1",128,
  "5","3",131,
  "6","2",172,
  "6","3",172)))
colnames(Prod)=c("id","GC","Prot")
Prod$Prot=as.numeric(Prod$Prot)
Prod$GC=as.numeric(Prod$GC)
Prod$VacaGC=paste0(Prod$id,"-",Prod$GC)
```

```
Prod
```

```
##   id GC Prot VacaGC
## 1  4  1  145    4-1
## 2  4  2  153    4-2
## 3  5  1  128    5-1
## 4  5  3  131    5-3
## 5  6  2  172    6-2
## 6  6  3  172    6-3
```

Base completa:

```
General=merge(Genealogia, Prod,all=T)
General
```

```
##   id sire  dam sex GC Prot VacaGC
## 1  1 <NA> <NA>   1 NA   NA   <NA>
## 2  2 <NA> <NA>   2 NA   NA   <NA>
## 3  3 <NA> <NA>   1 NA   NA   <NA>
## 4  4    1    2   2  1  145    4-1
## 5  4    1    2   2  2  153    4-2
## 6  5    3    4   2  1  128    5-1
## 7  5    3    4   2  3  131    5-3
## 8  6    1    2   2  2  172    6-2
## 9  6    1    2   2  3  172    6-3
```

Número de datos:

```
n=nrow(General)
n
```

```
## [1] 9
```

Montaje de las matrices X, Z y S y el vector y:

```
library(MatrixModels)

X=as.matrix(model.Matrix(~ as.factor(Prod$GC) -1))
rownames(X)=Prod$VacaGC
colnames(X)=c("GC_1","GC_2","GC_3")
X
```

```
##      GC_1 GC_2 GC_3
## 4-1     1    0    0
## 4-2     0    1    0
## 5-1     1    0    0
## 5-3     0    0    1
## 6-2     0    1    0
## 6-3     0    0    1
```

```
Pro=as.matrix(model.Matrix(~ as.factor(Prod$id) -1))
library(stringr)
colnames(Pro)=word(colnames(Pro), 2, sep = fixed(')'))
Pro
```

```
##   4 5 6
## 1 1 0 0
## 2 1 0 0
## 3 0 1 0
## 4 0 1 0
## 5 0 0 1
## 6 0 0 1
```

```
rownames(Pro)=Prod$id
Pro
```

```
##   4 5 6
## 4 1 0 0
## 4 1 0 0
## 5 0 1 0
## 5 0 1 0
## 6 0 0 1
## 6 0 0 1
```

```
Z=matrix(nrow=nrow(Prod),ncol=ncol(A),0)
colnames(Z)=colnames(A)
rownames(Z)=rownames(Pro)
Z[,colnames(Pro)]=Pro
Z
```

```
##   1 2 3 4 5 6
## 4 0 0 0 1 0 0
## 4 0 0 0 1 0 0
## 5 0 0 0 0 1 0
## 5 0 0 0 0 1 0
## 6 0 0 0 0 0 1
## 6 0 0 0 0 0 1
```

```
S=Pro
S
```

```
##    4 5 6
## 4 1 0 0
## 4 1 0 0
## 5 0 1 0
## 5 0 1 0
## 6 0 0 1
## 6 0 0 1
```

```
y=as.matrix(Prod$Prot)
colnames(y)=c("Prot")
rownames(y)=Prod$VacaGC
y
```

```
##      Prot
## 4-1   145
## 4-2   153
## 5-1   128
## 5-3   131
## 6-2   172
## 6-3   172
```

Montaje del sistema de ecuaciones:

```
XpX=t(X)%*%X
XpX
```

```
##       GC_1 GC_2 GC_3
## GC_1     2    0    0
## GC_2     0    2    0
## GC_3     0    0    2
```

```
XpZ=t(X)%*%Z
XpZ
```

```
##       1 2 3 4 5 6
## GC_1 0 0 0 1 1 0
## GC_2 0 0 0 1 0 1
## GC_3 0 0 0 0 1 1
```

```
ZpX=t(Z)%*%X
```

```
ZpZ=t(Z)%*%Z
ZpZ
```

```
##   1 2 3 4 5 6
## 1 0 0 0 0 0 0
## 2 0 0 0 0 0 0
## 3 0 0 0 0 0 0
## 4 0 0 0 2 0 0
## 5 0 0 0 0 2 0
## 6 0 0 0 0 0 2
```

```
XpS=t(X)%*%S
XpS
```

```
##      4 5 6
## GC_1 1 1 0
## GC_2 1 0 1
## GC_3 0 1 1
```

```
SpX=t(S)%*%X
```

```
ZpS=t(Z)%*%S
ZpS
```

```
##   4 5 6
## 1 0 0 0
## 2 0 0 0
## 3 0 0 0
## 4 2 0 0
## 5 0 2 0
## 6 0 0 2
```

```
SpZ=t(S)%*%Z
```

```
SpS=t(S)%*%S
SpS
```

```
##   4 5 6
## 4 2 0 0
## 5 0 2 0
## 6 0 0 2
```

```
Xpy=t(X)%*%y
Xpy
```

```
##       Prot
## GC_1  273
## GC_2  325
## GC_3  303
```

```
Zpy=t(Z)%*%y
Zpy
```

```
##   Prot
## 1    0
## 2    0
## 3    0
## 4  298
## 5  259
## 6  344
```

```
Spy=t(S)%*%y
Spy
```

```
##   Prot
## 4  298
## 5  259
## 6  344
```

Inclusión del valor de alfa:

```
h2=0.15
rep=0.38
alfa1=(1-rep)/h2
alfa1
```

```
## [1] 4.1
```

```
alfa2=(1-rep)/(rep-h2)
alfa2
```

```
## [1] 2.7
```

```
library(MASS)
Ainv=ginv(A)
round(Ainv,2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    2    1  0.0 -1.0    0   -1
## [2,]    1    2  0.0 -1.0    0   -1
## [3,]    0    0  1.5  0.5   -1    0
## [4,]   -1   -1  0.5  2.5   -1    0
## [5,]    0    0 -1.0 -1.0    2    0
## [6,]   -1   -1  0.0  0.0    0    2
```

```
alfaAinv=alfa1%x%Ainv
round(alfaAinv,2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]  8.3  4.1  0.0 -4.1  0.0 -4.1
## [2,]  4.1  8.3  0.0 -4.1  0.0 -4.1
## [3,]  0.0  0.0  6.2  2.1 -4.1  0.0
## [4,] -4.1 -4.1  2.1 10.3 -4.1  0.0
## [5,]  0.0  0.0 -4.1 -4.1  8.3  0.0
## [6,] -4.1 -4.1  0.0  0.0  0.0  8.3
```

```
ZpZmasalfaAinv=ZpZ+alfaAinv
round(ZpZmasalfaAinv,2)
```

```
##      1    2    3    4    5    6
## 1  8.3  4.1  0.0 -4.1  0.0 -4.1
## 2  4.1  8.3  0.0 -4.1  0.0 -4.1
## 3  0.0  0.0  6.2  2.1 -4.1  0.0
## 4 -4.1 -4.1  2.1 12.3 -4.1  0.0
## 5  0.0  0.0 -4.1 -4.1 10.3  0.0
## 6 -4.1 -4.1  0.0  0.0  0.0 10.3
```

```
I=matrix(nrow=nrow(t(S)),ncol=ncol(S),0);I
```

```
##      [,1] [,2] [,3]
## [1,]    0    0    0
## [2,]    0    0    0
## [3,]    0    0    0
```

```
diag(I)=1;colnames(I)=rownames(I)=colnames(S)
alfa2I=alfa2%x%I
round(alfa2I,2)
```

```
##      [,1] [,2] [,3]
## [1,]  2.7  0.0  0.0
## [2,]  0.0  2.7  0.0
## [3,]  0.0  0.0  2.7
```

```
SpSalfa2I=SpS+alfa2I
round(SpSalfa2I,2)
```

```
##     4   5   6
## 4 4.7 0.0 0.0
## 5 0.0 4.7 0.0
## 6 0.0 0.0 4.7
```

Construcción de la matriz de coeficientes (lado izquierdo del sistema):

```
Izq=rbind(
  cbind(XpX,XpZ,XpS),
  cbind(ZpX,ZpZmasalfaAinv,ZpS),
  cbind(SpX,SpZ,SpSalfa2I))
Izqinv=solve(Izq)
```

Lado derecho del sistema:

```
Der=rbind(Xpy,Zpy,Spy)
Der
```

```
##      Prot
## GC_1  273
## GC_2  325
## GC_3  303
## 1       0
## 2       0
## 3       0
## 4     298
## 5     259
## 6     344
## 4     298
## 5     259
## 6     344
```

Solución al sistema:

```
Sol=round(Izqinv%*%Der,1)
rownames(Sol)=c(rownames(XpX),paste0("a_",rownames(ZpZ)),
paste0("p_",rownames(SpS)))
Sol
```

```
##        Prot
## GC_1 141.0
## GC_2 158.1
## GC_3 151.4
## a_1    1.0
## a_2    1.0
## a_3   -1.9
## a_4   -0.1
## a_5   -2.9
## a_6    2.9
## p_4   -0.2
## p_5   -5.9
## p_6    6.1
```

Valores genéticos:

```
VG=as.matrix(Sol[paste0("a_",rownames(ZpZ)),])
VG
```

```
##      [,1]
## a_1  1.0
## a_2  1.0
## a_3 -1.9
## a_4 -0.1
## a_5 -2.9
## a_6  2.9
```

Diferencia esperada de progenie:

```
Deps=VG/2
```

Efecto de ambiente permanente:

```
AP=as.matrix(Sol[paste0("p_",rownames(SpS)),])
AP
```

```
##      [,1]
## p_4 -0.2
## p_5 -5.9
## p_6  6.1
```

5.2. Modelo animal con efectos maternos

De acuerdo con Quaas [32] y Mrode [21], algunas características como el peso al destete en el ganado de carne tienen una expresión fenotípica que está influida por el ambiente que provee la madre a sus crías. Los efectos maternos tienen un componente genético y un componente ambiental. Willham [33] aseguró que el componente genético materno se puede particionar en efectos aditivos, dominantes y epistáticos y que el componente ambiental se puede dividir en efectos permanentes y temporales. En este punto resulta de especial importancia destacar que se transmite el componente genético aditivo maternal a toda su descendencia, pero se expresa únicamente cuando las hijas se convierten en madres y cuidan de sus crías.

En concordancia con lo anteriormente expresado, debe reiterarse que el desempeño de un individuo al cuidado de su madre, está determinado por efectos ambientales, el efecto genético del propio individuo (efectos directos) y la habilidad de la madre para cuidarlo. Este cuidado se relaciona con el comportamiento materno (etología) y en el caso de los mamíferos con la producción y calidad de la leche que las madres proporcionan a sus crías. Por consiguiente, se evalúa genéticamente a las madres por el desempeño de sus hijos y a los machos por el de los descendientes de sus hijas.

A manera de síntesis, se puede afirmar que cuando se realizan evaluaciones genéticas para características con influencia materna, debe tenerse en cuenta que el desempeño de un individuo en la etapa en que está al cuidado de su madre depende del ambiente, del efecto genético directo (proporcionado por el genotipo heredado de su padre y madre) y el efecto genético materno (genética de la madre, proporcionada por los abuelos maternos), así que en estos casos es indispensable disponer de la información genealógica correspondiente a los abuelos maternos de las crías, para estimar el efecto genético materno.

Adicionalmente, se requiere considerar la existencia de la correlación genética, entre el efecto genético directo y el efecto materno. Cuando se evalúan varios partos de una misma hembra, se debe incluir, además, el efecto ambiental permanente de la madre.

Según Quintanilla y Piedrafita [34], la influencia materna en la expresión de los fenotipos es un efecto estrictamente ambiental en relación a la descendencia, constituyendo tan solo un componente del valor fenotípico de ella.

En la FIGURA NRO. 5.2 se representa de manera gráfica este efecto genético, de especial importancia en los mamíferos.

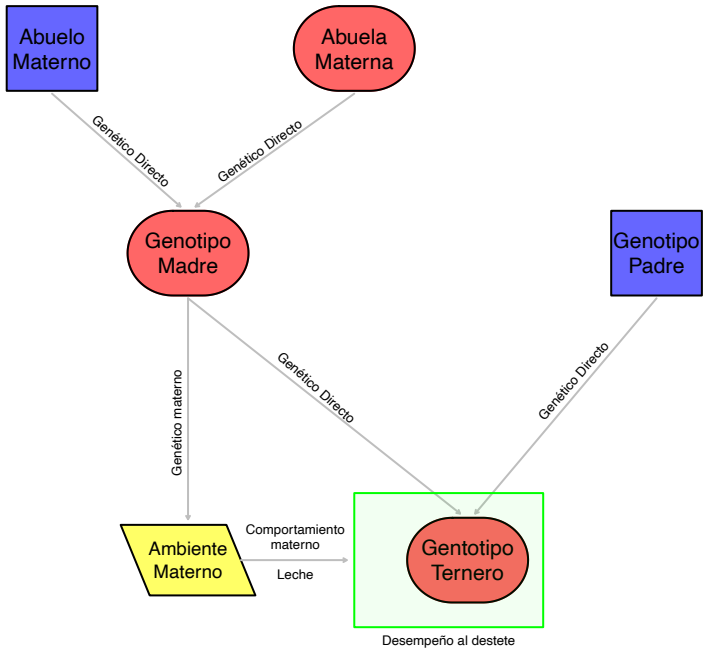

***Figura* 5.2:** Representación gráfica del efecto materno sobre la expresión fenotípica de características en mamíferos.
Fuente: Adaptado de Cerón-Muñoz et al [35]

Según Mrode [21], el modelo animal para características influenciadas por efectos maternos se expresa de la siguiente manera:

$$y = X\beta + Za + Wm + Sp + e$$

Donde:

y= vector de observaciones

β=vector de efectos fijos

a=vector de los efectos genéticos aditivos directos

m= vector de efectos genéticos aditivos maternos

p= vector de efectos de ambiente permanente, cuando se tiene información de varios partos de las madres

e= vector de errores

X = matriz de diseño que relaciona los registros con los efectos fijos

Z= matriz de diseño que relaciona los registros con los efectos genéticos aditivos directos

W= matriz de diseño que relaciona los registros con los efectos genéticos aditivos maternos

S= matriz de diseño que relaciona los registros con los efectos aleatorios ambientales permanentes.

Se asume que:

$$Var\begin{bmatrix} a \\ m \\ p \\ s \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A & 0 & 0 \\ g_{21}A & g_{22}A & 0 & 0 \\ 0 & 0 & \sigma_p^2 I & 0 \\ 0 & 0 & 0 & \sigma_e^2 I \end{bmatrix}$$

Donde:

g_{11}= varianza genética aditiva directa

g_{12}= covarianza entre los efectos aditivos directos y los genéticos aditivos maternos

g_{22}= Varianza genética aditiva materna

p_2 = varianza debida a los efectos ambientales permanentes

σ_e^2= varianza del error.

La varianza fenotípica, de acuerdo con Mrode [21], adopta la siguiente forma:

$$Var[y] = \begin{bmatrix} Z \\ W \end{bmatrix}\begin{bmatrix} g_{11}A & g_{12}A \\ g_{21}A & g_{22}A \end{bmatrix}[Z^T \quad W^T] + \sigma_p^2 SS^T + \sigma_e^2 I$$

Como lo indica Mrode [21], bajo este modelo el fenotipo se particiona en los siguientes elementos:

1) Efectos que transmiten el padre y la madre a sus descendientes y que se denominan efectos genéticos aditivos

2) Habilidad genética aditiva de la madre para proporcionar un ambiente favorable y que se denomina efecto indirecto o efecto genético materno

3) Efectos que incluyen la influencia ambiental permanente relacionada con la habilidad materna para proporcionar un ambiente favorable a sus crías, más los efectos genéticos no aditivos que transmiten las madres

4) Otros efectos ambientales y genéticos que se encuentran en el vector de errores

El sistema de ecuaciones del modelo mixto tendrá la siguiente estructura:

$$\begin{bmatrix} X^TX & X^TZ & X^TW & X^TS \\ Z^TX & Z^TZ+\alpha_1A^{-1} & Z^TW+\alpha_2A^{-1} & Z^TS \\ W^TX & W^TZ+\alpha_2A^{-1} & W^TW+\alpha_3A^{-1} & W^TS \\ S^TX & S^TZ & S^TW & S^TS+\alpha_4I \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \\ \hat{m} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X^TY \\ Z^TY \\ W^TY \\ S^TY \end{bmatrix}$$

Los valores de los α serían:

$$\begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{bmatrix} = \sigma_e^2 G_0^{-1}$$

$$G_0 = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

$$G_0^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix}$$

$$\alpha_4 = \frac{\sigma_e^2}{\sigma_p^2}$$

Para facilitar la comprensión del modelo, la construcción del sistema de ecuaciones del modelo mixto, los resultados y sus interpretaciones, en el presente capítulo se utilizarán datos arbitrarios del peso al destete en bovinos de carne. Para desarrollar el EJEMPLO 5.2 correspondiente a este modelo se toma la información mínima requerida para desarrollar los procedimientos de cálculo y análisis, es decir el peso al destete en kilogramos de bovinos de carne, con dos padres, dos madres y cuatro crías y considerando el sexo como el único efecto fijo (ver TABLA NRO. 5.2).

Para calcular los distintos valores de α se toman los siguientes valores:

$$G_0 = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} 100 & -60 \\ -60 & 70 \end{bmatrix}$$

$$G_0^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix} = \begin{bmatrix} 0.02058 & 0.01764 \\ 0.01764 & 0.02941 \end{bmatrix}$$

$$\begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{bmatrix} = \sigma_e^2 \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix} = 300 \begin{bmatrix} 0.02058 & 0.01764 \\ 0.01764 & 0.02941 \end{bmatrix} = \begin{bmatrix} 6.7941 & 5.8235 \\ 5.8235 & 9.7058 \end{bmatrix}$$

Tabla 5.2: Información genealógica y peso al destete (kg) de una población vacuna para carne

Animal	Padre	Madre	Sexo	Peso Destete
1			1	
2			2	
3			1	
4			2	
5			1	
6			2	
7			1	
8			2	
9	1	2	1	
10	3	4	2	
11	5	6	1	
12	7	8	2	
13	9	10	1	180
14	11	10	2	160
15	11	12	1	190
16	9	12	2	100

Fuente: elaboración propia (2024).

$$\alpha_4 = \frac{\sigma_e^2}{\sigma_p^2} = \frac{300}{80} = 4.125$$

Para estimar las medias de cada sexo y determinar el valor genético de los animales, considerando el efecto genético aditivo directo además, del efecto genético aditivo maternal, al igual que el efecto ambiental permanente, que contendrá la estimación de efectos no aditivos de las madres y otros efectos ambientales permanentes, se requiere construir el sistema de ecuaciones correspondientes a este modelo de la siguiente manera:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La matriz X, para el Ejemplo 5.2 se construye representando en las columnas los niveles del efecto fijo sexo que en este caso son dos, machos en la primera columna y hembras en la segunda columna; mientras que en las filas se representan los animales con registro.

$$Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

En todos los modelos la matriz Z se construye de la misma manera, es decir incluyendo los animales con y sin registro que aparecen en el pedigrí.

Para construir la matriz W, en las filas se representan los individuos con registros y en las columnas todos los animales que aparecen en la evaluación tengan o no tengan registro. Se marcará con uno los elementos de la matriz correspondientes a la relación de cada animal con su madre.

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Para construir la matriz S, en las filas se representan los individuos con registros y en las columnas las madres que tienen dos o más progenies con registros. Se marcará con uno los puntos de corte de cada animal con su respectiva madre. Los individuos con registro son hijos de las hembras 10 y 12.

$$S = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

En todos los casos, la matriz A contiene la información correspondiente al valor de las relaciones genéticas aditivas entre todos los animales incluidos en la evaluación. El procedimiento para construirla es exactamente el mismo que se explicó en el modelo básico.

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0.5 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 \\ 0 & 1 & 0 & \cdots & 0 & 0.5 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0.5 & 0 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0.5 & 0 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.25 & 0.25 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.25 & 0.25 & 0.25 & \cdots & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0.25 & 0 & 0.25 \\ 0 & 0 & 0.25 & \cdots & 0 & 0 & 0.5 & 0.5 & 0 & 0.25 & 1 & 0.25 & 0 \\ 0 & 0 & 0 & \cdots & 0.25 & 0 & 0 & 0.5 & 0.5 & 0 & 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 0 & \cdots & 0.25 & 0.5 & 0 & 0 & 0.5 & 0.25 & 0 & 0.25 & 1 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 1.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 1.5 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 1.5 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.5 & 0.5 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

El vector y contiene los datos, así:

$$y = \begin{bmatrix} 180 \\ 160 \\ 190 \\ 100 \end{bmatrix}$$

El lado izquierdo del sistema de ecuaciones del modelo mixto (LHS), contiene la siguiente información, obtenida de cada una de las matrices que se indicaron anteriormente:

$$X^T X = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$X^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$W^T X = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$S^T X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$Z^T W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$S^T W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$W^T W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S^T S = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$Z^T Z + \alpha_1 A^{-1} = \begin{bmatrix} 10.19 & 3.4 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 3.4 & 10.19 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10.19 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.4 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 14.59 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 14.59 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -6.79 & 0 & 0 & 14.59 & 0 \\ 0 & 0 & 0 & \cdots & -6.79 & 0 & 0 & 0 & 14.59 \end{bmatrix}$$

$$Z^T W + \alpha_2 A^{-1} = \begin{bmatrix} 8.74 & 2.91 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 2.91 & 8.74 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8.74 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.91 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -5.82 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 11.65 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -5.82 & 0 & 0 & 11.65 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -5.82 & -4.82 & 0 & 0 & 11.65 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -4.82 & 0 & 0 & 0 & 11.65 \end{bmatrix}$$

$$W^TW+\alpha_3A^{-1}=\begin{bmatrix} 14.56 & 4.85 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 4.85 & 14.56 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 14.56 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4.85 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -9.71 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 19.41 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -9.71 & 0 & 0 & 19.41 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -9.71 & -9.71 & 0 & 0 & 19.41 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -9.71 & 0 & 0 & 0 & 19.41 \end{bmatrix}$$

$$S^TS+\alpha_4I=\begin{bmatrix} 6.1 & 0 \\ 0 & 6.1 \end{bmatrix}$$

Para construir el lado derecho del sistema de ecuaciones del modelo mixto, deben realizarse las siguientes operaciones matriciales:

$$X^Ty=\begin{bmatrix} 370 \\ 260 \end{bmatrix}$$

$$Z^Ty=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 180 \\ 160 \\ 190 \\ 100 \end{bmatrix}$$

$$W^Ty=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 340 \\ 0 \\ 290 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$S^T y = \begin{bmatrix} 340 \\ 290 \end{bmatrix}$$

Las soluciones requeridas se obtienen al multiplicar la matriz inversa del LHS, por el vector del lado derecho del sistema de ecuaciones del modelo mixto (RHS).

$$\begin{bmatrix} \widehat{S_1} \\ \widehat{S_2} \\ \widehat{a_1} \\ \widehat{a_2} \\ \vdots \\ \widehat{m_1} \\ \widehat{m_2} \\ \vdots \\ \widehat{p_{10}} \\ \widehat{p_{12}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 10.19 & 3.4 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.4 & 10.19 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10.19 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 19.41 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 19.41 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 6.12 & 0 \\ 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 6.12 \end{bmatrix}^{-1} \begin{bmatrix} 370 \\ 260 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 340 \\ 290 \end{bmatrix}$$

Los resultados del vector solución se explican de la siguiente manera: los dos primeros valores corresponden a los promedios de los dos sexos e indican que los machos tienen un valor numéricamente mayor al de las hembras.

$$\begin{bmatrix} \widehat{S_1} \\ \widehat{S_2} \end{bmatrix} = \begin{bmatrix} 185 \\ 130 \end{bmatrix}$$

Los valores genéticos aditivos se obtienen para los ocho animales que aparecen en la base de datos del EJEMPLO 5.2, es decir para la población base y su descendencia con registros.

$$\begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \vdots \\ \widehat{a_{13}} \\ \widehat{a_{14}} \\ \widehat{a_{15}} \\ \widehat{a_{16}} \end{bmatrix} = \begin{bmatrix} -2.03 \\ -2.03 \\ -0.20 \\ -0.20 \\ 2.03 \\ \vdots \\ -3.25 \\ 4.89 \\ 3.25 \\ -4.89 \end{bmatrix}$$

Posteriormente aparece el valor genético aditivo materno.

$$\begin{bmatrix} \widehat{m_1} \\ \widehat{m_2} \\ \widehat{m_3} \\ \widehat{m_4} \\ \widehat{m_5} \\ \vdots \\ \widehat{m_{13}} \\ \widehat{m_{14}} \\ \widehat{m_{15}} \\ \widehat{m_{16}} \end{bmatrix} = \begin{bmatrix} 1.22 \\ 1.22 \\ 0.82 \\ 0.82 \\ -1.22 \\ \vdots \\ 2.65 \\ -2.24 \\ -2.65 \\ 2.23 \end{bmatrix}$$

Finalmente, los dos últimos valores en el vector solución contienen las cifras que corresponden al efecto ambiental permanente:

$$\begin{bmatrix} \widehat{p_{10}} \\ \widehat{p_{12}} \end{bmatrix} = \begin{bmatrix} 3.28 \\ -3.28 \end{bmatrix}$$

La utilidad práctica de este modelo consiste, fundamentalmente, en ofrecer la posibilidad de seleccionar los reproductores por línea materna que se escogerán por el mayor mérito genético aditivo directo y aditivo maternal. En el EJEMPLO 5.2 es claro que se selecciona la hembra 14 por efecto directo y la 13 por efecto genético materno (TABLA NRO. 5.3).

TABLA 5.3: Valoración genética de peso al destete (kg) de una población bovina de carne

Id	Directo	Materno	Ambiente permanente
14	4.89	-2.24	
11	4.07	-2.44	
15	3.25	-2.65	
5	2.03	-1.22	
6	2.03	-1.22	
12	0.41	-1.64	-3.28
8	0.20	-0.82	
7	0.20	-0.82	
3	-0.20	0.82	
4	-0.20	0.82	
10	-0.41	1.64	3.28
1	-2.03	1.22	
2	-2.03	1.22	
13	-3.25	2.65	
9	-4.07	2.44	
16	-4.89	2.24	

Fuente: elaboración propia (2024).

La programación en R-project para el EJEMPLO 5.2 se desarrolla de la siguiente manera:

5.2.1. Ejercicios en R-project

```
Genealogia=data.frame(matrix(ncol=4,byrow=TRUE,c(
1,NA,NA,1,
2,NA,NA,2,
3,NA,NA,1,
4,NA,NA,2,
5,NA,NA,1,
6,NA,NA,2,
7,NA,NA,1,
8,NA,NA,2,
9,1,2,1,
10,3,4,2,
11,5,6,1,
12,7,8,2,
13,9,10,1,
14,11,10,2,
15,11,12,1,
16,9,12,2
)))
colnames(Genealogia)=c("id","sire","dam","sex")
```

```
Genealogia
```

```
##    id sire dam sex
## 1   1   NA  NA   1
## 2   2   NA  NA   2
## 3   3   NA  NA   1
## 4   4   NA  NA   2
## 5   5   NA  NA   1
## 6   6   NA  NA   2
## 7   7   NA  NA   1
## 8   8   NA  NA   2
## 9   9    1   2   1
## 10 10    3   4   2
## 11 11    5   6   1
## 12 12    7   8   2
## 13 13    9  10   1
## 14 14   11  10   2
## 15 15   11  12   1
## 16 16    9  12   2
```

Relaciones entre padres y madres:

```
table(Genealogia$sire,Genealogia$dam)
```

```
##
##       2 4 6 8 10 12
##    1  1 0 0 0  0  0
##    3  0 1 0 0  0  0
##    5  0 0 1 0  0  0
##    7  0 0 0 1  0  0
##    9  0 0 0 0  1  1
##    11 0 0 0 0  1  1
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id = Genealogia$id, dadid = Genealogia$sire,
                momid = Genealogia$dam,
                sex=as.numeric(Genealogia$sex))
```

Montaje de la matriz de parentesco:

```
A=2*kinship(Genealogia$id,Genealogia$sire,Genealogia$dam)
A[8:16,8:16]
```

```
##       8   9  10  11  12   13   14   15   16
## 8  1.00 0.0 0.0 0.0 0.5 0.00 0.00 0.25 0.25
## 9  0.00 1.0 0.0 0.0 0.0 0.50 0.00 0.00 0.50
## 10 0.00 0.0 1.0 0.0 0.0 0.50 0.50 0.00 0.00
## 11 0.00 0.0 0.0 1.0 0.0 0.00 0.50 0.50 0.00
## 12 0.50 0.0 0.0 0.0 1.0 0.00 0.00 0.50 0.50
## 13 0.00 0.5 0.5 0.0 0.0 1.00 0.25 0.00 0.25
## 14 0.00 0.0 0.5 0.5 0.0 0.25 1.00 0.25 0.00
## 15 0.25 0.0 0.0 0.5 0.5 0.00 0.25 1.00 0.25
## 16 0.25 0.5 0.0 0.0 0.5 0.25 0.00 0.25 1.00
```

```
bitSize(Genea1)


## $bitSize
## [1] 8
##
## $nFounder
## [1] 8
##
## $nNonFounder
## [1] 8
```

Utilizando el comando «plot» obtendremos el árbol genealógico (FIGURA NRO. 5.3):

```
plot(Genea1,mar=c(bottom=0, left=1, top=1, right=1))
```

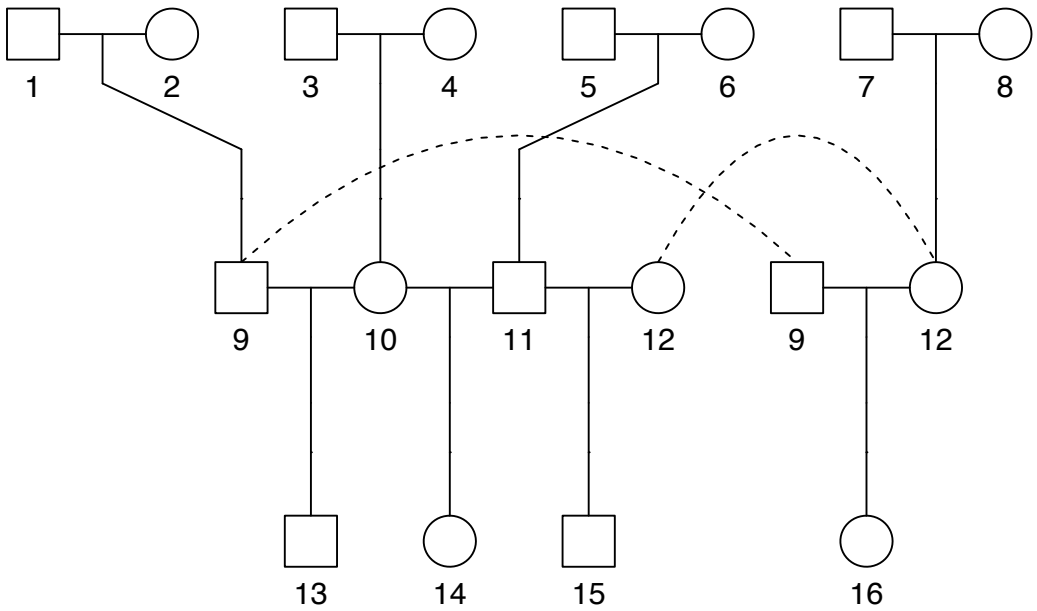

***Figura* 5.3:** Genealogía de los 16 animales para el ejercicio de modelo animal con efecto genético materno.
Fuente: elaboración propia generada en R-project [25].

Información de producciones:

```
Prod=data.frame(matrix(ncol=4,byrow=TRUE,c(
13,10,1,180,
14,10,2,160,
15,12,1,190,
16,12,2,100
)))
colnames(Prod)=c("id","dam","GSx","PD")
Prod$PD=as.numeric(Prod$PD)
Prod
```

```
##   id dam GSx  PD
## 1 13  10   1 180
## 2 14  10   2 160
## 3 15  12   1 190
## 4 16  12   2 100
```

Número de animales:

```
n=nrow(Genealogia)
n
```

```
## [1] 16
```

Montaje de las matrices X, Z y W y el vector y:

```
library(MatrixModels)

X=as.matrix(model.Matrix(~ as.factor(Prod$GSx) -1))
rownames(X)=Prod$id
colnames(X)=c("GSx_1","GSx_2")
X
```

```
##    GSx_1 GSx_2
## 13     1     0
## 14     0     1
## 15     1     0
## 16     0     1
```

```
Pro=as.matrix(model.Matrix(~ as.factor(Prod$id) -1))
library(stringr)
colnames(Pro)=word(colnames(Pro), 2, sep = fixed(')'))
rownames(Pro)=Prod$id
Pro
```

```
##    13 14 15 16
## 13  1  0  0  0
## 14  0  1  0  0
## 15  0  0  1  0
## 16  0  0  0  1
```

```
Z=matrix(nrow=nrow(Pro),ncol=ncol(A),0)
colnames(Z)=colnames(A)
rownames(Z)=colnames(Pro)
Z[colnames(Pro),colnames(Pro)]=Pro
Z[,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 13 0 0  0  0  0  1  0  0  0
## 14 0 0  0  0  0  0  1  0  0
## 15 0 0  0  0  0  0  0  1  0
## 16 0 0  0  0  0  0  0  0  1
```

```
AmbMat=as.matrix(model.Matrix(~ as.factor(Prod$dam) -1))
```

```
colnames(AmbMat)=word(colnames(AmbMat), 2, sep = fixed(')'))
rownames(AmbMat)=Prod$id
AmbMat
```

```
##    10 12
## 13  1  0
## 14  1  0
## 15  0  1
## 16  0  1
```

```
W=matrix(nrow=nrow(AmbMat),ncol=ncol(A),0)
colnames(W)=colnames(A)
rownames(W)=rownames(AmbMat)
W[rownames(AmbMat),colnames(AmbMat)]=AmbMat
W[,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 13 0 0  1  0  0  0  0  0  0
## 14 0 0  1  0  0  0  0  0  0
## 15 0 0  0  0  1  0  0  0  0
## 16 0 0  0  0  1  0  0  0  0
```

```
AmbMat
```

```
##    10 12
## 13  1  0
## 14  1  0
## 15  0  1
## 16  0  1
```

```
S=AmbMat
S
```

```
##    10 12
## 13  1  0
## 14  1  0
## 15  0  1
## 16  0  1
```

```
y=as.matrix(Prod$PD)
colnames(y)=c("PD")
rownames(y)=Prod$VacaGSx
y
```

```
##        PD
## [1,] 180
## [2,] 160
## [3,] 190
## [4,] 100
```

Montaje del sistema de ecuaciones:

```
XpX=t(X)%*%X
XpX
```

```
##        GSx_1 GSx_2
## GSx_1      2     0
## GSx_2      0     2
```

```
XpZ=t(X)%*%Z
XpZ
```

```
##        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
## GSx_1 0 0 0 0 0 0 0 0 0  0  0  0  1  0  1  0
## GSx_2 0 0 0 0 0 0 0 0 0  0  0  0  0  1  0  1
```

```
ZpX=t(Z)%*%X
```

```
ZpZ=t(Z)%*%Z
ZpZ[8:16,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 8  0 0  0  0  0  0  0  0  0
## 9  0 0  0  0  0  0  0  0  0
## 10 0 0  0  0  0  0  0  0  0
## 11 0 0  0  0  0  0  0  0  0
## 12 0 0  0  0  0  0  0  0  0
## 13 0 0  0  0  0  1  0  0  0
## 14 0 0  0  0  0  0  1  0  0
## 15 0 0  0  0  0  0  0  1  0
## 16 0 0  0  0  0  0  0  0  1
```

```
XpW=t(X)%*%W
XpW
```

```
##        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
## GSx_1 0 0 0 0 0 0 0 0 0  1  0  1  0  0  0  0
## GSx_2 0 0 0 0 0 0 0 0 0  1  0  1  0  0  0  0
```

```
WpX=t(W)%*%X
WpX
```

```
##     GSx_1 GSx_2
## 1       0     0
## 2       0     0
## 3       0     0
## 4       0     0
## 5       0     0
## 6       0     0
## 7       0     0
## 8       0     0
## 9       0     0
## 10      1     1
## 11      0     0
## 12      1     1
## 13      0     0
## 14      0     0
## 15      0     0
## 16      0     0
```

```
XpS=t(X)%*%S
XpS
```

```
##        10 12
## GSx_1   1  1
## GSx_2   1  1
```

```
SpX=t(S)%*%X
```

```
ZpW=t(Z)%*%W
ZpW[8:16,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 8  0 0  0  0  0  0  0  0  0
## 9  0 0  0  0  0  0  0  0  0
## 10 0 0  0  0  0  0  0  0  0
## 11 0 0  0  0  0  0  0  0  0
## 12 0 0  0  0  0  0  0  0  0
## 13 0 0  1  0  0  0  0  0  0
## 14 0 0  1  0  0  0  0  0  0
## 15 0 0  0  0  1  0  0  0  0
## 16 0 0  0  0  1  0  0  0  0
```

```
WpZ=t(W)%*%Z
```

```
ZpS=t(Z)%*%S
ZpS
```

```
##    10 12
## 1   0  0
## 2   0  0
## 3   0  0
## 4   0  0
## 5   0  0
## 6   0  0
## 7   0  0
## 8   0  0
## 9   0  0
## 10  0  0
## 11  0  0
## 12  0  0
## 13  1  0
## 14  1  0
## 15  0  1
## 16  0  1
```

```
SpZ=t(S)%*%Z
```

```
WpS=t(W)%*%S
WpS
```

```
##    10 12
## 1   0  0
## 2   0  0
## 3   0  0
## 4   0  0
## 5   0  0
## 6   0  0
## 7   0  0
## 8   0  0
## 9   0  0
## 10  2  0
## 11  0  0
## 12  0  2
## 13  0  0
## 14  0  0
```

```
## 15  0  0
## 16  0  0
```

```
SpW=t(S)%*%W
SpW[,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 10 0 0  2  0  0  0  0  0  0
## 12 0 0  0  0  2  0  0  0  0
```

```
WpW=t(W)%*%W
WpW[8:16,8:16]
```

```
##    8 9 10 11 12 13 14 15 16
## 8  0 0  0  0  0  0  0  0  0
## 9  0 0  0  0  0  0  0  0  0
## 10 0 0  2  0  0  0  0  0  0
## 11 0 0  0  0  0  0  0  0  0
## 12 0 0  0  0  2  0  0  0  0
## 13 0 0  0  0  0  0  0  0  0
## 14 0 0  0  0  0  0  0  0  0
## 15 0 0  0  0  0  0  0  0  0
## 16 0 0  0  0  0  0  0  0  0
```

```
SpS=t(S)%*%S
SpS
```

```
##    10 12
## 10  2  0
## 12  0  2
```

```
Xpy=t(X)%*%y
Xpy
```

```
##         PD
## GSx_1 370
## GSx_2 260
```

```
Zpy=t(Z)%*%y
Zpy
```

```
##      PD
## 1     0
## 2     0
## 3     0
## 4     0
## 5     0
## 6     0
## 7     0
## 8     0
## 9     0
## 10    0
## 11    0
## 12    0
## 13  180
## 14  160
## 15  190
## 16  100
```

```
Wpy=t(W)%*%y
Wpy
```

```
##      PD
## 1     0
## 2     0
## 3     0
## 4     0
## 5     0
## 6     0
## 7     0
## 8     0
## 9     0
## 10  340
## 11    0
## 12  290
## 13    0
## 14    0
## 15    0
## 16    0
```

```
Spy=t(S)%*%y
Spy
```

```
##     PD
## 10 340
## 12 290
```

Inclusión de los valores α:

```
var_a=100
var_m=70
cov_am=-60
var_p=80
var_e=330
G=matrix(nrow=2,ncol=2,byrow=TRUE,c(
  var_a,cov_am,
  cov_am,var_m
))
G
```

```
##      [,1] [,2]
## [1,]  100  -60
## [2,]  -60   70
```

```
Ginv=solve(G)
Ginv
```

```
##       [,1]  [,2]
```

```
## [1,] 0.021 0.018
## [2,] 0.018 0.029
```

```
alpha123=var_e%x%Ginv
alpha123
```

```
##      [,1] [,2]
## [1,]  6.8  5.8
## [2,]  5.8  9.7
```

```
alpha4=var_e/var_p
alpha4
```

```
## [1] 4.1
```

```
library(MASS)
Ainv=ginv(A)
round(Ainv,2)[8:16,8:16]
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,]  1.5  0.0  0.0  0.0 -1.0    0    0    0    0
##  [2,]  0.0  3.0  0.5  0.0  0.5   -1    0    0   -1
##  [3,]  0.0  0.5  3.0  0.5  0.0   -1   -1    0    0
##  [4,]  0.0  0.0  0.5  3.0  0.5    0   -1   -1    0
##  [5,] -1.0  0.5  0.0  0.5  3.0    0    0   -1   -1
##  [6,]  0.0 -1.0 -1.0  0.0  0.0    2    0    0    0
##  [7,]  0.0  0.0 -1.0 -1.0  0.0    0    2    0    0
##  [8,]  0.0  0.0  0.0 -1.0 -1.0    0    0    2    0
##  [9,]  0.0 -1.0  0.0  0.0 -1.0    0    0    0    2
```

```
ZpZalfaAinv=ZpZ+alpha123[1,1]%x%Ainv
round(ZpZalfaAinv,1)[8:16,8:16]
```

```
##       8    9   10   11   12   13   14   15   16
## 8  10.2  0.0  0.0  0.0 -6.8  0.0  0.0  0.0  0.0
## 9   0.0 20.4  3.4  0.0  3.4 -6.8  0.0  0.0 -6.8
## 10  0.0  3.4 20.4  3.4  0.0 -6.8 -6.8  0.0  0.0
## 11  0.0  0.0  3.4 20.4  3.4  0.0 -6.8 -6.8  0.0
## 12 -6.8  3.4  0.0  3.4 20.4  0.0  0.0 -6.8 -6.8
## 13  0.0 -6.8 -6.8  0.0  0.0 14.6  0.0  0.0  0.0
## 14  0.0  0.0 -6.8 -6.8  0.0  0.0 14.6  0.0  0.0
## 15  0.0  0.0  0.0 -6.8 -6.8  0.0  0.0 14.6  0.0
## 16  0.0 -6.8  0.0  0.0 -6.8  0.0  0.0  0.0 14.6
```

```
ZpWalfaAinv=ZpW+alpha123[1,2]%x%Ainv
round(ZpWalfaAinv,1)[8:16,8:16]
```

```
##       8    9   10   11   12   13   14   15   16
## 8   8.7  0.0  0.0  0.0 -5.8  0.0  0.0  0.0  0.0
## 9   0.0 17.5  2.9  0.0  2.9 -5.8  0.0  0.0 -5.8
## 10  0.0  2.9 17.5  2.9  0.0 -5.8 -5.8  0.0  0.0
## 11  0.0  0.0  2.9 17.5  2.9  0.0 -5.8 -5.8  0.0
## 12 -5.8  2.9  0.0  2.9 17.5  0.0  0.0 -5.8 -5.8
## 13  0.0 -5.8 -4.8  0.0  0.0 11.6  0.0  0.0  0.0
## 14  0.0  0.0 -4.8 -5.8  0.0  0.0 11.6  0.0  0.0
## 15  0.0  0.0  0.0 -5.8 -4.8  0.0  0.0 11.6  0.0
## 16  0.0 -5.8  0.0  0.0 -4.8  0.0  0.0  0.0 11.6
```

```
WpZalfaAinv=WpZ+alpha123[2,1]%x%Ainv
round(WpZalfaAinv,1)[8:16,8:16]
```

```
##       8    9   10   11   12   13   14   15   16
## 8   8.7  0.0  0.0  0.0 -5.8  0.0  0.0  0.0  0.0
## 9   0.0 17.5  2.9  0.0  2.9 -5.8  0.0  0.0 -5.8
## 10  0.0  2.9 17.5  2.9  0.0 -4.8 -4.8  0.0  0.0
## 11  0.0  0.0  2.9 17.5  2.9  0.0 -5.8 -5.8  0.0
## 12 -5.8  2.9  0.0  2.9 17.5  0.0  0.0 -4.8 -4.8
## 13  0.0 -5.8 -5.8  0.0  0.0 11.6  0.0  0.0  0.0
## 14  0.0  0.0 -5.8 -5.8  0.0  0.0 11.6  0.0  0.0
## 15  0.0  0.0  0.0 -5.8 -5.8  0.0  0.0 11.6  0.0
## 16  0.0 -5.8  0.0  0.0 -5.8  0.0  0.0  0.0 11.6
```

```
WpWalfaAinv=WpW+alpha123[2,2]%x%Ainv
round(WpWalfaAinv,1)[8:16,8:16]
```

```
##       8    9   10   11   12   13   14   15   16
## 8  14.6  0.0  0.0  0.0 -9.7  0.0  0.0  0.0  0.0
## 9   0.0 29.1  4.9  0.0  4.9 -9.7  0.0  0.0 -9.7
## 10  0.0  4.9 31.1  4.9  0.0 -9.7 -9.7  0.0  0.0
## 11  0.0  0.0  4.9 29.1  4.9  0.0 -9.7 -9.7  0.0
## 12 -9.7  4.9  0.0  4.9 31.1  0.0  0.0 -9.7 -9.7
## 13  0.0 -9.7 -9.7  0.0  0.0 19.4  0.0  0.0  0.0
## 14  0.0  0.0 -9.7 -9.7  0.0  0.0 19.4  0.0  0.0
## 15  0.0  0.0  0.0 -9.7 -9.7  0.0  0.0 19.4  0.0
## 16  0.0 -9.7  0.0  0.0 -9.7  0.0  0.0  0.0 19.4
```

```
I=matrix(nrow=nrow(t(S)),ncol=ncol(S),0);I
```

```
##      [,1] [,2]
## [1,]    0    0
## [2,]    0    0
```

```
diag(I)=1;colnames(I)=rownames(I)=colnames(S)
Ialpha4=I%x%alpha4
Ialpha4
```

```
##      [,1] [,2]
## [1,]  4.1  0.0
## [2,]  0.0  4.1
```

```
SpSalfa4I=SpS+Ialpha4
round(SpSalfa4I,2)


##      10  12
## 10 6.1 0.0
## 12 0.0 6.1
```

Construcción del lado izquierdo del sistema:

```
Izq=rbind(
  cbind(XpX,XpZ,XpW,XpS),
  cbind(ZpX,ZpZalfaAinv,ZpWalfaAinv,ZpS),
  cbind(WpX,WpZalfaAinv,WpWalfaAinv,WpS),
  cbind(SpX,SpZ,SpW,SpSalfa4I))
Izqinv=solve(Izq)
```

Lado derecho del sistema:

```
Der=rbind(Xpy,Zpy,Wpy,Spy)
Der
##         PD
## GSx_1  370
## GSx_2  260
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
## 7        0
## 8        0
## 9        0
## 10       0
## 11       0
## 12       0
## 13     180
## 14     160
## 15     190
## 16     100
## 1        0
## 2        0
## 3        0
## 4        0
```

```
## 5        0
## 6        0
## 7        0
## 8        0
## 9        0
## 10     340
## 11       0
## 12     290
## 13       0
## 14       0
## 15       0
## 16       0
## 10     340
## 12     290
```

Solución al sistema:

```
Sol=Izqinv%*%Der
rownames(Sol)=c(rownames(XpX),paste0("a_",rownames(ZpZ)),
paste0("m_",rownames(WpW)),paste0("p_",rownames(SpS)))
round(Sol,1)
##           PD
## GSx_1  185.0
## GSx_2  130.0
## a_1     -2.0
## a_2     -2.0
## a_3     -0.2
## a_4     -0.2
## a_5      2.0
## a_6      2.0
## a_7      0.2
## a_8      0.2
## a_9     -4.1
## a_10    -0.4
## a_11     4.1
## a_12     0.4
## a_13    -3.3
## a_14     4.9
## a_15     3.3
## a_16    -4.9
## m_1      1.2
## m_2      1.2
## m_3      0.8
## m_4      0.8
```

```
## m_5    -1.2
## m_6    -1.2
## m_7    -0.8
## m_8    -0.8
## m_9     2.4
## m_10    1.6
## m_11   -2.4
## m_12   -1.6
## m_13    2.6
## m_14   -2.2
## m_15   -2.6
## m_16    2.2
## p_10    3.3
## p_12   -3.3
```

```
Sexo=as.matrix(Sol[rownames(XpX),])
Sexo
```

```
##       [,1]
## GSx_1  185
## GSx_2  130
```

```
Directo=as.matrix(Sol[paste0("a_",rownames(ZpZ)),])
Directo
```

```
##       [,1]
## a_1  -2.03
## a_2  -2.03
## a_3  -0.20
## a_4  -0.20
## a_5   2.03
## a_6   2.03
## a_7   0.20
## a_8   0.20
## a_9  -4.07
## a_10 -0.41
## a_11  4.07
## a_12  0.41
## a_13 -3.25
## a_14  4.89
## a_15  3.25
## a_16 -4.89
```

```
Materno=as.matrix(Sol[paste0("m_",rownames(WpW)),])
Materno
```

```
##        [,1]
## m_1    1.22
## m_2    1.22
## m_3    0.82
## m_4    0.82
## m_5   -1.22
## m_6   -1.22
## m_7   -0.82
## m_8   -0.82
## m_9    2.44
## m_10   1.64
## m_11  -2.44
## m_12  -1.64
## m_13   2.65
## m_14  -2.24
## m_15  -2.65
## m_16   2.24
```

```
Permanente=as.matrix(Sol[paste0("p_",rownames(SpS)),])
Permanente
```

```
##       [,1]
## p_10   3.3
## p_12  -3.3
```

5.3. Modelo animal con efectos ambientales comunes

En ciertos casos, los animales se desempeñan en un ambiente común que contribuye a incrementar la similitud entre los miembros de una misma familia o grupo de indivíduos que comparten este ambiente, como en el caso de los cerdos, conejos, perros, gatos y cuyes, especies en las que los nacimientos se presentan en camada, dando lugar a la existencia de una covarianza adicional entre los miembros de la misma familia y al incremento de la varianza entre las distintas familias (Mrode [21]), por lo tanto, las evaluaciones genéticas tendrán como finalidad la predicción del valor genético y la estimación del efecto ambiental común. En el caso de especies como el cerdo, a menudo se hace crianza cruzada, caso en el cual, se mezclan animales de diferente madre para crear camadas homogéneas de tamaño. Así,los individuos

que comparten el ambiente no necesariamente son familiares. Otro ambiente común es el que se presenta, en los casos donde los individuos tienen una participación social cuando comparten un ambiente determinado, como un corral, una jaula o una caballeriza, entre otras.

El modelo en notación matricial para estos casos se define de la siguiente manera:

$$y = X\beta + Za + Cc + e$$

Donde:

y= vector de observaciones.

β= vector de efectos fijos.

a= vector de efectos aleatorios del animal (valores genéticos aditivos directos).

c= vector de efectos aleatorios de ambiente común.

e= vector de errores.

X = matriz de diseño que relaciona los registros con los efectos fijos.

Z= matriz de diseño que relaciona los registros con el efecto aleatorio del animal.

C= matriz de diseño que relaciona los registros con el efecto aleatorio ambiental común.

Bajo este modelo, se asume que los efectos ambientales comunes y los errores se distribuyen normal, idéntica e independientemente con media 0 y varianzas σ_c^2 y σ_e^2, respectivamente, por lo tanto, se tienen las siguientes matrices de covarianza:

$$Var(c) = \sigma_c^2 I.$$

$$Var(e) = \sigma_c^2 I.$$

$$Var(a) = \sigma_a^2 A$$

Donde A es la matriz de parentesco.

Para construir las ecuaciones de modelos mixtos, deben considerarse las siguientes razones entre componentes de varianza:

$$\alpha_1 = \frac{\sigma_e^2}{\sigma_a^2}$$

$$\alpha_2 = \frac{\sigma_e^2}{\sigma_c^2}$$

El sistema de ecuaciones del modelo mixto tiene la siguiente estructura:

$$\begin{bmatrix} X^TX & X^TZ & X^TC \\ Z^TX & Z^TZ+\alpha_1^{-1}A & Z^TC \\ C^TX & C^TZ & C^TC+\alpha_2 I \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X^TY \\ Z^TY \\ C^TY \end{bmatrix}$$

Ahora desarrollemos el EJEMPLO 5.3, con la base de datos indicada en la TABLA NRO. 5.4. Se considera como efecto fijo el sexo de las crías y como característica a mejorar el peso de los cuyes al destete, en un plantel productivo de Colombia, utilizando información reportada por Solarte et al [26].

Para calcular los valores de α requeridos se toman los siguientes datos, tomando como base el trabajo realizado por Solarte et al [27]:

$\sigma_a^2 = 245$

$\sigma_c^2 = 190$

$\sigma_e^2 = 680$

$h^2 = \frac{245}{1115} = 0.22$

TABLA 5.4: Información genelógica y peso al destete (g) de una población de cuyes

Animal	Padre	Madre	Sexo	Camada	Peso Destete
1			1		
2			2		
3			1		
4			2		
5	1	2	1	1	210
6	1	2	2	1	170
7	3	4	1	2	160
8	3	4	2	2	130
9	1	2	1	3	230
10	1	2	2	3	180
11	3	4	1	4	150
12	3	4	2	4	120

Fuente: elaboración propia (2024).

$$\alpha_1 = \frac{680}{245} = 2.78$$

$$\alpha_2 = \frac{680}{190} = 3.58$$

La matriz de parentesco sería:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 1 \end{bmatrix}$$

Para construir el sistema del ecuaciones de modelo mixto, será necesario definir las matrices que a continuación se presentan, y llevar a cabo las correspondientes operaciones de transposición, multiplicación, adición e inversión, según se requiera en cada caso. La matriz X que como ya se dijo relaciona los efectos fijos (el sexo en el EJEMPLO 5.3) con los registros de los animales, por lo tanto, en las columnas se presentan los niveles del sexo, en la primera los machos y en la segunda las hembras; mientras que en las filas se ubican los animales con registro. En consecuencia, esta matriz tiene la siguiente estructura:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La matriz Z que relaciona los registros con los efectos genéticos aditivos directos, se construye ubicando en las filas los animales con registro y en las columnas todos los animales, independientemente de si tienen o no registros.

$$Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La matriz C relaciona los registros con los efectos aleatorios del ambiente común de camada, tendrá en las filas la representación de los animales con registro y en las columnas las camadas, por lo que esta matriz adopta la siguiente estructura:

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

El vector y corresponde a las observaciones, es decir, los datos de los registros del peso vivo a las ocho semanas de edad de los animales.

$$y = \begin{bmatrix} 210 \\ 170 \\ 160 \\ 130 \\ 230 \\ 180 \\ 150 \\ 120 \end{bmatrix}$$

Las matrices que contiene el sistema de ecuaciones son:

$$X^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$X^T C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$C^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$C^T C = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 750 \\ 600 \end{bmatrix}$$

$$Z^T y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 210 \\ \vdots \\ 230 \\ 180 \\ 150 \\ 120 \end{bmatrix}$$

$$C^T y = \begin{bmatrix} 380 \\ 290 \\ 410 \\ 270 \end{bmatrix}$$

$$Z^T Z + \alpha_1 A^{-1} = \begin{bmatrix} 8.33 & 5.55 & \cdots & 0 & 0 & -2.78 & -2.78 & 0 & 0 \\ 5.55 & 8.33 & \cdots & 0 & 0 & -2.78 & -2.78 & 0 & 0 \\ 0 & 0 & \cdots & -2.78 & -2.78 & 0 & 0 & -2.78 & -2.78 \\ 0 & 0 & \cdots & -2.78 & -2.78 & 0 & 0 & -2.78 & -2.78 \\ -2.78 & -2.78 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -2.78 & -2.78 & \cdots & 0 & 0 & 6.55 & 0 & 0 & 0 \\ -2.78 & -2.78 & \cdots & 0 & 0 & 0 & 6.55 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 6.55 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 6.55 \end{bmatrix}$$

$$C^T C + \alpha_2 I = \begin{bmatrix} 5.58 & 0 & 0 & 0 \\ 0 & 5.58 & 0 & 0 \\ 0 & 0 & 5.58 & 0 \\ 0 & 0 & 0 & 5.58 \end{bmatrix}$$

Al realizar las operaciones de inversión de la matriz del lado izquierdo de las ecuaciones de modelo mixto, esa matriz inversa multiplicada por el vector del lado derecho produce las siguientes soluciones:

$$\begin{bmatrix} \widehat{x_1} \\ \widehat{x_2} \\ \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \vdots \\ \widehat{c_1} \\ \widehat{c_2} \\ \widehat{c_3} \\ \widehat{c_4} \end{bmatrix} = \begin{bmatrix} 187.5 \\ 150 \\ 8.4 \\ 8.4 \\ -8.4 \\ \vdots \\ 4.1 \\ -4.9 \\ 8.9 \\ -8.1 \end{bmatrix}$$

Los resultados indicados en la fila 1 y en la fila 2 corresponden a la solución para los efectos fijos, en este caso, la media de los machos y las hembras. De la tercera a la sexta fila aparecen los valores genéticos para los animales que constituyen la población base, es decir, los individuos que aparecen como progenitores pero no tienen registro ni padres identificados.

Finalmente, las últimas tres filas contienen las cifras correspondientes al efecto común ambiental de camada. No debe perderse de vista en esta interpretación que en este efecto se incluye tanto la influencia de las hembras, por su habilidad materna, como los factores ambientales, por el hecho de que los hermanos completos y medios hermanos paternos comparten el mismo ambiente, que puede resultar beneficioso para la expresión de la característica o por el contrario perjudicar la expresión.

Los valores genéticos serían:

$$\begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \vdots \\ \widehat{a_9} \\ \widehat{a_{10}} \\ \widehat{a_{11}} \\ \widehat{a_{12}} \end{bmatrix} = \begin{bmatrix} 8.4 \\ 8.4 \\ -8.4 \\ -8.4 \\ 9.9 \\ \vdots \\ 12.3 \\ 10.4 \\ -11.6 \\ -10.5 \end{bmatrix}$$

Las DEPs serían:

$$\frac{1}{2}\begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \\ \widehat{a_4} \\ \widehat{a_5} \\ \vdots \\ \widehat{a_9} \\ \widehat{a_{10}} \\ \widehat{a_{11}} \\ \widehat{a_{12}} \end{bmatrix} = \begin{bmatrix} 4.2 \\ 4.2 \\ -4.2 \\ -4.2 \\ 4.95 \\ \vdots \\ 6.15 \\ 5.20 \\ -5.80 \\ -5.25 \end{bmatrix}$$

Las soluciones del ambiente común serían:

$$\begin{bmatrix} \widehat{c_1} \\ \widehat{c_2} \\ \widehat{c_3} \\ \widehat{c_4} \end{bmatrix} = \begin{bmatrix} 4.1 \\ -4.9 \\ 8.9 \\ -8.1 \end{bmatrix}$$

En TABLA NRO. 5.5 se presentan los animales ordenados por valor genético.

TABLA 5.5: Catálogo de cuyes para peso al destete (g)

Animal	Valor Genético
9	12.30
10	10.40
5	9.90
6	9.60
1	8.40
2	8.40
3	-8.40
4	-8.40
8	-9.40
7	-10.50
12	-10.60
11	-11.60

Fuente: elaboración propia generada en R-project [25].

5.3.1. Ejercicios en R-project

```
Genealogia=data.frame(matrix(ncol=4,byrow=TRUE,c(
1,NA,NA,1,
2,NA,NA,2,
3,NA,NA,1,
4,NA,NA,2,
5,1,2,1,
6,1,2,2,
7,3,4,1,
8,3,4,2,
9,1,2,1,
10,1,2,2,
11,3,4,1,
12,3,4,2
)))
colnames(Genealogia)=c("id","sire","dam","sex")
```

```
Genealogia
```

```
##    id sire dam sex
## 1   1   NA  NA   1
## 2   2   NA  NA   2
## 3   3   NA  NA   1
## 4   4   NA  NA   2
## 5   5    1   2   1
```

```
## 6    6     1    2    2
## 7    7     3    4    1
## 8    8     3    4    2
## 9    9     1    2    1
## 10  10     1    2    2
## 11  11     3    4    1
## 12  12     3    4    2
```

Relaciones entre padres y madres:

```
table(Genealogia$sire,Genealogia$dam)


##
##      2 4
##    1 4 0
##    3 0 4
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id = Genealogia$id, dadid = Genealogia$sire,
                momid = Genealogia$dam,
                sex=as.numeric(Genealogia$sex))
```

Montaje de la matriz de parentesco:

```
A=2*kinship(Genealogia$id,Genealogia$sire,Genealogia$dam)
A[1:6,1:12]


##      1   2 3 4   5   6   7   8   9  10  11  12
## 1 1.0 0.0 0 0 0.5 0.5 0.0 0.0 0.5 0.5 0.0 0.0
## 2 0.0 1.0 0 0 0.5 0.5 0.0 0.0 0.5 0.5 0.0 0.0
## 3 0.0 0.0 1 0 0.0 0.0 0.5 0.5 0.0 0.0 0.5 0.5
## 4 0.0 0.0 0 1 0.0 0.0 0.5 0.5 0.0 0.0 0.5 0.5
## 5 0.5 0.5 0 0 1.0 0.5 0.0 0.0 0.5 0.5 0.0 0.0
## 6 0.5 0.5 0 0 0.5 1.0 0.0 0.0 0.5 0.5 0.0 0.0
```

```
bitSize(Genea1)


## $bitSize
## [1] 12
##
## $nFounder
## [1] 4
##
## $nNonFounder
## [1] 8
```

El árbol genealógico está representado en la FIGURA NRO. 5.4:

```
plot(Genea1,mar=c(bottom=0, left=1, top=1, right=1))
```

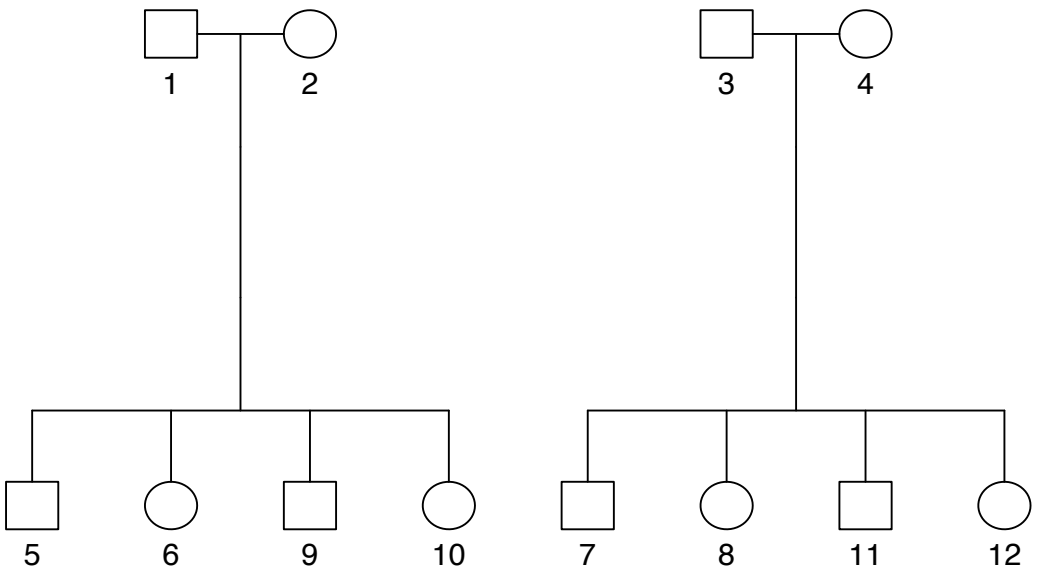

***Figura* 5.4:** Genealogía de los animales para el ejercio de modelo con ambiente común.
Fuente: elaboración propia generada en R-project [25].

Información de producciones:

```
Prod=data.frame(matrix(ncol=4,byrow=TRUE,c(
5,      1,      1,      210,
6,      1,      2,      170,
```

```
7,      2,      1,      160,
8,      2,      2,      130,
9,      3,      1,      230,
10,     3,      2,      180,
11,     4,      1,      150,
12,     4,      2,      120
  )))
colnames(Prod)=c("id","camada","GSx","PD")
Prod$PD=as.numeric(Prod$PD)
```

```
Prod
```

```
##   id camada GSx  PD
## 1  5      1   1 210
## 2  6      1   2 170
## 3  7      2   1 160
## 4  8      2   2 130
## 5  9      3   1 230
## 6 10      3   2 180
## 7 11      4   1 150
## 8 12      4   2 120
```

Número de animales:

```
n=nrow(Genealogia)
n
```

```
## [1] 12
```

Montaje de las matrices X, Z y C y el vector y:

```
library(MatrixModels)
X=as.matrix(model.Matrix(~ as.factor(Prod$GSx) -1))
rownames(X)=Prod$id
colnames(X)=c("GSx_1","GSx_2")
X
```

```
##   GSx_1 GSx_2
## 5     1     0
## 6     0     1
## 7     1     0
## 8     0     1
```

```
## 9       1      0
## 10      0      1
## 11      1      0
## 12      0      1
```

```
Pro=as.matrix(model.Matrix(~ as.factor(Prod$id) -1))
library(stringr)
colnames(Pro)=word(colnames(Pro), 2, sep = fixed(')'))
rownames(Pro)=Prod$id
Pro
```

```
##    5 6 7 8 9 10 11 12
## 5  1 0 0 0 0  0  0  0
## 6  0 1 0 0 0  0  0  0
## 7  0 0 1 0 0  0  0  0
## 8  0 0 0 1 0  0  0  0
## 9  0 0 0 0 1  0  0  0
## 10 0 0 0 0 0  1  0  0
## 11 0 0 0 0 0  0  1  0
## 12 0 0 0 0 0  0  0  1
```

```
Z=matrix(nrow=nrow(Pro),ncol=ncol(A),0)
colnames(Z)=colnames(A)
rownames(Z)=colnames(Pro)
Z[colnames(Pro),colnames(Pro)]=Pro
Z
```

```
##    1 2 3 4 5 6 7 8 9 10 11 12
## 5  0 0 0 0 1 0 0 0 0  0  0  0
## 6  0 0 0 0 0 1 0 0 0  0  0  0
## 7  0 0 0 0 0 0 1 0 0  0  0  0
## 8  0 0 0 0 0 0 0 1 0  0  0  0
## 9  0 0 0 0 0 0 0 0 1  0  0  0
## 10 0 0 0 0 0 0 0 0 0  1  0  0
## 11 0 0 0 0 0 0 0 0 0  0  1  0
## 12 0 0 0 0 0 0 0 0 0  0  0  1
```

```
Comun=as.matrix(model.Matrix(~ as.factor(Prod$camada) -1))
colnames(Comun)=word(colnames(Comun), 2, sep = fixed(')'))
rownames(Comun)=Prod$id
Comun
```

```
##    1 2 3 4
## 5  1 0 0 0
## 6  1 0 0 0
## 7  0 1 0 0
## 8  0 1 0 0
## 9  0 0 1 0
## 10 0 0 1 0
## 11 0 0 0 1
## 12 0 0 0 1
```

```
C=matrix(nrow=nrow(Comun),ncol=ncol(Comun),0)
colnames(C)=colnames(Comun)
rownames(C)=rownames(Comun)
C[rownames(Comun),colnames(Comun)]=Comun
C
```

```
##    1 2 3 4
## 5  1 0 0 0
## 6  1 0 0 0
## 7  0 1 0 0
## 8  0 1 0 0
## 9  0 0 1 0
## 10 0 0 1 0
## 11 0 0 0 1
## 12 0 0 0 1
```

```
y=as.matrix(Prod$PD)
colnames(y)=c("PD")
rownames(y)=Prod$VacaGSx
y
```

```
##       PD
## [1,] 210
## [2,] 170
## [3,] 160
## [4,] 130
## [5,] 230
## [6,] 180
## [7,] 150
## [8,] 120
```

Montaje del sistema de ecuaciones:

```
XpX=t(X)%*%X
XpX
```

```
##        GSx_1 GSx_2
## GSx_1      4     0
## GSx_2      0     4
```

```
XpZ=t(X)%*%Z
XpZ
```

```
##        1 2 3 4 5 6 7 8 9 10 11 12
## GSx_1  0 0 0 0 1 0 1 0 1  0  1  0
## GSx_2  0 0 0 0 0 1 0 1 0  1  0  1
```

```
ZpX=t(Z)%*%X
```

```
ZpZ=t(Z)%*%Z
ZpZ
```

```
##    1 2 3 4 5 6 7 8 9 10 11 12
## 1  0 0 0 0 0 0 0 0 0  0  0  0
## 2  0 0 0 0 0 0 0 0 0  0  0  0
## 3  0 0 0 0 0 0 0 0 0  0  0  0
## 4  0 0 0 0 0 0 0 0 0  0  0  0
## 5  0 0 0 0 1 0 0 0 0  0  0  0
## 6  0 0 0 0 0 1 0 0 0  0  0  0
## 7  0 0 0 0 0 0 1 0 0  0  0  0
## 8  0 0 0 0 0 0 0 1 0  0  0  0
## 9  0 0 0 0 0 0 0 0 1  0  0  0
## 10 0 0 0 0 0 0 0 0 0  1  0  0
## 11 0 0 0 0 0 0 0 0 0  0  1  0
## 12 0 0 0 0 0 0 0 0 0  0  0  1
```

```
XpC=t(X)%*%C
XpC
```

```
##        1 2 3 4
## GSx_1  1 1 1 1
## GSx_2  1 1 1 1
```

```
CpX=t(C)%*%X
```

```
ZpC=t(Z)%*%C
ZpC
```

```
##    1 2 3 4
## 1  0 0 0 0
## 2  0 0 0 0
## 3  0 0 0 0
## 4  0 0 0 0
## 5  1 0 0 0
## 6  1 0 0 0
## 7  0 1 0 0
## 8  0 1 0 0
## 9  0 0 1 0
## 10 0 0 1 0
## 11 0 0 0 1
## 12 0 0 0 1
```

```
CpZ=t(C)%*%Z
```

```
CpC=t(C)%*%C
CpC
```

```
##   1 2 3 4
## 1 2 0 0 0
## 2 0 2 0 0
## 3 0 0 2 0
## 4 0 0 0 2
```

```
Xpy=t(X)%*%y
Xpy
```

```
##         PD
## GSx_1 750
## GSx_2 600
```

```
Zpy=t(Z)%*%y
Zpy
```

```
##     PD
## 1    0
## 2    0
## 3    0
## 4    0
## 5  210
## 6  170
## 7  160
## 8  130
## 9  230
## 10 180
## 11 150
## 12 120
```

```
Cpy=t(C)%*%y
Cpy
```

```
##    PD
## 1 380
## 2 290
## 3 410
## 4 270
```

Inclusión de los alfas:

```
vara=245
varc=190
vare=680
h2=vara/(vara+varc+vare)
h2
```

```
## [1] 0.22
```

```
alpha1=vare/vara
alpha1
```

```
## [1] 2.8
```

```
alpha2=vare/varc
alpha2
```

```
## [1] 3.6
```

```
library(MASS)
Ainv=ginv(A)
round(Ainv[1:8,1:8],2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    3    2    0    0   -1   -1    0    0
## [2,]    2    3    0    0   -1   -1    0    0
## [3,]    0    0    3    2    0    0   -1   -1
## [4,]    0    0    2    3    0    0   -1   -1
## [5,]   -1   -1    0    0    2    0    0    0
## [6,]   -1   -1    0    0    0    2    0    0
## [7,]    0    0   -1   -1    0    0    2    0
## [8,]    0    0   -1   -1    0    0    0    2
```

```
alfa1InvA=alpha1%x%Ainv
round(alfa1InvA[1:8,1:8],2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]  8.3  5.5  0.0  0.0 -2.8 -2.8  0.0  0.0
## [2,]  5.5  8.3  0.0  0.0 -2.8 -2.8  0.0  0.0
## [3,]  0.0  0.0  8.3  5.5  0.0  0.0 -2.8 -2.8
## [4,]  0.0  0.0  5.5  8.3  0.0  0.0 -2.8 -2.8
## [5,] -2.8 -2.8  0.0  0.0  5.5  0.0  0.0  0.0
## [6,] -2.8 -2.8  0.0  0.0  0.0  5.5  0.0  0.0
## [7,]  0.0  0.0 -2.8 -2.8  0.0  0.0  5.5  0.0
## [8,]  0.0  0.0 -2.8 -2.8  0.0  0.0  0.0  5.5
```

```
ZpZalfa1InvA=ZpZ+alfa1InvA
round(ZpZalfa1InvA[1:8,1:8],1)
```

```
##      1    2    3    4    5    6    7    8
## 1  8.3  5.6  0.0  0.0 -2.8 -2.8  0.0  0.0
## 2  5.6  8.3  0.0  0.0 -2.8 -2.8  0.0  0.0
## 3  0.0  0.0  8.3  5.6  0.0  0.0 -2.8 -2.8
## 4  0.0  0.0  5.6  8.3  0.0  0.0 -2.8 -2.8
## 5 -2.8 -2.8  0.0  0.0  6.6  0.0  0.0  0.0
## 6 -2.8 -2.8  0.0  0.0  0.0  6.6  0.0  0.0
## 7  0.0  0.0 -2.8 -2.8  0.0  0.0  6.6  0.0
## 8  0.0  0.0 -2.8 -2.8  0.0  0.0  0.0  6.6
```

```
I=matrix(nrow=nrow(t(C)),ncol=ncol(C),0);I
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    0    0    0
## [2,]    0    0    0    0
## [3,]    0    0    0    0
## [4,]    0    0    0    0
```

```
diag(I)=1;colnames(I)=rownames(I)=colnames(C)
alpha2I=alpha2%x%I
round(alpha2I,2)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  3.6  0.0  0.0  0.0
## [2,]  0.0  3.6  0.0  0.0
## [3,]  0.0  0.0  3.6  0.0
## [4,]  0.0  0.0  0.0  3.6
```

```
CpCalfa2I=CpC+alpha2I
round(CpCalfa2I,2)
```

```
##     1   2   3   4
## 1 5.6 0.0 0.0 0.0
## 2 0.0 5.6 0.0 0.0
## 3 0.0 0.0 5.6 0.0
## 4 0.0 0.0 0.0 5.6
```

Construcción del lado izquierdo del sistema:

```
Izq=rbind(
  cbind(XpX,XpZ,XpC),
  cbind(ZpX,ZpZalfa1InvA,ZpC),
  cbind(CpX,CpZ,CpCalfa2I))
Izqinv=solve(Izq)
```

Lado derecho del sistema:

```
Der=rbind(Xpy,Zpy,Cpy)
Der
```

```
##        PD
```

```
## GSx_1 750
## GSx_2 600
## 1       0
## 2       0
## 3       0
## 4       0
## 5     210
## 6     170
## 7     160
## 8     130
## 9     230
## 10    180
## 11    150
## 12    120
## 1     380
## 2     290
## 3     410
## 4     270
```

Solución al sistema:

```
Sol=round(Izqinv%*%Der,1)
rownames(Sol)=c(rownames(XpX),paste0("a_",rownames(ZpZ)),
paste0("c_",rownames(CpC)))
Sol
```

```
##          PD
## GSx_1 187.5
## GSx_2 150.0
## a_1     8.4
## a_2     8.4
## a_3    -8.4
## a_4    -8.4
## a_5     9.9
## a_6     9.6
## a_7   -10.6
## a_8    -9.4
## a_9    12.3
## a_10   10.4
## a_11  -11.6
## a_12  -10.5
## c_1     4.1
## c_2    -4.9
## c_3     8.9
## c_4    -8.1
```

Valores genéticos:

```
VG=as.matrix(Sol[paste0("a_",rownames(ZpZ)),])
VG
```

```
##        [,1]
## a_1     8.4
## a_2     8.4
## a_3    -8.4
## a_4    -8.4
## a_5     9.9
## a_6     9.6
## a_7   -10.6
## a_8    -9.4
## a_9    12.3
## a_10   10.4
## a_11  -11.6
## a_12  -10.5
```

Diferencia esperada de progenie:

```
Deps=VG/2
```

Ambiente común:

```
AC=as.matrix(Sol[paste0("c_",rownames(CpC)),])
AC
```

```
##      [,1]
## c_1   4.1
## c_2  -4.9
## c_3   8.9
## c_4  -8.1
```

5.4. Modelo animal para varias características (multicaracter)

En este caso el interés se centra en el análisis conjunto de dos o más características. Bajo esta condición se determina la influencia de la acción genética aditiva para cada rasgo y las correlaciones entre características, tanto de tipo genético aditivo, como de orden ambiental; aunque es preciso indicar que, en algunos casos, se asume que no existen covarianzas de tipo ambiental entre los rasgos considerados y en dicha situación

se reducen notablemente los cálculos necesarios para estimar el mérito genético de los animales incluidos en las evaluaciones.

Para simplificar la presentación del modelo, nos enfocamos en el caso de dos características. El modelo animal correspondiente se expresa de la siguiente manera:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

Donde:

y_1 y y_2 representan los vectores que contienen las observaciones de las dos características;

β_1 es el vector de efectos fijos de la característica 1;

β_2 es el vector de efectos fijos de la característica 2.

Los efectos pueden ser los mismos o diferentes para las dos características;

a_1 es el vector de valores genéticos aditivos directos para la característica 1;

a_2 es vector de valores genéticos aditivos directos l para la característica 2;

e_1 el vector de errores para la característica 1;

e_2 el vector de errores para la característica 2.

$$Var \begin{bmatrix} a_1 \\ a_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A & 0 & 0 \\ g_{21}A & g_{22}A & 0 & 0 \\ 0 & 0 & e_{11}I & e_{12}I \\ 0 & 0 & e_{21}I & e_{22}I \end{bmatrix}$$

Donde: A es la matriz de parentesco; g_{11} es la varianza genética aditiva para la característica 1; g_{22} es la varianza genética aditiva para la característica 2 y $g_{12} = g_{21}$, es la covarianza genética aditiva entre las características 1 y 2, I es la matriz identidad; r_{11} es la varianza del error para la característica 1; r_{22} es la varianza del error para la característica 2 y $r_{12} = r_{21}$, es la covarianza del error entre las características 1 y 2.

Sea:

$$G_0 = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

Entonces:

$$G^{-1} = G_0^{-1} \otimes A^{-1}$$

Sea:

$$R_0 = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

Entonces:

$$R^{-1} = R_0^{-1} \otimes I$$

Donde $\otimes$ es el producto, conocido como Kronecker o multiplicación directa de dos matrices. Las ecuaciones del modelo mixto se expresan de la siguiente manera, según Henderson [36]

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{a} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

Donde:

$$X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}$$

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta_1} \\ \widehat{\beta_1} \end{bmatrix}$$

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}$$

$$\widehat{a} = \begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Al escribir las ecuaciones para cada característica, el sistema de ecuaciones del modelo mixto es el siguiente, según lo describe Mrode [21]:

$$\begin{bmatrix} X_1^T r^{11} X_1 & X_1^T r^{12} X_2 & X_1^T r^{11} Z_1 & X_1^T r^{12} Z_2 \\ X_2^T r^{21} X_1 & X_2^T r^{22} X_2 & X_2^T r^{21} Z_1 & X_2^T r^{22} Z_2 \\ Z_1^T r^{11} X_1 & Z_1^T r^{12} X_2 & Z_1^T r^{11} Z_1 + A^{-1} g^{11} & Z_1^T r^{12} Z_2 + A^{-1} g^{12} \\ Z_2^T r^{21} X_1 & Z_2^T r^{22} X_2 & Z_2^T r^{21} Z_1 + A^{-1} g^{21} & Z_2^T r^{22} Z_2 + A^{-1} g^{22} \end{bmatrix} \begin{bmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \\ \widehat{a_1} \\ \widehat{a_2} \end{bmatrix}$$

$$= \begin{bmatrix} X_1^T r^{11} y_1 + X_1^T r^{12} y_2 \\ X_2^T r^{21} y_1 + X_2^T r^{22} y_2 \\ Z_1^T r^{11} y_1 + Z_1^T r^{12} y_2 \\ Z_2^T r^{21} y_1 + Z_2^T r^{22} y_2 \end{bmatrix}$$

Donde g^{ij} y r^{ij} son elementos de la matriz inversa de covarianzas genéticas aditivas y residuales, respectivamente.

Como puede apreciarse, bajo este modelo se generará un sistema de ecuaciones lineales cuyo tamaño aumenta con el número de fenotipos evaluados. En el EJEMPLO 5.4 se explica paso a paso el procedimiento para construir y resolver el sistema de ecuaciones utilizando los datos de pedigrí y pesos a la octava y decimosegunda semana en cuyes machos (*Cavia porcellus Rodentia: caviidae*).

La estructura de la base de datos contiene la información genealógica (animal, padre y madre), el sexo , el grupo contemporáneo y los pesos a las 8 y 12 semanas de edad (TABLA NRO. 5.6). El desarrollo del sistema de ecuaciones tiene en cuenta que las dos características tienen diferente número de registros.

TABLA 5.6: Información de pedigrí y peso a las ocho y doce semanas (*g*) de cuyes (*Cavia porcellus Rodentia: caviidae*)

Animal	Padre	Madre	Sexo	Grupo 8	Grupo 12	Peso 8	Peso 12
1			1				
2			2				
3			1				
4			2		1		930
5	1	2	1	1	2	850	900
6	1	2	1	2	1	800	980
7	3	4	1	1	1	650	850
8	3	4	1	2	2	700	880

Fuente: elaboración propia (2024).

La información de varianzas y covarianzas corresponde a valores de referencia obtenidos por Solarte et al [27] que encontraron correlación positiva tanto en la parte genética aditiva como en la del error , entre el peso a la octava y decimosegunda semana de edad:

$$G_0 = \begin{bmatrix} 352 & 300 \\ 300 & 533 \end{bmatrix}$$

$$R_0 = \begin{bmatrix} 660 & 610 \\ 610 & 1170 \end{bmatrix}$$

La matriz de parentesco es:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.5 & 1 \end{bmatrix}$$

La parte de la matriz X asociada al grupo contemporáneo se compone de dos submatrices, una para cada característica:

$$X_{P8} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$X_{P12} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

En la matriz Z, las filas representan los animales con registros en una o en las dos características y en las columnas todos los animales. Se tienen dos submatrices Z:

$$Z_{P8} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$Z_{P12} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El vector y expresado en dos subvectores es:

$$y_{P8} = \begin{bmatrix} 850 \\ 800 \\ 650 \\ 700 \end{bmatrix}$$

$$y_{P12} = \begin{bmatrix} 930 \\ 900 \\ 980 \\ 850 \\ 880 \end{bmatrix}$$

Para construir el resto de los elementos matriciales que constituyen el sistema de ecuaciones del modelo mixto, se procede disponer de los elementos de G^1 y R^{-1} así:

Las matrices del sistema serían:

$$X^T R^{-1} X = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$Z^T R^{-1} X = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.0015 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0.0009 \\ 0 & 0 & 0.0009 & 0 \\ 0 & 0 & 0.0009 & 0 \\ 0 & 0 & 0 & 0.0009 \end{bmatrix}$$

$$Z^T R^{-1} Z + G_0^{-1} \otimes A^{-1} = \begin{bmatrix} 0.01 & 0.01 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0.01 & 0.01 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & \cdots & 0 & 0 & 0 & 0 \\ -0.01 & -0.01 & 0 & \cdots & -0.01 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0.01 \end{bmatrix}$$

$$X_{P8}^T R^{-1} y_{P8} = \begin{bmatrix} 1.72 \\ 1.55 \end{bmatrix}$$

$$X_{P12}^{T}R^{-1}y_{P12} = \begin{bmatrix} 2.34 \\ 0.57 \end{bmatrix}$$

$$Z_{P8}^{T}R^{-1}y_{P8} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1.11 \\ 0.85 \\ 0.6 \\ 0.71 \end{bmatrix}$$

$$Z_{P12}^{T}R^{-1}y_{P12} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1.53 \\ 0.19 \\ 0.4 \\ 0.41 \\ 0.38 \end{bmatrix}$$

Las soluciones del sistema de ecuaciones son:

$$\begin{bmatrix} P8 : \widehat{x_1} \\ P8 : \widehat{x_2} \\ P12 : \widehat{x_1} \\ P12 : \widehat{x_2} \\ P8 : \widehat{a_1} \\ P8 : \widehat{a_2} \\ P8 : \widehat{a_3} \\ P8 : \widehat{a_4} \\ P8 : \widehat{a_5} \\ P8 : \widehat{a_6} \\ P8 : \widehat{a_7} \\ P8 : \widehat{a_8} \\ P12 : \widehat{a_1} \\ P12 : \widehat{a_2} \\ P12 : \widehat{a_3} \\ P12 : \widehat{a_4} \\ P12 : \widehat{a_5} \\ P12 : \widehat{a_6} \\ P12 : \widehat{a_7} \\ P12 : \widehat{a_8} \end{bmatrix} = \begin{bmatrix} 765 \\ 736 \\ 939 \\ 868 \\ 22 \\ 22 \\ -23 \\ -22 \\ 36 \\ 31 \\ -42 \\ -26 \\ 11 \\ 11 \\ -12 \\ -10 \\ 16 \\ 17 \\ -27 \\ -7.3 \end{bmatrix}$$

Las soluciones para las medias del grupo contemporáneo son:

$$\begin{bmatrix} P8 : \widehat{x_1} \\ P8 : \widehat{x_2} \\ P12 : \widehat{x_1} \\ P12 : \widehat{x_2} \end{bmatrix} = \begin{bmatrix} 765 \\ 736 \\ 939 \\ 868 \end{bmatrix}$$

Ahora calculamos la confiabilidad, la exactitud y la raiz cuadrada de la varianza del error de predicción (a la que denominaremos *SEP*), empleando el procedimiento que se usó en el capítulo anterior. El vector d para el EJEMPLO 5.4, se tendría:

$$d_i = \begin{bmatrix} P8 : a_1 \\ P8 : a_2 \\ P8 : a_3 \\ P8 : a_4 \\ P8 : a_5 \\ P8 : a_6 \\ P8 : a_7 \\ P8 : a_8 \\ P12 : a_1 \\ P12 : a_2 \\ P12 : a_3 \\ P12 : a_4 \\ P12 : a_5 \\ P12 : a_6 \\ P12 : a_7 \\ P12 : a_8 \end{bmatrix} = \begin{bmatrix} 324 \\ 324 \\ 320 \\ 298 \\ 275 \\ 276 \\ 280 \\ 280 \\ 491 \\ 491 \\ 478 \\ 410 \\ 424 \\ 412 \\ 432 \\ 431 \end{bmatrix}$$

Los resultados de la evaluación genética se presenta en la TABLA NRO. 5.7 para peso a las 8 semanas y en la TABLA NRO. 5.8 para peso a las doce semanas.

Tabla 5.7: Valoración genética de cuyes (*Cavia porcellus Rodentia: caviidae*) para peso a las 8 semanas de edad

Valor Genético	animal	DEP	Confiabilidad	Exactitud	SEP
36.14	5	18.07	0.22	0.47	16.60
31.36	6	15.68	0.22	0.47	16.61
22.50	1	11.25	0.08	0.28	18.00
22.50	2	11.25	0.08	0.28	18.00
-22.14	4	-11.07	0.15	0.39	17.28
-22.85	3	-11.43	0.09	0.30	17.88
-25.80	8	-12.90	0.20	0.45	16.73
-42.06	7	-21.03	0.20	0.45	16.73

Nota: DEP=diferencia esperada de progenie y SEP=raiz cuadrada de la varianza del error de predicción.
Fuente: elaboración propia generada en R-project [25].

Tabla 5.8: Valoración genética de cuyes (*Cavia porcellus Rodentia: caviidae*) para el peso a las 12 semanas de edad

Valor genético	animal	DEP	Confiabilidad	Exactitud	SEP
17.17	6	8.59	0.23	0.48	20.29
16.09	5	8.04	0.20	0.45	20.60
11.08	1	5.54	0.08	0.28	22.15
11.08	2	5.54	0.08	0.28	22.15
-7.28	8	-3.64	0.19	0.44	20.77
-10.45	4	-5.22	0.23	0.48	20.24
-11.72	3	-5.86	0.10	0.32	21.86
-26.61	7	-13.30	0.19	0.44	20.79

Nota: DEP=diferencia esperada de progenie y SEP=raiz cuadrada de la varianza del error de predicción.
Fuente: elaboración propia generada en R-project [25].

5.4.1. Ejercicios en R-project

```
Genealogia=data.frame(matrix(ncol=4,byrow=TRUE,c(
  1,NA,NA,1,
  2,NA,NA,2,
  3,NA,NA,1,
  4,NA,NA,2,
  5,1,2,1,
  6,1,2,1,
  7,3,4,1,
  8,3,4,1
  )))
  colnames(Genealogia)=c("id","sire","dam","sex")
```

```
Genealogia
```

```
##   id sire dam sex
## 1  1   NA  NA   1
## 2  2   NA  NA   2
## 3  3   NA  NA   1
## 4  4   NA  NA   2
## 5  5    1   2   1
## 6  6    1   2   1
## 7  7    3   4   1
## 8  8    3   4   1
```

Tabla de relaciones entre padres y madres:

```
table(Genealogia$sire,Genealogia$dam)
```

```
##
##     2 4
##   1 2 0
##   3 0 2
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id = Genealogia$id, dadid = Genealogia$sire,
```

```
                momid = Genealogia$dam,
                sex=as.numeric(Genealogia$sex))
```

Montaje de la matriz de parentesco:

```
A=2*kinship(Genealogia$id,Genealogia$sire,Genealogia$dam)
A[1:4,]
```

```
##   1 2 3 4   5   6   7   8
## 1 1 0 0 0 0.5 0.5 0.0 0.0
## 2 0 1 0 0 0.5 0.5 0.0 0.0
## 3 0 0 1 0 0.0 0.0 0.5 0.5
## 4 0 0 0 1 0.0 0.0 0.5 0.5
```

```
bitSize(Genea1)
```

```
## $bitSize
## [1] 4
##
## $nFounder
## [1] 4
##
## $nNonFounder
## [1] 4
```

El árbol genealógico está representado en la FIGURA NRO. 5.5:

```
plot(Genea1,mar=c(bottom=0, left=1, top=1, right=1))
```

Información de producciones:

```
Prod=data.frame(matrix(ncol=5,byrow=TRUE,c(
4,NA,1,0,930,
5,1,2,850,900,
6,2,1,800,980,
7,1,1,650,850,
8,2,2,700,880
)))
colnames(Prod)=c("id","GC_p8","GC_p12","P8","P12")
```

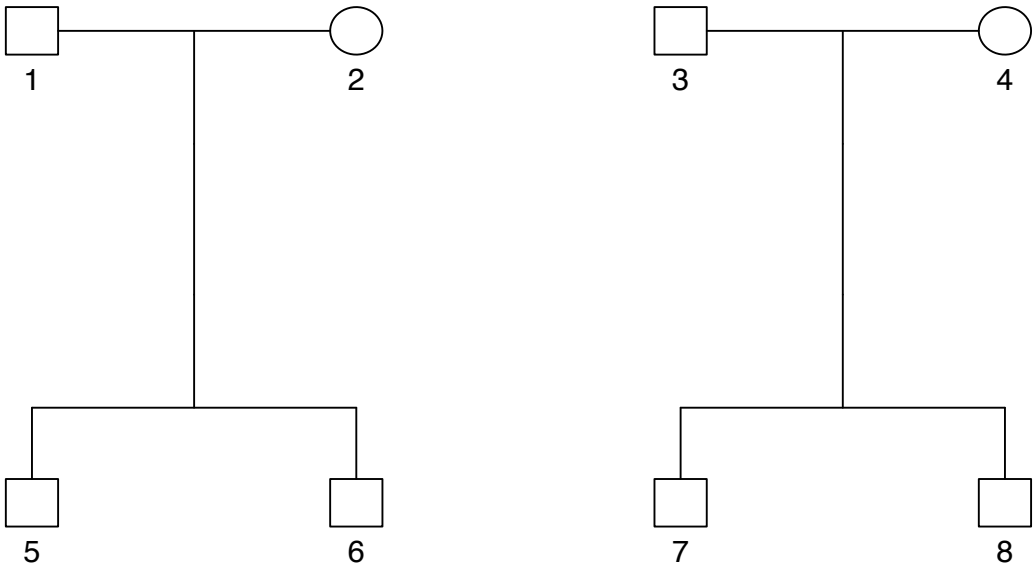

***Figura* 5.5:** Genealogía de los animales para el ejercicio de modelo animal multicaracterístico.
Fuente: elaboración propia generada en R-project [25].

```
Prod
```

```
##   id GC_p8 GC_p12  P8 P12
## 1  4    NA      1   0 930
## 2  5     1      2 850 900
## 3  6     2      1 800 980
## 4  7     1      1 650 850
## 5  8     2      2 700 880
```

```
mean(Prod$P8, na.rm=T)
```

```
## [1] 600
```

```
sd(Prod$P8, na.rm=T)^2
```

```
## [1] 118750
```

```
mean(Prod$P12, na.rm=T)
```

```
## [1] 908
```

```
sd(Prod$P12, na.rm=T)^2
```

```
## [1] 2470
```

Número de animales y datos:

```
n=nrow(Genealogia)
n
```

```
## [1] 8
```

```
nd=nrow(Prod)
nd
```

```
## [1] 5
```

Bases de datos para las dos características:

```
P8=subset(Prod,Prod$P8>=0,select=c("id","GC_p8","P8"))
P8
```

```
##   id GC_p8  P8
## 1  4    NA   0
## 2  5     1 850
## 3  6     2 800
## 4  7     1 650
## 5  8     2 700
```

```
P12=subset(Prod,Prod$P12>=0,select=c("id","GC_p12","P12"))
P12
```

```
##   id GC_p12 P12
## 1  4      1 930
## 2  5      2 900
## 3  6      1 980
## 4  7      1 850
## 5  8      2 880
```

Montaje de las matrices *G* y *R*:

```
var_a11=352
cov_a12=cov_a21=300
var_a22=533

var_e11=660
cov_e12=cov_e21=610
var_e22=1170

h2_1=var_a11/(var_a11+var_e11);round(h2_1,2)


## [1] 0.35


h2_2=var_a22/(var_a22+var_e22);round(h2_2,2)


## [1] 0.31


cor_G=cov_a12/sqrt(var_a11*var_a22);round(cor_G,2)


## [1] 0.69


cor_R=cov_e12/sqrt(var_e11*var_e22);round(cor_R,2)


## [1] 0.69
```

```
G=matrix(nrow=2,ncol=2,byrow=TRUE,c(
  var_a11,cov_a12,
  cov_a21,var_a22
))
round(G,3)


##      [,1] [,2]
## [1,]  352  300
## [2,]  300  533
```

```
Ginv=solve(G)
round(Ginv,3)


##        [,1]   [,2]
```

```
## [1,]  0.005 -0.003
## [2,] -0.003  0.004
```

```
R=matrix(nrow=2,ncol=2,byrow=TRUE,c(
  var_e11,cov_e12,
  cov_e21,var_e22
))
round(R,3)
```

```
##      [,1] [,2]
## [1,]  660  610
## [2,]  610 1170
```

```
Rinv=solve(R)
round(Rinv,3)
```

```
##        [,1]   [,2]
## [1,]  0.003 -0.002
## [2,] -0.002  0.002
```

Montaje de las matrices X y Z y el vector y:

```
X_p8=t(table(P8$GC_p8, P8$id))
colnames(X_p8)=paste0("G_p8",colnames(X_p8))
X_p8
```

```
##
##     G_p81 G_p82
##   4     0     0
##   5     1     0
##   6     0     1
##   7     1     0
##   8     0     1
```

```
X_p12=t(table(P12$GC_p12, P12$id))
colnames(X_p12)=paste0("G_p12",colnames(X_p12))
X_p12
```

```
##
##     G_p121 G_p122
```

```
##    4       1       0
##    5       0       1
##    6       1       0
##    7       1       0
##    8       0       1
```

```
Peso_p8=t(table(P8$id, P8$id))
sinreg=subset(Prod,Prod$P8==0,select=c("id"))
Peso_p8[as.factor(sinreg$id),as.factor(sinreg$id)]=0
Peso_p8
```

```
##
##      4 5 6 7 8
##    4 0 0 0 0 0
##    5 0 1 0 0 0
##    6 0 0 1 0 0
##    7 0 0 0 1 0
##    8 0 0 0 0 1
```

```
Z_p8=matrix(nrow=nrow(Peso_p8),ncol=ncol(A),0)
colnames(Z_p8)=colnames(A)
rownames(Z_p8)=colnames(Peso_p8)
Z_p8[rownames(Peso_p8),colnames(Peso_p8)]=Peso_p8
round(Z_p8,2)
```

```
##    1 2 3 4 5 6 7 8
## 4  0 0 0 0 0 0 0 0
## 5  0 0 0 0 1 0 0 0
## 6  0 0 0 0 0 1 0 0
## 7  0 0 0 0 0 0 1 0
## 8  0 0 0 0 0 0 0 1
```

```
Peso_p12=t(table(P12$id, P12$id))
sinreg=subset(Prod,Prod$P12==0,select=c("id"))
Peso_p12[as.factor(sinreg$id),as.factor(sinreg$id)]=0
Peso_p12
```

```
##
##      4 5 6 7 8
##    4 1 0 0 0 0
##    5 0 1 0 0 0
```

```
##   6 0 0 1 0 0
##   7 0 0 0 1 0
##   8 0 0 0 0 1
```

```
Z_p12=matrix(nrow=nrow(Peso_p12),ncol=ncol(A),0)
colnames(Z_p12)=colnames(A)
rownames(Z_p12)=colnames(Peso_p12)
Z_p12[rownames(Peso_p12),colnames(Peso_p12)]=Peso_p12
round(Z_p12,2)
```

```
##   1 2 3 4 5 6 7 8
## 4 0 0 0 1 0 0 0 0
## 5 0 0 0 0 1 0 0 0
## 6 0 0 0 0 0 1 0 0
## 7 0 0 0 0 0 0 1 0
## 8 0 0 0 0 0 0 0 1
```

```
y_p8=as.matrix(P8$P8)
colnames(y_p8)=c("P8")
rownames(y_p8)=P8$id
y_p8
```

```
##    P8
## 4   0
## 5 850
## 6 800
## 7 650
## 8 700
```

```
y_p12=as.matrix(P12$P12)
colnames(y_p12)=c("P12")
rownames(y_p12)=P12$id
y_p12
```

```
##   P12
## 4 930
## 5 900
## 6 980
## 7 850
## 8 880
```

Montaje del sistema de ecuaciones:

```
I=matrix(nrow=nd,ncol=nd,0);diag(I)=1
I
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
X_p8RinvX_p8=t(X_p8)%*%(I%x%Rinv[1,1])%*%X_p8
round(X_p8RinvX_p8,3)
```

```
##
##         G_p81 G_p82
##   G_p81 0.006 0.000
##   G_p82 0.000 0.006
```

```
X_p12RinvX_p12=t(X_p12)%*%(I%x%Rinv[2,2])%*%X_p12
round(X_p12RinvX_p12,3)
```

```
##
##          G_p121 G_p122
##   G_p121  0.005  0.000
##   G_p122  0.000  0.003
```

```
X_p8RinvX_p12=t(X_p8)%*%(I%x%Rinv[1,2])%*%X_p12
round(X_p8RinvX_p12,3)
```

```
##
##         G_p121 G_p122
##   G_p81 -0.002 -0.002
##   G_p82 -0.002 -0.002
```

```
X_p12RinvX_p8=t(X_p12)%*%(I%x%Rinv[2,1])%*%X_p8
round(X_p12RinvX_p8,3)
```

```
##
##           G_p81  G_p82
##   G_p121 -0.002 -0.002
##   G_p122 -0.002 -0.002
```

```
XpRinvX=rbind(
  cbind(X_p8RinvX_p8,X_p8RinvX_p12),
  cbind(X_p12RinvX_p8,X_p12RinvX_p12))
round(XpRinvX,3)
```

```
##         G_p81  G_p82 G_p121 G_p122
## G_p81   0.006  0.000 -0.002 -0.002
## G_p82   0.000  0.006 -0.002 -0.002
## G_p121 -0.002 -0.002  0.005  0.000
## G_p122 -0.002 -0.002  0.000  0.003
```

```
X_p8RinvZ_p8=t(X_p8)%*%(I%x%Rinv[1,1])%*%Z_p8
round(X_p8RinvZ_p8,2)
```

```
##
##         1 2 3 4 5 6 7 8
##   G_p81 0 0 0 0 0 0 0 0
##   G_p82 0 0 0 0 0 0 0 0
```

```
X_p8RinvZ_p12=t(X_p8)%*%(I%x%Rinv[1,2])%*%Z_p12
round(X_p8RinvZ_p12,2)
```

```
##
##         1 2 3 4 5 6 7 8
##   G_p81 0 0 0 0 0 0 0 0
##   G_p82 0 0 0 0 0 0 0 0
```

```
X_p12RinvZ_p8=t(X_p12)%*%(I%x%Rinv[2,1])%*%Z_p8
round(X_p12RinvZ_p8,2)
```

```
##
##          1 2 3 4 5 6 7 8
##   G_p121 0 0 0 0 0 0 0 0
##   G_p122 0 0 0 0 0 0 0 0
```

```
X_p12RinvZ_p12=t(X_p12)%*%(I%x%Rinv[2,2])%*%Z_p12
round(X_p12RinvZ_p12,2)
```

```
##
```

```
##           1 2 3 4 5 6 7 8
##   G_p121  0 0 0 0 0 0 0 0
##   G_p122  0 0 0 0 0 0 0 0
```

```
XpRinvZ=rbind(
  cbind(X_p8RinvZ_p8,X_p8RinvZ_p12),
  cbind(X_p12RinvZ_p8,X_p12RinvZ_p12))
round(XpRinvZ[,1:10],3)
```

```
##         1 2 3 4      5      6      7      8 1 2
## G_p81   0 0 0 0  0.003  0.000  0.003  0.000 0 0
## G_p82   0 0 0 0  0.000  0.003  0.000  0.003 0 0
## G_p121  0 0 0 0  0.000 -0.002 -0.002  0.000 0 0
## G_p122  0 0 0 0 -0.002  0.000  0.000 -0.002 0 0
```

```
ZpRinvX=t(XpRinvZ)
```

```
Ainv=solve(A)
round(Ainv,2)
```

```
##    1  2  3  4  5  6  7  8
## 1  2  1  0  0 -1 -1  0  0
## 2  1  2  0  0 -1 -1  0  0
## 3  0  0  2  1  0  0 -1 -1
## 4  0  0  1  2  0  0 -1 -1
## 5 -1 -1  0  0  2  0  0  0
## 6 -1 -1  0  0  0  2  0  0
## 7  0  0 -1 -1  0  0  2  0
## 8  0  0 -1 -1  0  0  0  2
```

```
Z_p8RinvZ_p8AinvGinv=(t(Z_p8)%*%(I%x%Rinv[1,1])%*%Z_p8)+
  Ainv%x%Ginv[1,1]
round(Z_p8RinvZ_p8AinvGinv[1:8,1:4],4)
```

```
##          1       2       3       4
## 1   0.0109  0.0055  0.0000  0.0000
## 2   0.0055  0.0109  0.0000  0.0000
## 3   0.0000  0.0000  0.0109  0.0055
## 4   0.0000  0.0000  0.0055  0.0109
## 5 -0.0055 -0.0055  0.0000  0.0000
```

```
## 6 -0.0055 -0.0055  0.0000  0.0000
## 7  0.0000  0.0000 -0.0055 -0.0055
## 8  0.0000  0.0000 -0.0055 -0.0055
```

```
Z_p8RinvZ_p12AinvGinv=(t(Z_p8)%*%(I%x%Rinv[1,2])%*%Z_p12)+
  Ainv%x%Ginv[1,2]
round(Z_p8RinvZ_p12AinvGinv[1:8,1:4],4)
```

```
##         1       2       3       4
## 1 -0.0061 -0.0031  0.0000  0.0000
## 2 -0.0031 -0.0061  0.0000  0.0000
## 3  0.0000  0.0000 -0.0061 -0.0031
## 4  0.0000  0.0000 -0.0031 -0.0061
## 5  0.0031  0.0031  0.0000  0.0000
## 6  0.0031  0.0031  0.0000  0.0000
## 7  0.0000  0.0000  0.0031  0.0031
## 8  0.0000  0.0000  0.0031  0.0031
```

```
Z_p12RinvZ_p8AinvGinv=(t(Z_p12)%*%(I%x%Rinv[2,1])%*%Z_p8)+
  Ainv%x%Ginv[2,1]
round(Z_p12RinvZ_p8AinvGinv[1:8,1:4],4)
```

```
##         1       2       3       4
## 1 -0.0061 -0.0031  0.0000  0.0000
## 2 -0.0031 -0.0061  0.0000  0.0000
## 3  0.0000  0.0000 -0.0061 -0.0031
## 4  0.0000  0.0000 -0.0031 -0.0061
## 5  0.0031  0.0031  0.0000  0.0000
## 6  0.0031  0.0031  0.0000  0.0000
## 7  0.0000  0.0000  0.0031  0.0031
## 8  0.0000  0.0000  0.0031  0.0031
```

```
Z_p12RinvZ_p12AinvGinv=(t(Z_p12)%*%(I%x%Rinv[2,2])%*%Z_p12)+
  Ainv%x%Ginv[2,2]
round(Z_p12RinvZ_p12AinvGinv[1:8,1:4],4)
```

```
##         1       2       3       4
## 1  0.0072  0.0036  0.0000  0.0000
## 2  0.0036  0.0072  0.0000  0.0000
## 3  0.0000  0.0000  0.0072  0.0036
## 4  0.0000  0.0000  0.0036  0.0089
```

```
## 5 -0.0036 -0.0036  0.0000  0.0000
## 6 -0.0036 -0.0036  0.0000  0.0000
## 7  0.0000  0.0000 -0.0036 -0.0036
## 8  0.0000  0.0000 -0.0036 -0.0036
```

```
ZpRinvZAinvGinv=rbind(
  cbind(Z_p8RinvZ_p8AinvGinv,Z_p8RinvZ_p12AinvGinv),
  cbind(Z_p12RinvZ_p8AinvGinv,Z_p12RinvZ_p12AinvGinv))
round(ZpRinvZAinvGinv[c(1:2,9:10),c(1:2,9:10)],4)
```

```
##         1       2       1       2
## 1  0.0109  0.0055 -0.0061 -0.0031
## 2  0.0055  0.0109 -0.0031 -0.0061
## 1 -0.0061 -0.0031  0.0072  0.0036
## 2 -0.0031 -0.0061  0.0036  0.0072
```

```
X_p8pRinvy_p8=(t(X_p8)%*%(I%x%Rinv[1,1])%*%y_p8)+
  (t(X_p8)%*%(I%x%Rinv[1,2])%*%y_p12)
round(X_p8pRinvy_p8,3)
```

```
##
##          P8
##   G_p81 1.7
##   G_p82 1.6
```

```
X_p12pRinvy_p12=(t(X_p12)%*%(I%x%Rinv[2,1])%*%y_p8)+
  (t(X_p12)%*%(I%x%Rinv[2,2])%*%y_p12)
round(X_p12pRinvy_p12,3)
```

```
##
##            P8
##   G_p121 2.34
##   G_p122 0.57
```

```
Z_p8pRinvy_p8=(t(Z_p8)%*%(I%x%Rinv[1,1])%*%y_p8)+
  (t(Z_p8)%*%(I%x%Rinv[1,2])%*%y_p12)
round(Z_p8pRinvy_p8,2)
```

```
##     P8
## 1 0.00
```

```
## 2 0.00
## 3 0.00
## 4 0.00
## 5 1.11
## 6 0.85
## 7 0.60
## 8 0.71
```

```
Z_p12pRinvy_p12=(t(Z_p12)%*%(I%x%Rinv[2,1])%*%y_p8)+
  (t(Z_p12)%*%(I%x%Rinv[2,2])%*%y_p12)
round(Z_p12pRinvy_p12,2)


##     P8
## 1 0.00
## 2 0.00
## 3 0.00
## 4 1.53
## 5 0.19
## 6 0.40
## 7 0.41
## 8 0.38
```

Lado izquierdo del sistema:

```
library(MASS)
Izq=rbind(
  cbind(XpRinvX,XpRinvZ),
  cbind(ZpRinvX,ZpRinvZAinvGinv))
Izqinv=ginv(Izq)
```

Lado derecho del sistema:

```
Der=rbind(
  X_p8pRinvy_p8,X_p12pRinvy_p12,
  Z_p8pRinvy_p8,Z_p12pRinvy_p12)
```

Solución al sistema:

```
Sol=as.matrix(round(Izqinv%*%Der,2))
colnames(Sol)=c("Sol")
rownames(Sol)=paste0(rownames(Der),
    c(rep("_p8",ncol(X_p8)),rep("_p12",ncol(X_p12)),
      rep("_p8",ncol(Z_p8)),rep("_p12",ncol(Z_p12))))
```

```
Sol
```

```
##                Sol
## G_p81_p8    765.0
## G_p82_p8    735.9
## G_p121_p12 939.0
## G_p122_p12 867.7
## 1_p8         22.5
## 2_p8         22.5
## 3_p8        -22.9
## 4_p8        -22.1
## 5_p8         36.1
## 6_p8         31.4
## 7_p8        -42.1
## 8_p8        -25.8
## 1_p12        11.1
## 2_p12        11.1
## 3_p12       -11.7
## 4_p12       -10.4
## 5_p12        16.1
## 6_p12        17.2
## 7_p12       -26.6
## 8_p12        -7.3
```

Efectos fijos:

```
solfijos=as.matrix(Sol[1:4,])
```

Valores genéticos:

```
library(stringr)
VG_p8=data.frame(Sol[paste0(colnames(Z_p8),"_p8"),])
colnames(VG_p8)=c("VG_p8")
VG_p8$animal=word(rownames(VG_p8), 1, sep = fixed('_'))
```

```
VG_p12=data.frame(Sol[paste0(colnames(Z_p12),"_p12"),])
colnames(VG_p12)=c("VG_p12")
VG_p12$animal=word(rownames(VG_p12), 1, sep = fixed('_'))
```

Diferencia esperada de progenie:

```
VG_p8$DEP_p8=VG_p8$VG_p8/2
```

```
VG_p12$DEP_p12=VG_p12$VG_p12/2
```

Valores de la diagonal del lado izquierdo de las ecuaciones de modelo mixto relacionados con los animales:

```
diagonal=as.matrix(diag(Izqinv[5:20,5:20]))
```

Exactitudes, confiabilidades y errores estándar de las predicciones:

```
VG_p8$Confiab_p8=round(1-(diagonal[1:8]%x%(1/var_a11)),2)
```

```
VG_p12$Confiab_p12=round(1-(diagonal[9:16]%x%(1/var_a22)),2)
```

```
#para tenerla en cuenta, la matriz inversa de G:
#G*G^-1=I
#Entonces el primer elemeto es:
#g11=
round((1-(cov_a12*Ginv[2,1]))/var_a11,3)


## [1] 0.005


round((1-(cov_a12*Ginv[2,1]))/var_a22,3)


## [1] 0.004


round(1/var_a11,3)


## [1] 0.003


# valor de la inversa de G del modelo unicaracter
```

```
VG_p8$Exact_p8=sqrt(VG_p8$Confiab_p8)
VG_p12$Exact_p12=sqrt(VG_p12$Confiab_p12)
```

```
VG_p8$SEP_p8=round(sqrt(diagonal[1:8]),2)
VG_p12$SEP_p12=round(sqrt(diagonal[9:16]),2)
```

6 CAPÍTULO SEIS

CRUZAMIENTO

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

6.1. Generalidades

El cruzamiento comprende el apareamiento de individuos pertenecientes a poblaciones diferentes, lo que conlleva a que los individuos cruzados tengan mayores probabilidades de que presenten mayor heterocigosis (H). La diferencia entre el promedio del desempeño de poblaciones cruzadas y el promedio de desempeño de las poblaciones puras se denomina heterosis (η) o también denominada vigor híbrido, donde se espera que los individuos cruzados tengan un mejor desarrollo y adaptación, teniendo como base la idea de que individuos heterocigóticos presentan mayor versatilidad [37, 38].

El cruzamiento de individuos puede ser entre razas o poblaciones de una misma especie o entre especies. En producción animal, lo más frecuente es el cruzamiento entre razas, tema que aboradermos en este capítulo. Veamos una representación del desempeño de animales puros y cruzados en la FIGURA NRO. 6.1

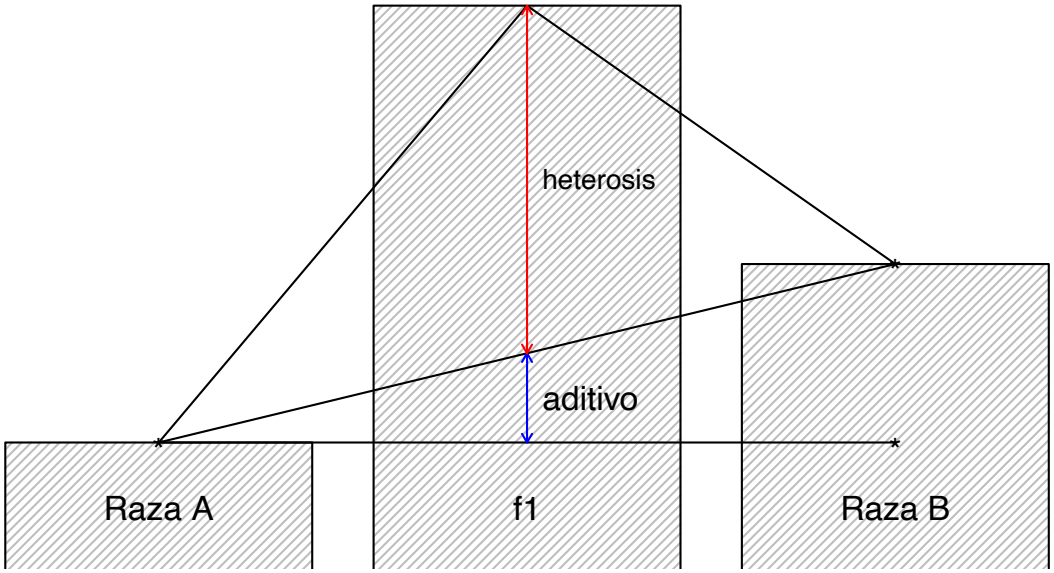

***Figura* 6.1:** Efecto genético aditivo y de heterosis de grupos genéticos puros y cruzados.
Fuente: elaboración propia generada en R-project [25].

Un modelo general para estimar heterosis puede ser descrito de la siguiente forma:

$$y = A\alpha + B\beta + H\eta + e, \text{ tal que } : A + B = 1 \tag{6.1}$$

Donde:

y es la variable respuesta de los n individuos.

A es el vector que contiene las proporciones de la raza A que tiene cada individuo. Por ejemplo, si el individuo es puro de la raza A la proporción es 1; si es $F1$ sería 0.5 y si es un $\frac{3}{4}$ de A $\frac{1}{4}$ de B, sería 0.75.

B es el vector que contiene las proporciones de la raza B que tiene cada individuo. Teniendo en cuenta los tres individuos anteriores, las proporciones serían 0, 0.5 y 0.25.

En los tres casos, la suma de las proporciones de A y B de cada individuo es 1.

H es un vector que contiene las proporciones de heterocigosis de cada individuo, en el caso de los tres individuos anteriores, sería 0, 1.0 y 0.5 (ver cálculos más adelante).

α, β son escalares correspondientes a las incógnitas relacionadas a los efectos de razas (A y B).

η es un escalar correspondiente a la incógnita asociada a la heterosis.

Si se emplea el método de los mínimos cuadrados ordinarios para estimar los parámetros del modelo, se tendría:

$$[X^T X]\begin{bmatrix}\widehat{\alpha}\\ \widehat{\beta}\\ \widehat{\eta}\end{bmatrix} = [X^T y]$$

Donde:

$$X_{n\text{x}3} = \begin{bmatrix} x_{1A} & x_{1B} & x_{1H} \\ x_{2A} & x_{1B} & x_{1H} \\ . & . & . \\ . & . & . \\ x_{nA} & x_{nB} & x_{nH} \end{bmatrix}$$

Esto puede escribirse como:

$$y = X\beta + e$$

$$X = [A \vdots B \vdots H]$$

$$\beta = \begin{bmatrix}\alpha\\ \beta\\ \eta\end{bmatrix}$$

El parámetro η estará dado por la proporción de individuos que son heterocigóticos, producto del apareamiento de individuos de diferentes razas, teniendo en cuenta el siguiente esquema general para el caso de dos razas:

		Madres	
		A_m	B_m
Padres	A_p |	A_pA_m	A_pB_m
	B_p |	B_pA_m	B_pB_m

La heterocigosis de cada individuo (H) se calcula mediante la fórmula:

$$H = 1 - [(A_p * A_m) + (B_p * B_m)]$$

Para k razas:

$$H = 1 - \sum_{i=1}^{k} R_{p_i} * R_{m_i}$$

Donde R_{p_i} y R_{m_i} son las fracciones de cada una de las razas involucradas de cada progenitor del individuo.

En el caso de un individuo $\frac{1}{2}A\frac{1}{2}B$, donde el padre es A y la madre B, H sería:

$$H = 1 - [(A_p * A_m) + (B_p * B_m)]$$

$$H = 1 - [(1 * 0) + (0 * 1)]$$

$$H = 1 - 0 = 1$$

En el caso de un individuo $\frac{1}{2}A\frac{1}{2}B$, donde el padre es $\frac{1}{2}A\frac{1}{2}B$ y la madre $\frac{1}{2}A\frac{1}{2}B$, H, se sería:

$$H = 1 - [(A_p * A_m) + (B_p * B_m)]$$

$$H = 1 - [(0.5 * 0.5) + (0.5 * 0.5)]$$

$$H = 1 - 0.5 = 0.5$$

En el caso de un individuo $\frac{3}{4}A\frac{1}{4}B$, donde el padre es A y la madre $\frac{1}{2}A\frac{1}{2}B$, la H, se tendría:

$$H = 1 - [(A_p * A_m) + (B_p * B_m)]$$

$$H = 1 - [(1 * 0.5) + (0 * 0.5)]$$

$$H = 1 - 0.5 = 0.5$$

Veamos EJEMPLO 6.1 con información de peso de 8 novillas a los 18 meses para evaluar el cruzamiento de la raza A y la raza B, donde la población $\frac{1}{2}A\frac{1}{2}B$ presentó un mayor peso que el promedio de los individuos A y B. ¿Cuáles serían las soluciones del sistema de ecuaciones para α, β y η?. La base de datos se presenta a TABLA NRO. 6.1.

En la TABLA NRO. 6.2 se presenta la información de la composición genética de individuos cruzados.

La base de datos con la información de animales, proporciones raciales y heterocigosis está en la TABLA NRO. 6.3:

Tabla 6.1: Información de animales puros y cruzados

Identificación	Sexo	GG	Peso
A1A2	2	A	401
A3A4	2	A	401
B1B2	2	B	457
B3B4	2	B	412
A1B2	2	AB	473
A3B4	2	AB	477
B1A2	2	BA	499
B3A4	2	BA	483

Nota: GG= grupo genético puro o cruzado de las razas *A* y *B*.
Fuente: elaboración propia (2024).

Tabla 6.2: Proporciones de las razas *A* y *B* y de heterocigosis (*H*) de grupos genéticos.

GG	Bp	Ap	Bm	Am	A	B	H
A	0.00	1.00	0.00	1.00	1.00	0.00	0.00
B	1.00	0.00	1.00	0.00	0.00	1.00	0.00
BA	1.00	0.00	0.00	1.00	0.50	0.50	1.00
AB	0.00	1.00	1.00	0.00	0.50	0.50	1.00

Nota: GG= grupo genético puro o cruzado de las razas *A* y *B*,
p= padre y m= madre.
Fuente: elaboración propia (2024).

Tabla 6.3: Proporciones de la raza *A* y *B* y heterocigosis (*H*) de animales puros y cruzados.

GG	id	Sexo	Peso	Bp	Ap	Bm	Am	A	B	H
A	A1A2	2	401.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
A	A3A4	2	401.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
B	B1B2	2	457.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
B	B3B4	2	412.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
AB	A1B2	2	473.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
AB	A3B4	2	477.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
BA	B1A2	2	499.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00
BA	B3A4	2	483.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00

Nota: GG= grupo genético puro o cruzado de las razas *A* y *B*, id=identificación del animal, p= padre y m= madre.
Fuente: elaboración propia (2024).

Las matrices y el sistema de ecuaciones serían:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0.5 & 0.5 & 1 \\ 0.5 & 0.5 & 1 \\ 0.5 & 0.5 & 1 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

$$y = \begin{bmatrix} 401 \\ 401 \\ 457 \\ 412 \\ 473 \\ 477 \\ 499 \\ 483 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 3 & 2 \\ 2 & 2 & 4 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1768 \\ 1835 \\ 1932 \end{bmatrix}$$

Sistema de ecuaciones:

$$\begin{bmatrix} 3 & 1 & 2 \\ 1 & 3 & 2 \\ 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \end{bmatrix} = \begin{bmatrix} 1768 \\ 1835 \\ 1932 \end{bmatrix}$$

El vector solución sería:

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \end{bmatrix} = \begin{bmatrix} 401 \\ 434.5 \\ 65.25 \end{bmatrix}$$

Por consiguiente, los promedios de las razas A y B fueron 401 y 434.5 kg, respectivamente, y la heterosis fue $\eta = 65.25$ kg. Indicando que los individuos cruzados pesaron 65.25 kg más que los individuos de las razas A y B.

Ahora consideremos el caso en el que se tienen otros efectos fijos que influyen en la expresión de la característica. Para el EJEMPLO 6.1 tenemos la existencia de dos fincas como se indica en TABLA NRO. 6.4.

TABLA 6.4: Proporciones de la raza A y B y heterocigosis de animales puros y cruzados en dos fincas.

GG	id	sexo	Finca	Peso	Bp	Ap	Bm	Am	A	B	H
A	A1A2	2	f1	401.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
A	A3A4	2	f1	401.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
B	B1B2	2	f1	457.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
B	B3B4	2	f1	412.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
AB	A1B2	2	f1	473.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
AB	A3B4	2	f1	477.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
BA	B1A2	2	f1	499.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00
BA	B3A4	2	f1	483.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00
A	A1A4	2	f2	440.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
A	A3A2	2	f2	450.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
B	B1B4	2	f2	451.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
B	B3B2	2	f2	446.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
AB	A1B4	2	f2	403.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
AB	A3B2	2	f2	407.00	0.00	1.00	1.00	0.00	0.50	0.50	1.00
BA	B1A4	2	f2	401.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00
BA	B3A2	2	f2	402.00	1.00	0.00	0.00	1.00	0.50	0.50	1.00

Nota: GG= grupo genético puro o cruzado de las razas A y B, id=identificación del animal, p= padre y m= madre y el sexo 2 corresponde a hembras.
Fuente: elaboración propia (2024).

El modelo sería:

$$y = Ff + A\alpha + B\beta + H\eta + e, \text{ con } A + B = 1$$

Donde F es el parámetro asociado a la finca ($f = (1, 2)$), como se mencionó en capítulos anteriores, corresponde a las medias de cada finca.

Las dos primeras columnas de X estarían asociadas a f_1 y f_2 y las siguientes serían A, B y H:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0.5 & 0.5 & 1 \\ 1 & 0 & 0.5 & 0.5 & 1 \\ 1 & 0 & 0.5 & 0.5 & 1 \\ 1 & 0 & 0.5 & 0.5 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0.5 & 0.5 & 1 \\ 0 & 1 & 0.5 & 0.5 & 1 \\ 0 & 1 & 0.5 & 0.5 & 1 \\ 0 & 1 & 0.5 & 0.5 & 1 \end{bmatrix}$$

$$y = \begin{bmatrix} 401 \\ 401 \\ 457 \\ 412 \\ 473 \\ 477 \\ 499 \\ 483 \\ 440 \\ 450 \\ 451 \\ 446 \\ 403 \\ 407 \\ 401 \\ 402 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 8 & 0 & 4 & 4 & 4 \\ 0 & 8 & 4 & 4 & 4 \\ 4 & 4 & 6 & 2 & 4 \\ 4 & 4 & 2 & 6 & 4 \\ 4 & 4 & 4 & 4 & 8 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 3603 \\ 3400 \\ 3464.5 \\ 3538.5 \\ 3545 \end{bmatrix}$$

Sistema de ecuaciones:

$$\begin{bmatrix} 8 & 0 & 4 & 4 & 4 \\ 0 & 8 & 4 & 4 & 4 \\ 4 & 4 & 6 & 2 & 4 \\ 4 & 4 & 2 & 6 & 4 \\ 4 & 4 & 4 & 4 & 8 \end{bmatrix} \begin{bmatrix} \widehat{f_1} \\ \widehat{f_2} \\ \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \end{bmatrix} = \begin{bmatrix} 3603 \\ 3400 \\ 3464.5 \\ 3538.5 \\ 3545 \end{bmatrix}$$

El vector solución sería:

$$\begin{bmatrix} \widehat{f_1} \\ \widehat{f_2} \\ \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \end{bmatrix} = \begin{bmatrix} 228.81 \\ 203.44 \\ 206.88 \\ 225.38 \\ 10.87 \end{bmatrix}$$

Los valores esperados para los individuos según el grupo genético y la finca serían:

Raza A en la finca 1: $\widehat{f_1} + \widehat{\alpha} = 228.81 + 206.88 = 435.69$

Raza B en la finca 1: $\widehat{f_1} + \widehat{\beta} = 228.81 + 225.38 = 454.19$

Raza A en la finca 2: $\widehat{f_2} + \widehat{\alpha} = 203.44 + 206.88 = 410.32$

Raza B en la finca 2: $\widehat{f_2} + \widehat{\beta} = 203.44 + 225.38 = 428.82$

Cruzados $(\frac{1}{2}A\frac{1}{2}B)$ en la finca 1: $\widehat{f_1} + \frac{1}{2}\widehat{\alpha} + \frac{1}{2}\widehat{\beta} + \widehat{\eta} = 228.81 + \frac{1}{206.88} + \frac{1}{225.38} + 10.87 = 455.81$

Cruzados $(\frac{1}{2}A\frac{1}{2}B)$ en la finca 2: $\widehat{f_2} + \frac{1}{2}\widehat{\alpha} + \frac{1}{2}\widehat{\beta} + \widehat{\eta} = 203.44 + \frac{1}{206.88} + \frac{1}{225.38} + 10.87 = 430.44$

Si tenemos el efecto de la interacción (I) genotipo (grupo genético) y ambiente (finca), el modelo sería:

$$y = Ff + A\alpha + B\beta + H\eta + I_{FA}(f\alpha) + I_{FB}(f\beta) + I_{FH}(f\eta) + e, \text{ con } A + B = 1$$

La base de datos se presenta en la TABLA NRO. 6.5, donde las columnas correspondientes a las interacciones se obtienen multiplicando las columnas implicadas en cada caso, por ejemplo, la columna nueve (B1) corresponde a la multiplicación de los valores de la segunda columna (finca 1) con la quinta columna (columna 5) (raza *B*):

TABLA 6.5: Proporciones de la raza *A* y *B* y heterocigosis (*H*) de animales puros y cruzados en dos fincas (1 y 2), teniendo en cuenta la interacción genotipo y ambiente.

id	finca 1	finca 2	A	B	H	A1	A2	B1	B2	H1	H2
A1A2	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
A3A4	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
B1B2	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
B3B4	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
A1B2	1.00	0.00	0.50	0.50	1.00	0.50	0.00	0.50	0.00	1.00	0.00
A3B4	1.00	0.00	0.50	0.50	1.00	0.50	0.00	0.50	0.00	1.00	0.00
B1A2	1.00	0.00	0.50	0.50	1.00	0.50	0.00	0.50	0.00	1.00	0.00
B3A4	1.00	0.00	0.50	0.50	1.00	0.50	0.00	0.50	0.00	1.00	0.00
A1A4	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
A3A2	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
B1B4	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
B3B2	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
A1B4	0.00	1.00	0.50	0.50	1.00	0.00	0.50	0.00	0.50	0.00	1.00
A3B2	0.00	1.00	0.50	0.50	1.00	0.00	0.50	0.00	0.50	0.00	1.00
B1A4	0.00	1.00	0.50	0.50	1.00	0.00	0.50	0.00	0.50	0.00	1.00
B3A2	0.00	1.00	0.50	0.50	1.00	0.00	0.50	0.00	0.50	0.00	1.00

Nota: id=identificación de animales puros y cruzados de las razas *A* y *B*.
Fuente: elaboración propia (2024).

Las matriz X relaciona en sus dos primeras columnas a f_1 y f_2 (finca 1 y finca 2) y las siguientes estarían relacionadas a: α (correspondiente a la raza A), β (correspondiente a la raza B), η (correspondiente a la heterosis), $\alpha f_1 A$ (correspondiente a la raza A en la finca 1), $\alpha f_2 A$ (correspondiente a la raza A en la finca 2), $\beta f_1 B$ (correspondiente a la raza B en la finca 1), $\beta f_2 B$ (correspondiente a la raza B en la finca 2), $\eta f_1 H$ (correspondiente a la heterosis en la finca 1) y $\eta f_2 H$ (correspondiente a la heterosis en la finca 2):

$$X = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 8 & 0 & . & . & 4 & 0 \\ 0 & 8 & . & . & 0 & 4 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 4 & 0 & . & . & 4 & 0 \\ 0 & 4 & . & . & 0 & 4 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 3603 \\ 3400 \\ 3464.5 \\ 3538.5 \\ 3545 \\ 1768 \\ 1696.5 \\ 1835 \\ 1703.5 \\ 1932 \\ 1613 \end{bmatrix}$$

El sistema de ecuaciones sería:

$$\begin{bmatrix} 8 & 0 & . & . & 4 & 0 \\ 0 & 8 & . & . & 0 & 4 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 4 & 0 & . & . & 4 & 0 \\ 0 & 4 & . & . & 0 & 4 \end{bmatrix} \begin{bmatrix} \widehat{f_1} \\ \widehat{f_2} \\ \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \\ \widehat{\alpha_{f1}} \\ \widehat{\alpha_{f2}} \\ \widehat{\beta_{f1}} \\ \widehat{\beta_{f2}} \\ \widehat{\eta_{f1}} \\ \widehat{\eta_{f2}} \end{bmatrix} = \begin{bmatrix} 3603 \\ 3400 \\ 3464.5 \\ 3538.5 \\ 3545 \\ 1768 \\ 1696.5 \\ 1835 \\ 1703.5 \\ 1932 \\ 1613 \end{bmatrix}$$

El vector de soluciones sería:

$$\begin{bmatrix} \widehat{f_1} \\ \widehat{f_2} \\ \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\eta} \\ \widehat{\alpha_{f1}} \\ \widehat{\alpha_{f2}} \\ \widehat{\beta_{f1}} \\ \widehat{\beta_{f2}} \\ \widehat{\eta_{f1}} \\ \widehat{\eta_{f2}} \end{bmatrix} = \begin{bmatrix} 163.23 \\ 182.57 \\ 166.73 \\ 179.07 \\ 7.25 \\ 71.03 \\ 95.7 \\ 92.2 \\ 86.87 \\ 58 \\ -50.75 \end{bmatrix}$$

En características influenciadas por el ambiente materno, es necesario estudiar la habilidad materna atribuida al efecto de la heterocigosis por cruzamiento de la madre (H_m). La diferencia entre el promedio de desempeño de individuos con madres cruzadas e individuos con madres puras se denomina heterosis materna (η_m). Para el cado de dos razas, la H_m se calcula teniendo en cuenta el grupo genético de los abuelos maternos, asi:

$$H_m = 1 - [(A_{p_m} * A_{m_m}) + (B_{p_m} * B_{m_m})]$$

Para k razas se tendría:

$$H_m = 1 - \sum_{i=1}^{k} R_{p_m i} * R_{m_m i}$$

Donde $R_{p_m i}$ y $R_{m_m i}$ son las fracciones de cada una de las razas involucradas en el abuelo materno y la abuela materna del i-ésimo individuo.

En el caso de un individuo con composición $\frac{1}{2}A\frac{1}{2}B$, donde el padre es $\frac{1}{2}A\frac{1}{2}B$, la madre $\frac{1}{2}A\frac{1}{2}B$, el abuelo materno A y la abuela materna B, se tendría:

$$H_m = 1 - [(A_{p_m} * A_{m_m}) + (B_p * B_{m_m})]$$

$$H_m = 1 - [(1 * 0) + (0 * 1)]$$

$$H_m = 1 - 0 = 1$$

En el caso de un individuo $\frac{3}{4}A\frac{1}{4}B$, donde el padre es A, la madre $\frac{1}{2}A\frac{1}{2}B$, el abuelo materno A y la abuela materna B, se tendría:

$$H_m = 1 - [(A_{p_m} * A_{m_m}) + (B_p * B_{m_m})]$$

$$H_m = 1 - [(1 * 0) + (0 * 1)]$$

$$H_m = 1 - 0 = 1$$

Cuando se analizan grupos genéticos posteriores a la primera generación cruzada (conocidad como $F1$), es necesario tener en cuenta la recombinación genética (Ψ) causada por la segregación independiente de combinaciones diferentes a los gametos parentales.

$$\Psi = [\frac{1}{2}F_p + \frac{1}{2}F_m]$$

Donde F_p y F_M son las fracciones esperadas de las recombinaciones de los gametos del padre y de la madre, respectivamente.

En el caso de un individuo $\frac{1}{2}A\frac{1}{2}B$, donde el padre es $\frac{1}{2}A\frac{1}{2}B$, la madre $\frac{1}{2}A\frac{1}{2}B$, el abuelo materno A y la abuela materna B, se tendría:

$$\Psi = [\frac{1}{2}0.5 + \frac{1}{2}0.5] = 0.5$$

En el caso de un individuo $\frac{3}{4}A\frac{1}{4}B$, donde el padre es A, la madre $\frac{1}{2}A\frac{1}{2}B$, el abuelo materno A y la abuela materna B, se tendría:

$$\Psi = [\frac{1}{2}0.0 + \frac{1}{2}0.5] = 0.25$$

6.2. Modelo animal multirracial

Para la realización de evaluaciones genéticas de animales puros y cruzados utilizando un modelo animal con efectos aditivos y no aditivos, utilizaremos los procedimientos descritos por Arnold et al [39]. Uno modelo multirracial básico, es descrito de la siguiente forma:

$$y = X\beta + ZQ\phi + Za + Ps\eta + PTh + e \tag{6.2}$$

Donde:

y es el vector de observaciones, X es la matriz de diseño que relaciona los efectos fijos β con los registros, Z es la matriz de diseño que relaciona los registros con el efecto aleatorio genético aditivo (a) y e es un vector de errores.

Las matrices y los vectores anteriores ya los hemos visto en modelos animales estudiados previamente. Detallaremos a continuación Q, P T y s:

Q es la matriz de diseño que relaciona los individuos con el efecto fijo de grupo genético (ϕ). En las filas están los individuos y en las columnas las razas. Cada elemento tendrá la proporción correspondiente de cada raza y varía de 0 a 1. Por ejemplo, si hay dos razas (A y B) y cuatro individuos de los grupos genéticos A, B, $\frac{1}{2}A\frac{1}{2}B$ y $\frac{1}{4}A\frac{3}{4}B$; Q sería:

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}$$

P es la matriz de diseño que relaciona los individuos con el efecto genético no aditivo (h). En las filas estarán los individuos con registro, en las columnas estarán los individuos de la genealogía, y tendrá valores de ceros y unos. Los unos estarán en los elementos que corresponden a los individuos cruzados, continuando con el ejemplo anterior, sería:

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

El vector s contiene la relación de los individuos con el efecto fijo de heterosis (η). En las filas estarán los individuos y los elementos tendrán valores de 0 a 1 con la proporción de heterocigosis por cruzamiento, así:

$$s = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0.5 \end{bmatrix}$$

La matriz T es una matriz cuadrada de tamaño igual al número de individuos. Sus elementos son 0, excepto en las filas correspondientes a los animales cruzados y las columnas de sus padres; las cuales tienen los valores de heterocigosis por cruzamiento del individuo.

Continuando con el ejemplo, y considerando que los individuos de los grupos genéticos A y B son los padres del tercer individuo de grupo genético $\frac{1}{2}A\frac{1}{2}B$ y que el cuarto individuo, es de grupo genético $\frac{1}{4}A\frac{3}{4}B$ cuyo padre es el tercer individuo, la matriz T sería:

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{bmatrix}$$

La matriz del lado izquierdo del sistema de ecuaciones de modelo mixto sería:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}ZQ & X^TR^{-1}Z & X^TR^{-1}Ps & X^TR^{-1}PT \\ Q^TZ^TR^{-1}X & Q^TZ^TR^{-1}ZQ & Q^TZ^TR^{-1}Z & Q^TZ^TR^{-1}Ps & Q^TZ^TR^{-1}PT \\ Z^TR^{-1}X & Z^TR^{-1}ZQ & Z^TR^{-1}Z + G^{-1} & Z^TR^{-1}Ps & Z^TR^{-1}PT \\ s^TP^TR^{-1}X & s^TP^TR^{-1}ZQ & s^TP^TR^{-1}Z & s^TP^TR^{-1}Ps & s^TP^TR^{-1}PT \\ T^TP^TR^{-1}X & T^TP^TR^{-1}ZQ & T^TP^TR^{-1}Z & T^TP^TR^{-1}Ps & T^TP^TR^{-1}PT + H^{-1} \end{bmatrix}$$

El vector de incógnitas sería:

El vector de incógnitas sería:

$$\begin{bmatrix} \widehat{\beta} \\ \widehat{\phi} \\ \widehat{a} \\ \widehat{\eta} \\ \widehat{h} \end{bmatrix}$$

El vector del lado derecho sería:

$$\begin{bmatrix} X^T R^{-1} y \\ Q^T Z^T R^{-1} y \\ Z^T R^{-1} y \\ s^T P^T R^{-1} y \\ T^T P^T R^{-1} y \end{bmatrix}$$

La inversa de la matriz de relaciones genéticas G^{-1} es el resultado de la combinaciones de A^{-1} con las inversas de las varianzas genéticas de cada una de los grupos genéticos (mayores detalles en el capítulo 14). Las matrices H^{-1} y R^{-1} contienen en la diagonal las varianzas de cada grupo genético para los efectos de la heterocigosis y de los errores, respectivamente.

El valor genético aditivo (VG_a) de un individuo estaría dado por las soluciones del efecto fijo genético aditivo según sus proporciones raciales y la solución del efecto aleatorio del individuo ($VG_a = Q\phi + a$). El valor genético no aditivo (VG_h) de un individuo estaría dado por la solución de la heterosis según su proporción de heterocigosis por cruzamiento y la solución genética no aditiva correspondiente ($VG_h = s\eta + Th$).

Ahora desarrollaremos el EJEMPLO 6.2 con un modelo animal multirracial para la variable ≪peso≫ con individuos de dos razas y sus cruces, teniendo en cuenta que:

Hay tres animales de la raza A ($A1$, $A2$ y $A1A2$).

Cuatro de la raza B ($B1$,$B2$,$B3$, y $B1B2$).

Un animal $\frac{1}{2}A\frac{1}{2}B$.

Un animal $\frac{1}{2}B\frac{1}{2}A$.

Un animal $\frac{3}{4}B\frac{1}{4}A$.

Hay 10 individuos en la genealogía y un individuo sin registro de producción y el efecto fijo es el sexo. Veamos los datos en TABLA NRO. 6.6.

TABLA 6.6: Información genéalógica y productiva de invididuos puros y cruzados.

	Identificación	padre	madre	sexo	Grupo genético	Peso
1	A1			1	A	432
2	A2			2	A	
3	B1			1	B	401
4	B2			2	B	411
5	B3			1	B	423
6	A1A2	A1	A2	2	A	432
7	B1B2	B1	B2	2	B	423
8	A1B2	A1	B2	2	AB	421
9	B3A2	B3	A2	1	BA	432
10	B3A1B2	B3	A1B2	2	BBA	332

Nota: información de animales puros y cruzados de las razas A y B, el sexo 1 es macho y el sexo 2 es hembra.
Fuente: elaboración propia (2024).

La matriz de parentesco (A) tendría un tamaño igual al número de individuos en la genealogía (10), y sería:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0.25 \\ 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 & 0 & 1 & 0 & 0.25 & 0.25 & 0.12 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0.25 & 0 & 0.12 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0.25 & 0.25 & 1 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0 & 0.5 & 0.25 & 0 & 0 & 1 & 0.25 \\ 0.25 & 0 & 0 & 0.25 & 0.5 & 0.12 & 0.12 & 0.5 & 0.25 & 1 \end{bmatrix}$$

La matriz Z relaciona los individuos con registros (filas) y los individuos en la genealogía (columnas), tendrá valores de uno en los elementos de la matriz que

relaciona al individuo con su registro.

$$Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La matriz X relaciona los individuos con registro (filas) y los niveles de efecto fijo, que en este caso son los sexos.

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La matriz P relaciona los individuos con registros (filas) y los individuos en la genealogía (columnas), tendrá valores de uno en los elementos que relacionan a los correspondientes individuos cruzados.

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La matriz Q relaciona los individuos (filas) con sus proporciones raciales (columnas), quedando:

$$Q = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}$$

El vector s relaciona los registros (filas) con su correspondiente proporción de heterocigosis debida al cruzamiento.

$$s = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0.5 \end{bmatrix}$$

La matriz cuadrada T de tamaño igual al número de individuos en la genealogía, tendrá los valores de heterocigosis en la relación de hijos (filas) con sus padres (columnas).

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 \end{bmatrix}$$

El vector de observaciones y, será:

$$y = \begin{bmatrix} 432 \\ 401 \\ 411 \\ 423 \\ 432 \\ 423 \\ 421 \\ 432 \\ 332 \end{bmatrix}$$

Cada grupo racial tendrá sus valores de varianzas, sin embargo, debido a la complejidad que presentan los modelos de evaluación genética multirracial y a la estructura de las bases de datos, y con el propósito de simplificar el ejemplo, podemos considerar que la población pura y cruzada tiene las mismas varianzas, siendo: $\sigma_a^2 = 20$, $\sigma_h^2 = 10$ y $\sigma_e^2 = 70$. Por consiguiente, las matrices que involucran las varianzas serían:

$$G^{-1} = \frac{1}{\sigma_a^2} A^{-1} = 0.1A^{-1} =$$

$$\begin{bmatrix} 0.1 & 0.03 & 0 & 0.03 & 0 & -0.05 & 0 & -0.05 & 0 & 0 \\ 0.03 & 0.1 & 0 & 0 & 0.03 & -0.05 & 0 & 0 & -0.05 & 0 \\ 0 & 0 & 0.08 & 0.03 & 0 & 0 & -0.05 & 0 & 0 & 0 \\ 0.03 & 0 & 0.03 & 0.1 & 0 & 0 & -0.05 & -0.05 & 0 & 0 \\ 0 & 0.03 & 0 & 0 & 0.1 & 0 & 0 & 0.03 & -0.05 & -0.05 \\ -0.05 & -0.05 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.05 & -0.05 & 0 & 0 & 0.1 & 0 & 0 & 0 \\ -0.05 & 0 & 0 & -0.05 & 0.03 & 0 & 0 & 0.12 & 0 & -0.05 \\ 0 & -0.05 & 0 & 0 & -0.05 & 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & -0.05 & 0 & 0 & -0.05 & 0 & 0.1 \end{bmatrix}$$

$$R^{-1} = \begin{bmatrix} 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 \end{bmatrix}$$

$$H^{-1} = \begin{bmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}$$

Las operaciones entre matrices que conforman el sistema de ecuaciones serían:

$$X^T R^{-1} X = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.07 \end{bmatrix}$$

$$X^T R^{-1} ZQ = \begin{bmatrix} 0.02 & 0.04 \\ 0.03 & 0.05 \end{bmatrix}$$

$$X^T R^{-1} Z = \begin{bmatrix} 0.01 & 0 & 0.01 & 0 & 0.01 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 0.01 & 0.01 & 0.01 & 0 & 0.01 \end{bmatrix}$$

$$X^T R^{-1} Ps = \begin{bmatrix} 0.01 \\ 0.02 \end{bmatrix}$$

$$X^T R^{-1} PT = \begin{bmatrix} 0 & 0.01 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0.01 & 0.01 & 0 & 0 & 0.01 & 0 & 0 \end{bmatrix}$$

$$Q^T Z^T R^{-1} ZQ = \begin{bmatrix} 0.04 & 0.01 \\ 0.01 & 0.07 \end{bmatrix}$$

$$Q^T Z^T R^{-1} Z = \begin{bmatrix} 0.01 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0.01 & 0.01 & 0 \\ 0 & 0 & 0.01 & 0.01 & 0.01 & 0 & 0.01 & 0.01 & 0.01 & 0.01 \end{bmatrix}$$

$$Q^T Z^T R^{-1} Ps = \begin{bmatrix} 0.02 \\ 0.02 \end{bmatrix}$$

$$Q^T Z^T R^{-1} PT = \begin{bmatrix} 0.01 & 0.01 & 0 & 0.01 & 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0.01 & 0 & 0.01 & 0.01 & 0 & 0 & 0.01 & 0 & 0 \end{bmatrix}$$

$$Z^T R^{-1} Z + G^{-1} =$$

$$\begin{bmatrix} 0.11 & 0.03 & 0 & 0.03 & 0 & -0.05 & 0 & -0.05 & 0 & 0 \\ 0.03 & 0.1 & 0 & 0 & 0.03 & -0.05 & 0 & 0 & -0.05 & 0 \\ 0 & 0 & 0.09 & 0.03 & 0 & 0 & -0.05 & 0 & 0 & 0 \\ 0.03 & 0 & 0.03 & 0.11 & 0 & 0 & -0.05 & -0.05 & 0 & 0 \\ 0 & 0.03 & 0 & 0 & 0.11 & 0 & 0 & 0.03 & -0.05 & -0.05 \\ -0.05 & -0.05 & 0 & 0 & 0 & 0.11 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.05 & -0.05 & 0 & 0 & 0.11 & 0 & 0 & 0 \\ -0.05 & 0 & 0 & -0.05 & 0.03 & 0 & 0 & 0.14 & 0 & -0.05 \\ 0 & -0.05 & 0 & 0 & -0.05 & 0 & 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 & -0.05 & 0 & 0 & -0.05 & 0 & 0.11 \end{bmatrix}$$

$$Z^T R^{-1} Ps = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.01 \\ 0.01 \\ 0.01 \end{bmatrix}$$

$$Z^T R^{-1} PT = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0.01 & 0 & 0 \end{bmatrix}$$

$$s^T P^T R^{-1} Ps = \begin{bmatrix} 0.03 \end{bmatrix}$$

$$s^T P^T R^{-1} PT = \begin{bmatrix} 0.01 & 0.01 & 0 & 0.01 & 0.02 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$T^T P^T R^{-1} PT + H^{-1} = \begin{bmatrix} 0.11 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.11 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0.11 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}$$

$$XTR^{-1}y = \begin{bmatrix} 24.11 \\ 28.84 \end{bmatrix}$$

$$Q^T Z^T R^{-1} y = \begin{bmatrix} 19.62 \\ 33.34 \end{bmatrix}$$

$$Z^T R^{-1} y = \begin{bmatrix} 6.17 \\ 0 \\ 5.73 \\ 5.87 \\ 6.04 \\ 6.17 \\ 6.04 \\ 6.01 \\ 6.17 \\ 4.74 \end{bmatrix}$$

$$s^T P^T R^{-1} y = \begin{bmatrix} 14.56 \end{bmatrix}$$

$$T^T P^T R^{-1} y = \begin{bmatrix} 6.01 \\ 6.17 \\ 0 \\ 6.01 \\ 8.54 \\ 0 \\ 0 \\ 2.37 \\ 0 \\ 0 \end{bmatrix}$$

Las soluciones del efecto fijo no genético son:

$$\begin{bmatrix} \widehat{Sexo_1} \\ \widehat{Sexo_2} \end{bmatrix} = \begin{bmatrix} 219.2 \\ 200.4 \end{bmatrix}$$

Soluciones al efecto fijo genético aditivo:

$$\begin{bmatrix} \widehat{Raza_A} \\ \widehat{Raza_B} \end{bmatrix} = \begin{bmatrix} 223 \\ 196.7 \end{bmatrix}$$

Soluciones al efecto genético aditivo:

$$\begin{bmatrix} \widehat{A1_a} \\ \widehat{A2_a} \\ \widehat{B1_a} \\ \widehat{B2_a} \\ \widehat{B3_a} \\ \widehat{A1A2_a} \\ \widehat{B1B2_a} \\ \widehat{A1B2_a} \\ \widehat{B3A2_a} \\ \widehat{B3A1B2_a} \end{bmatrix} = \begin{bmatrix} -2.7 \\ 2.7 \\ -1 \\ 4.2 \\ -3.2 \\ 1.1 \\ 4.6 \\ -0.8 \\ 1.3 \\ -9.8 \end{bmatrix}$$

Soluciones al efecto fijo de la heterosis (η):

$$\widehat{\eta} = \begin{bmatrix} -8.9 \end{bmatrix}$$

Soluciones al efecto genético no aditivo (h):

$$\begin{bmatrix} \widehat{A1_h} \\ \widehat{A2_h} \\ \widehat{B1_h} \\ \widehat{B2_h} \\ \widehat{B3_h} \\ \widehat{A1A2_h} \\ \widehat{B1B2_h} \\ \widehat{A1B2_h} \\ \widehat{B3A2_h} \\ \widehat{B3A1B2_h} \end{bmatrix} = \begin{bmatrix} 2.3 \\ 1.6 \\ 0 \\ 2.3 \\ -2.3 \\ 0 \\ 0 \\ -3.9 \\ 0 \\ 0 \end{bmatrix}$$

6.3. Ejercicios en R-project

Utilizaremos las librerías ≪doBy≫ [40] para organizar datos, ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular las inversas requeridas.

Iniciemos con el desarrollo del EJEMPLO 6.1:

```
BD=data.frame(matrix(ncol=4,byrow=TRUE, c(
  "A1A2",2,"A",401,
  "A3A4",2,"A",401,
  "B1B2",2,"B",457,
  "B3B4",2,"B",412,
  "A1B2",2,"AB",473,
  "A3B4",2,"AB",477,
  "B1A2",2,"BA",499,
  "B3A4",2,"BA",483)))
colnames(BD)=c("id","sex","GG","Peso")
BD$Peso=as.numeric(BD$Peso)
```

```
BD
```

```
##     id sex GG Peso
## 1 A1A2   2  A  401
## 2 A3A4   2  A  401
## 3 B1B2   2  B  457
## 4 B3B4   2  B  412
## 5 A1B2   2 AB  473
## 6 A3B4   2 AB  477
## 7 B1A2   2 BA  499
## 8 B3A4   2 BA  483
```

```
library(doBy)
medias=summaryBy(Peso~GG, data=BD); medias
```

```
##   GG Peso.mean
## 1  A       401
## 2 AB       475
## 3  B       434
## 4 BA       491
```

Composición genética general de individuos puros y cruzados de dos razas:

```
G=data.frame(matrix(ncol=5,byrow=TRUE,c(
  "A" ,0,1,0,1,
  "B" ,1,0,1,0,
  "BA",1,0,0,1,
  "AB",0,1,1,0
)))
variables=c("GG", "Bp", "Ap", "Bm", "Am")
colnames(G)=variables
```

```
G
```

```
##   GG Bp Ap Bm Am
## 1  A  0  1  0  1
## 2  B  1  0  1  0
## 3 BA  1  0  0  1
## 4 AB  0  1  1  0
```

```
G[,2:5] <- lapply(G[,2:5],
                  function(x) as.numeric(as.character(x)))
str(G)
```

```
## 'data.frame': 4 obs. of  5 variables:
##  $ GG: chr  "A" "B" "BA" "AB"
##  $ Bp: num  0 1 1 0
##  $ Ap: num  1 0 0 1
##  $ Bm: num  0 1 0 1
##  $ Am: num  1 0 1 0
```

```
G$A=(G$Ap+G$Am)/2
```

```
G$B=(G$Bp+G$Bm)/2
```

```
G
```

```
##   GG Bp Ap Bm Am   A   B
## 1  A  0  1  0  1 1.0 0.0
## 2  B  1  0  1  0 0.0 1.0
## 3 BA  1  0  0  1 0.5 0.5
## 4 AB  0  1  1  0 0.5 0.5
```

```
G$H=1-((G$Bp*G$Bm)+(G$Ap*G$Am))
G
```

```
##   GG Bp Ap Bm Am   A   B H
## 1  A  0  1  0  1 1.0 0.0 0
## 2  B  1  0  1  0 0.0 1.0 0
## 3 BA  1  0  0  1 0.5 0.5 1
## 4 AB  0  1  1  0 0.5 0.5 1
```

```
BD$secuencia=seq(1,nrow(BD),1)
BD
```

```
##     id sex GG Peso secuencia
## 1 A1A2   2  A  401         1
## 2 A3A4   2  A  401         2
## 3 B1B2   2  B  457         3
## 4 B3B4   2  B  412         4
## 5 A1B2   2 AB  473         5
## 6 A3B4   2 AB  477         6
## 7 B1A2   2 BA  499         7
## 8 B3A4   2 BA  483         8
```

```
Base=merge(BD,G, all=TRUE)
library(doBy)
Base=orderBy(~secuencia, data=Base)
rownames(Base)=Base$secuencia
Base$secuencia=NULL
BD$secuencia=NULL
Base
```

```
##   GG   id sex Peso Bp Ap Bm Am   A   B H
## 1  A A1A2   2  401  0  1  0  1 1.0 0.0 0
## 2  A A3A4   2  401  0  1  0  1 1.0 0.0 0
## 3  B B1B2   2  457  1  0  1  0 0.0 1.0 0
## 4  B B3B4   2  412  1  0  1  0 0.0 1.0 0
## 5 AB A1B2   2  473  0  1  1  0 0.5 0.5 1
## 6 AB A3B4   2  477  0  1  1  0 0.5 0.5 1
## 7 BA B1A2   2  499  1  0  0  1 0.5 0.5 1
## 8 BA B3A4   2  483  1  0  0  1 0.5 0.5 1
```

Montaje de matrices:

```
X=cbind(Base$A,Base$B,Base$H)
colnames(X)=c("A","B","H")
XpX=t(X)%*%X
XpX


##   A B H
## A 3 1 2
## B 1 3 2
## H 2 2 4
```

```
y=as.matrix(Base$Peso)
y


##      [,1]
## [1,]  401
## [2,]  401
## [3,]  457
## [4,]  412
## [5,]  473
## [6,]  477
## [7,]  499
## [8,]  483
```

```
Xpy=t(X)%*%y
Xpy


##   [,1]
## A 1768
## B 1835
## H 1932
```

```
library(MASS)
XpXinv=ginv(XpX)
Sol=XpXinv%*%Xpy
rownames(Sol)=c("A","B","H")
Sol

##   [,1]
## A  401
## B  434
## H   65
```

Incluyendo el efecto de Finca:

```
BD=data.frame(matrix(ncol=5,byrow=TRUE, c(
  "A1A2",2,"f1","A",401,
  "A3A4",2,"f1","A",401,
  "B1B2",2,"f1","B",457,
  "B3B4",2,"f1","B",412,
  "A1B2",2,"f1","AB",473,
  "A3B4",2,"f1","AB",477,
  "B1A2",2,"f1","BA",499,
  "B3A4",2,"f1","BA",483,
  "A1A4",2,"f2","A",440,
  "A3A2",2,"f2","A",450,
  "B1B4",2,"f2","B",451,
  "B3B2",2,"f2","B",446,
  "A1B4",2,"f2","AB",403,
  "A3B2",2,"f2","AB",407,
  "B1A4",2,"f2","BA",401,
  "B3A2",2,"f2","BA",402)))
colnames(BD)=c("id","sex","Finca","GG","Peso")
BD$Peso=as.numeric(BD$Peso)
```

```
BD
```

```
##      id sex Finca GG Peso
## 1  A1A2   2    f1  A  401
## 2  A3A4   2    f1  A  401
## 3  B1B2   2    f1  B  457
## 4  B3B4   2    f1  B  412
## 5  A1B2   2    f1 AB  473
## 6  A3B4   2    f1 AB  477
## 7  B1A2   2    f1 BA  499
## 8  B3A4   2    f1 BA  483
## 9  A1A4   2    f2  A  440
## 10 A3A2   2    f2  A  450
## 11 B1B4   2    f2  B  451
## 12 B3B2   2    f2  B  446
## 13 A1B4   2    f2 AB  403
## 14 A3B2   2    f2 AB  407
## 15 B1A4   2    f2 BA  401
## 16 B3A2   2    f2 BA  402
```

Información de composición genética general para dos razas:

```
G=data.frame(matrix(ncol=5,byrow=TRUE,c(
  "A" ,0,1,0,1,
  "B" ,1,0,1,0,
  "BA",1,0,0,1,
  "AB",0,1,1,0
)))
variables=c("GG", "Bp", "Ap", "Bm", "Am")
colnames(G)=variables
```

```
G
```

```
##   GG Bp Ap Bm Am
## 1  A  0  1  0  1
## 2  B  1  0  1  0
## 3 BA  1  0  0  1
## 4 AB  0  1  1  0
```

```
G[,2:5] <- lapply(G[,2:5],
                  function(x) as.numeric(as.character(x)))
str(G)
```

```
## 'data.frame': 4 obs. of  5 variables:
##  $ GG: chr  "A" "B" "BA" "AB"
##  $ Bp: num  0 1 1 0
##  $ Ap: num  1 0 0 1
##  $ Bm: num  0 1 0 1
##  $ Am: num  1 0 1 0
```

```
G$A=(G$Ap+G$Am)/2
```

```
G$B=(G$Bp+G$Bm)/2
G
```

```
##   GG Bp Ap Bm Am   A   B
## 1  A  0  1  0  1 1.0 0.0
## 2  B  1  0  1  0 0.0 1.0
## 3 BA  1  0  0  1 0.5 0.5
## 4 AB  0  1  1  0 0.5 0.5
```

```
G$H=1-((G$Bp*G$Bm)+(G$Ap*G$Am))
G
```

```
##   GG Bp Ap Bm Am   A   B H
## 1  A  0  1  0  1 1.0 0.0 0
## 2  B  1  0  1  0 0.0 1.0 0
## 3 BA  1  0  0  1 0.5 0.5 1
## 4 AB  0  1  1  0 0.5 0.5 1
```

Base de datos de composición genética:

```
BD$secuencia=seq(1,nrow(BD),1)
BD
```

```
##      id sex Finca GG Peso secuencia
## 1  A1A2   2    f1  A  401         1
## 2  A3A4   2    f1  A  401         2
## 3  B1B2   2    f1  B  457         3
## 4  B3B4   2    f1  B  412         4
## 5  A1B2   2    f1 AB  473         5
## 6  A3B4   2    f1 AB  477         6
## 7  B1A2   2    f1 BA  499         7
## 8  B3A4   2    f1 BA  483         8
## 9  A1A4   2    f2  A  440         9
## 10 A3A2   2    f2  A  450        10
## 11 B1B4   2    f2  B  451        11
## 12 B3B2   2    f2  B  446        12
## 13 A1B4   2    f2 AB  403        13
## 14 A3B2   2    f2 AB  407        14
## 15 B1A4   2    f2 BA  401        15
## 16 B3A2   2    f2 BA  402        16
```

```
Base=merge(BD,G, all=TRUE)
library(doBy)
Base=orderBy(~secuencia, data=Base)
rownames(Base)=Base$secuencia
Base$secuencia=NULL
BD$secuencia=NULL
Base
```

```
##    GG   id sex Finca Peso Bp Ap Bm Am   A   B H
## 1   A A1A2   2    f1  401  0  1  0  1 1.0 0.0 0
```

```
## 2   A A3A4   2    f1  401  0  1  0  1 1.0 0.0 0
## 3   B B1B2   2    f1  457  1  0  1  0 0.0 1.0 0
## 4   B B3B4   2    f1  412  1  0  1  0 0.0 1.0 0
## 5  AB A1B2   2    f1  473  0  1  1  0 0.5 0.5 1
## 6  AB A3B4   2    f1  477  0  1  1  0 0.5 0.5 1
## 7  BA B1A2   2    f1  499  1  0  0  1 0.5 0.5 1
## 8  BA B3A4   2    f1  483  1  0  0  1 0.5 0.5 1
## 9   A A1A4   2    f2  440  0  1  0  1 1.0 0.0 0
## 10  A A3A2   2    f2  450  0  1  0  1 1.0 0.0 0
## 11  B B1B4   2    f2  451  1  0  1  0 0.0 1.0 0
## 12  B B3B2   2    f2  446  1  0  1  0 0.0 1.0 0
## 13 AB A1B4   2    f2  403  0  1  1  0 0.5 0.5 1
## 14 AB A3B2   2    f2  407  0  1  1  0 0.5 0.5 1
## 15 BA B1A4   2    f2  401  1  0  0  1 0.5 0.5 1
## 16 BA B3A2   2    f2  402  1  0  0  1 0.5 0.5 1
```

Montaje de matrices:

```
library (MatrixModels)

Finca=as.matrix(model.Matrix(~as.factor(Finca)-1,data=Base))
Finca


##    as.factor(Finca)f1 as.factor(Finca)f2
## 1                   1                  0
## 2                   1                  0
## 3                   1                  0
## 4                   1                  0
## 5                   1                  0
## 6                   1                  0
## 7                   1                  0
## 8                   1                  0
## 9                   0                  1
## 10                  0                  1
## 11                  0                  1
## 12                  0                  1
## 13                  0                  1
## 14                  0                  1
## 15                  0                  1
## 16                  0                  1
```

```
library(stringr)
colnames(Finca)=word(colnames(Finca), 2, sep = fixed(')'))
rownames(Finca)=Base$id
```

```
Finca
```

```
##      f1 f2
## A1A2  1  0
## A3A4  1  0
## B1B2  1  0
## B3B4  1  0
## A1B2  1  0
## A3B4  1  0
## B1A2  1  0
## B3A4  1  0
## A1A4  0  1
## A3A2  0  1
## B1B4  0  1
## B3B2  0  1
## A1B4  0  1
## A3B2  0  1
## B1A4  0  1
## B3A2  0  1
```

```
X=cbind(Finca,Base$A,Base$B,Base$H)
colnames(X)=c("f1","f2","A","B","H")
XpX=t(X)%*%X
XpX
```

```
##    f1 f2 A B H
## f1  8  0 4 4 4
## f2  0  8 4 4 4
## A   4  4 6 2 4
## B   4  4 2 6 4
## H   4  4 4 4 8
```

```
y=as.matrix(Base$Peso)
y
```

```
##       [,1]
##  [1,]  401
##  [2,]  401
##  [3,]  457
##  [4,]  412
##  [5,]  473
##  [6,]  477
```

```
##  [7,]  499
##  [8,]  483
##  [9,]  440
## [10,]  450
## [11,]  451
## [12,]  446
## [13,]  403
## [14,]  407
## [15,]  401
## [16,]  402
```

```
Xpy=t(X)%*%y
Xpy
```

```
##    [,1]
## f1 3603
## f2 3400
## A  3464
## B  3538
## H  3545
```

```
library(MASS)
XpXinv=ginv(XpX)
```

```
Sol=XpXinv%*%Xpy
rownames(Sol)=c("f1","f2","A","B","H")
Sol
```

```
##    [,1]
## f1  229
## f2  203
## A   207
## B   225
## H    11
```

```
Finca1RazaA=Sol[1,]+Sol[3,];Finca1RazaA
```

```
##  f1
## 436
```

```
Finca1RazaB=Sol[1,]+Sol[4,];Finca1RazaB


##  f1
## 454
```

```
Finca2RazaA=Sol[2,]+Sol[3,];Finca2RazaA


##  f2
## 410
```

```
Finca2RazaB=Sol[2,]+Sol[4,];Finca2RazaB


##  f2
## 429
```

```
Finca1Cruzado=Sol[1,]+Sol[3,]*0.5+Sol[4,]*0.5+Sol[5,]
Finca1Cruzado


##  f1
## 456
```

```
Finca2Cruzado=Sol[2,]+Sol[3,]*0.5+Sol[4,]*0.5+Sol[5,]
Finca2Cruzado


##  f2
## 430
```

Si tenemos en cuenta la información de la Finca 2:

```
#Finca f2
X=as.matrix(subset(Base,Base$Finca=="f2",
                   select=c("A","B","H")))
XpX=t(X)%*%X
```

```
y=as.matrix(subset(Base,Base$Finca=="f2",
                   select=c("Peso")))
y
```

```
##    Peso
## 9   440
## 10  450
## 11  451
## 12  446
## 13  403
## 14  407
## 15  401
## 16  402
```

```
Xpy=t(X)%*%y
XpXinv=ginv(XpX)
```

```
Sol=XpXinv%*%Xpy
rownames(Sol)=c("A","B","H")
Sol
```

```
##   Peso
## A  445
## B  448
## H  -44
```

Teniendo en cuenta la interacción de finca y grupo genético:

```
X=data.frame(Finca,Base$A,Base$B,Base$H)
colnames(X)=c("f1","f2","A","B","H")
X$A1=ifelse(X$f1==1,X$A,0)
X$A2=ifelse(X$f2==1,X$A,0)
X$B1=ifelse(X$f1==1,X$B,0)
X$B2=ifelse(X$f2==1,X$B,0)
X$H1=ifelse(X$f1==1,X$H,0)
X$H2=ifelse(X$f2==1,X$H,0)
X=as.matrix(X)
```

```
X
```

```
##       f1 f2   A   B H  A1  A2  B1  B2 H1 H2
## A1A2   1  0 1.0 0.0 0 1.0 0.0 0.0 0.0  0  0
## A3A4   1  0 1.0 0.0 0 1.0 0.0 0.0 0.0  0  0
## B1B2   1  0 0.0 1.0 0 0.0 0.0 1.0 0.0  0  0
## B3B4   1  0 0.0 1.0 0 0.0 0.0 1.0 0.0  0  0
```

```
## A1B2  1  0 0.5 0.5 1 0.5 0.0 0.5 0.0  1  0
## A3B4  1  0 0.5 0.5 1 0.5 0.0 0.5 0.0  1  0
## B1A2  1  0 0.5 0.5 1 0.5 0.0 0.5 0.0  1  0
## B3A4  1  0 0.5 0.5 1 0.5 0.0 0.5 0.0  1  0
## A1A4  0  1 1.0 0.0 0 0.0 1.0 0.0 0.0  0  0
## A3A2  0  1 1.0 0.0 0 0.0 1.0 0.0 0.0  0  0
## B1B4  0  1 0.0 1.0 0 0.0 0.0 0.0 1.0  0  0
## B3B2  0  1 0.0 1.0 0 0.0 0.0 0.0 1.0  0  0
## A1B4  0  1 0.5 0.5 1 0.0 0.5 0.0 0.5  0  1
## A3B2  0  1 0.5 0.5 1 0.0 0.5 0.0 0.5  0  1
## B1A4  0  1 0.5 0.5 1 0.0 0.5 0.0 0.5  0  1
## B3A2  0  1 0.5 0.5 1 0.0 0.5 0.0 0.5  0  1
```

```
XpX=t(X)%*%X
XpX
```

```
##    f1 f2 A B H A1 A2 B1 B2 H1 H2
## f1  8  0 4 4 4  4  0  4  0  4  0
## f2  0  8 4 4 4  0  4  0  4  0  4
## A   4  4 6 2 4  3  3  1  1  2  2
## B   4  4 2 6 4  1  1  3  3  2  2
## H   4  4 4 4 8  2  2  2  2  4  4
## A1  4  0 3 1 2  3  0  1  0  2  0
## A2  0  4 3 1 2  0  3  0  1  0  2
## B1  4  0 1 3 2  1  0  3  0  2  0
## B2  0  4 1 3 2  0  1  0  3  0  2
## H1  4  0 2 2 4  2  0  2  0  4  0
## H2  0  4 2 2 4  0  2  0  2  0  4
```

```
y=Base$Peso
Xpy=t(X)%*%y
Xpy
```

```
##    [,1]
## f1 3603
## f2 3400
## A  3464
## B  3538
## H  3545
## A1 1768
## A2 1696
## B1 1835
## B2 1704
```

```
## H1 1932
## H2 1613
```

```
XpXinv=ginv(XpX)
```

```
Sol=XpXinv%*%Xpy
rownames(Sol)=c("f1","f2","A","B","H",
                "A_f1","A_f2","B_f1","B_f2","H_f1","H_f2")
round(Sol,1)
```

```
##       [,1]
## f1   163.2
## f2   182.6
## A    166.7
## B    179.1
## H      7.3
## A_f1  71.0
## A_f2  95.7
## B_f1  92.2
## B_f2  86.9
## H_f1  58.0
## H_f2 -50.8
```

Desarrollo del EJEMPLO 6.2 con un modelo animal multirracial para peso a los 18 meses de individuos puros de la raza *A* y *B* y sus cruces:

```
Base=data.frame(matrix(ncol=6,byrow=TRUE, c(
  "A1",      NA,NA,1,"A",432,
  "A2",      NA,NA,2,"A",NA,
  "B1",      NA,NA,1,"B",401,
  "B2",      NA,NA,2,"B",411,
  "B3",      NA,NA,1,"B",423,
  "A1A2",   "A1","A2",2,"A",432,
  "B1B2",   "B1","B2",2,"B",423,
  "A1B2",   "A1","B2",2,"AB",421,
  "B3A2",   "B3","A2",1,"BA",432,
  "B3A1B2", "B3","A1B2",2,"BBA",332  )))
colnames(Base)=c("id", "sire", "dam","sex","GG","Peso")
rownames (Base)=seq(1,nrow(Base),1)
```

Información de los primeros seis animales:

```
head(Base)
```

```
##      id sire  dam sex GG Peso
## 1    A1 <NA> <NA>   1  A  432
## 2    A2 <NA> <NA>   2  A <NA>
## 3    B1 <NA> <NA>   1  B  401
## 4    B2 <NA> <NA>   2  B  411
## 5    B3 <NA> <NA>   1  B  423
## 6 A1A2   A1   A2   2  A  432
```

Base de datos de los individuos con información productiva:

```
BD=subset(Base, !is.na(Base$Peso))
BD
```

```
##         id sire  dam sex  GG Peso
## 1       A1 <NA> <NA>   1   A  432
## 3       B1 <NA> <NA>   1   B  401
## 4       B2 <NA> <NA>   2   B  411
## 5       B3 <NA> <NA>   1   B  423
## 6    A1A2   A1   A2   2   A  432
## 7    B1B2   B1   B2   2   B  423
## 8    A1B2   A1   B2   2  AB  421
## 9    B3A2   B3   A2   1  BA  432
## 10 B3A1B2   B3 A1B2   2 BBA  332
```

Información genealógica:

```
library(kinship2)
Genea1=pedigree(id = Base$id, dadid = Base$sire,
                momid = Base$dam, sex=as.numeric(Base$sex))
```

Con la siguiente programación obtendremos el árbol genealógico (FIGURA NRO. 6.2):

```
plot(Genea1, col=as.numeric(as.factor(Base$GG)),
   mar = c(bottom=0.5, left=1, top=4, right=1), cex=0.8)
```

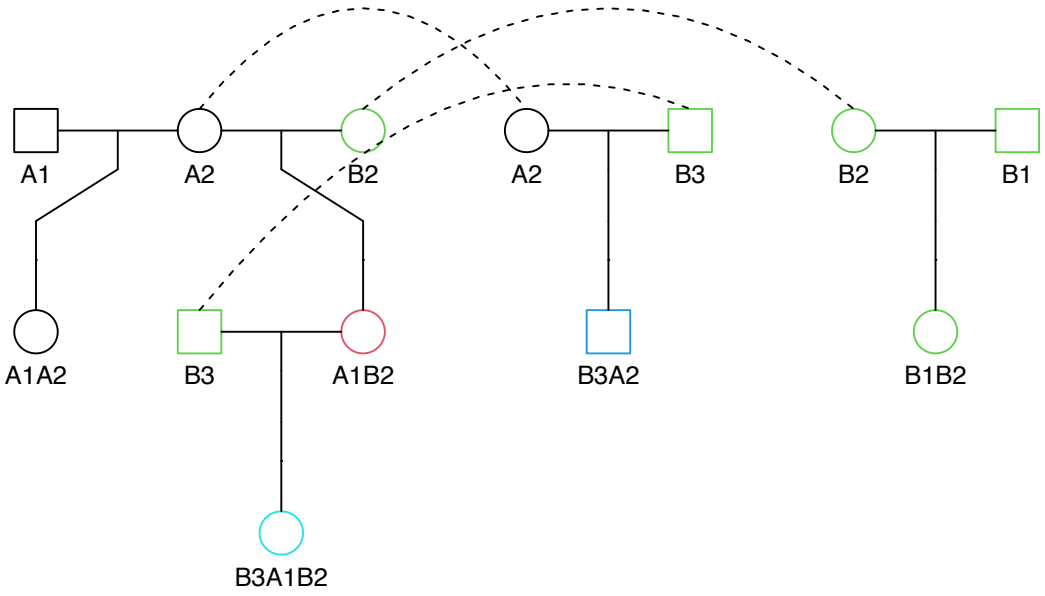

***Figura* 6.2:** Genealogía de los animales puros y cruzados.
Nota: Los círculos identifican a las hembras, los cuadros a los machos y los colores identifican el grupo genético.
Fuente: elaboración propia generada en R-project [25].

```
bitSize(Genea1)


## $bitSize
## [1] 5
##
## $nFounder
## [1] 5
```

```
##
## $nNonFounder
## [1] 5


A=2*kinship(Base$id,Base$sire,Base$dam)
```

```
A[1:10,1:10]#diez primeros animales


##          A1  A2  B1   B2  B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## A1     1.00 0.0 0.0 0.00 0.0 0.50 0.00 0.50 0.00   0.25
## A2     0.00 1.0 0.0 0.00 0.0 0.50 0.00 0.00 0.50   0.00
## B1     0.00 0.0 1.0 0.00 0.0 0.00 0.50 0.00 0.00   0.00
## B2     0.00 0.0 0.0 1.00 0.0 0.00 0.50 0.50 0.00   0.25
## B3     0.00 0.0 0.0 0.00 1.0 0.00 0.00 0.00 0.50   0.50
## A1A2   0.50 0.5 0.0 0.00 0.0 1.00 0.00 0.25 0.25   0.12
## B1B2   0.00 0.0 0.5 0.50 0.0 0.00 1.00 0.25 0.00   0.12
## A1B2   0.50 0.0 0.0 0.50 0.0 0.25 0.25 1.00 0.00   0.50
## B3A2   0.00 0.5 0.0 0.00 0.5 0.25 0.00 0.00 1.00   0.25
## B3A1B2 0.25 0.0 0.0 0.25 0.5 0.12 0.12 0.50 0.25   1.00
```

```
library (MatrixModels)
Pro=as.matrix(model.Matrix(~ as.factor(BD$id) -1))
library(stringr)
colnames(Pro)=word(colnames(Pro), 2, sep = fixed(')'))
rownames(Pro)=BD$id
```

```
Pro


##        A1 A1A2 A1B2 B1 B1B2 B2 B3 B3A1B2 B3A2
## A1      1    0    0  0    0  0  0       0    0
## B1      0    0    0  1    0  0  0       0    0
## B2      0    0    0  0    0  1  0       0    0
## B3      0    0    0  0    0  0  1       0    0
## A1A2    0    1    0  0    0  0  0       0    0
## B1B2    0    0    0  0    1  0  0       0    0
## A1B2    0    0    1  0    0  0  0       0    0
## B3A2    0    0    0  0    0  0  0       0    1
## B3A1B2  0    0    0  0    0  0  0       1    0
```

```
Z=matrix(nrow=nrow(BD),ncol=ncol(A),0)
colnames(Z)=colnames(A)
rownames(Z)=BD$id
```

```
Z[rownames(Pro),colnames(Pro)]=
  Pro[rownames(Pro),colnames(Pro)]
Z
```

```
##         A1 A2 B1 B2 B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## A1       1  0  0  0  0    0    0    0    0       0
## B1       0  0  1  0  0    0    0    0    0       0
## B2       0  0  0  1  0    0    0    0    0       0
## B3       0  0  0  0  1    0    0    0    0       0
## A1A2     0  0  0  0  0    1    0    0    0       0
## B1B2     0  0  0  0  0    0    1    0    0       0
## A1B2     0  0  0  0  0    0    0    1    0       0
## B3A2     0  0  0  0  0    0    0    0    1       0
## B3A1B2   0  0  0  0  0    0    0    0    0       1
```

```
Sexo=as.matrix(model.Matrix(~ as.factor(BD$sex) -1))
colnames(Sexo)=word(colnames(Sexo), 2, sep = fixed(')'))
rownames(Sexo)=BD$id
Sexo
```

```
##         1 2
## A1      1 0
## B1      1 0
## B2      0 1
## B3      1 0
## A1A2    0 1
## B1B2    0 1
## A1B2    0 1
## B3A2    1 0
## B3A1B2  0 1
```

```
X=matrix(nrow=nrow(BD),ncol=ncol(Sexo),0)
colnames(X)=colnames(Sexo)
rownames(X)=BD$id
X[rownames(Sexo),colnames(Sexo)]=
  Sexo[rownames(Sexo),colnames(Sexo)]
X
```

```
##         1 2
## A1      1 0
## B1      1 0
## B2      0 1
## B3      1 0
## A1A2    0 1
## B1B2    0 1
## A1B2    0 1
## B3A2    1 0
## B3A1B2  0 1
```

```
cruzados=subset(BD,BD$GG=="AB"|BD$GG=="BA"|BD$GG=="BBA")
Proc=as.matrix(model.Matrix(~ as.factor(cruzados$id) -1))
```

```
library(stringr)
colnames(Proc)=word(colnames(Proc), 2, sep = fixed(')'))
rownames(Proc)=cruzados$id
Proc
```

```
##         A1B2 B3A1B2 B3A2
## A1B2       1       0    0
## B3A2       0       0    1
## B3A1B2     0       1    0
```

```
n=nrow(Base)
P=matrix(nrow=nrow(BD),ncol=n,0)
rownames(P)=BD$id
colnames(P)=Base$id
```

```
P[cruzados$id,cruzados$id]=Proc[cruzados$id,cruzados$id]
P
```

```
##         A1  A2  B1  B2  B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## A1       0   0   0   0   0    0    0    0    0       0
## B1       0   0   0   0   0    0    0    0    0       0
## B2       0   0   0   0   0    0    0    0    0       0
## B3       0   0   0   0   0    0    0    0    0       0
## A1A2     0   0   0   0   0    0    0    0    0       0
## B1B2     0   0   0   0   0    0    0    0    0       0
## A1B2     0   0   0   0   0    0    0    1    0       0
## B3A2     0   0   0   0   0    0    0    0    1       0
## B3A1B2   0   0   0   0   0    0    0    0    0       1
```

```
Q=matrix(ncol=2,nrow=n)
Q[,1]=ifelse(Base$GG=="A",1,
             ifelse(Base$GG=="B",0,
                    ifelse(Base$GG=="BBA",0.25,0.5)))
Q[,2]=1-Q[,1]
rownames(Q)=Base$id
colnames(Q)=c("A","B")
```

```
Q
```

```
##            A    B
## A1      1.00 0.00
## A2      1.00 0.00
## B1      0.00 1.00
## B2      0.00 1.00
## B3      0.00 1.00
## A1A2    1.00 0.00
## B1B2    0.00 1.00
## A1B2    0.50 0.50
## B3A2    0.50 0.50
## B3A1B2 0.25 0.75
```

```
s=matrix(ncol=1,nrow=n)
s[,1]=ifelse(Base$GG=="A",0,
             ifelse(Base$GG=="B",0,
                    ifelse(Base$GG=="BBA",0.5,  1)))
rownames(s)=Base$id
colnames(s)=c("Heterocigosis")
```

```
s
```

```
##         Heterocigosis
## A1                0.0
## A2                0.0
## B1                0.0
## B2                0.0
## B3                0.0
## A1A2              0.0
## B1B2              0.0
## A1B2              1.0
## B3A2              1.0
## B3A1B2            0.5
```

```
BDc=subset(Base,Base$GG=="AB"|Base$GG=="BA"|Base$GG=="BBA")
BDc
```

```
##         id sire  dam sex  GG Peso
## 8    A1B2   A1   B2   2  AB  421
## 9    B3A2   B3   A2   1  BA  432
## 10 B3A1B2   B3 A1B2   2 BBA  332
```

```
pc=table(BDc$id,BDc$sire)
pc
```

```
##
##           A1 B3
##   A1B2     1  0
##   B3A1B2   0  1
##   B3A2     0  1
```

```
mc=table(BDc$id,BDc$dam)
mc
```

```
##
##           A1B2 A2 B2
##   A1B2       0  0  1
##   B3A1B2     1  0  0
##   B3A2       0  1  0
```

```
T=matrix(ncol=n,nrow=n,0)
colnames(T)=Base$id;rownames(T)=Base$id
T[rownames(pc),colnames(pc)]=pc[rownames(pc),colnames(pc)]
T[rownames(mc),colnames(mc)]=mc[rownames(mc),colnames(mc)]
T=ifelse(T==1,s,0)
T
```

```
##         A1  A2  B1  B2  B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## A1       0   0   0   0 0.0    0    0  0.0    0       0
## A2       0   0   0   0 0.0    0    0  0.0    0       0
## B1       0   0   0   0 0.0    0    0  0.0    0       0
## B2       0   0   0   0 0.0    0    0  0.0    0       0
## B3       0   0   0   0 0.0    0    0  0.0    0       0
## A1A2     0   0   0   0 0.0    0    0  0.0    0       0
## B1B2     0   0   0   0 0.0    0    0  0.0    0       0
```

```
## A1B2    1  0  0  1 0.0    0    0  0.0    0       0
## B3A2    0  1  0  0 1.0    0    0  0.0    0       0
## B3A1B2  0  0  0  0 0.5    0    0  0.5    0       0
```

Vector y:

```
y=matrix(nrow=nrow(BD),ncol=1,0)
colnames(y)=c("Peso")
rownames(y)=BD$id
y[BD$id,]=as.numeric(BD$Peso)
y
```

```
##        Peso
## A1      432
## B1      401
## B2      411
## B3      423
## A1A2    432
## B1B2    423
## A1B2    421
## B3A2    432
## B3A1B2  332
```

Montaje del sistema de ecuaciones:

```
var_e=70
var_h=10
var_a=20
```

```
Gin=solve(A)%x%solve (var_a)
```

```
Rin=as.matrix(1/var_e)
Rin
```

```
##       [,1]
## [1,] 0.014
```

```
I=matrix(nrow=nrow(BD),ncol=nrow(BD),0);diag(I)=1
#I de tamaño igual al número de animales con registro
IRin=I%x%Rin
round(IRin,2)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,] 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##  [2,] 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##  [3,] 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00
##  [4,] 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00
##  [5,] 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00
##  [6,] 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00
##  [7,] 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00
##  [8,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00
##  [9,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01
```

```
Hin=as.matrix(1/var_h)
Hin
```

```
##      [,1]
## [1,]  0.1
```

```
I=matrix(ncol=n,nrow=n,0);diag(I)=1
#I de tamaño igual al número total de animales
IHin=I%x%Hin
round(IHin,3)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]  0.1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0
##  [2,]  0.0  0.1  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0
##  [3,]  0.0  0.0  0.1  0.0  0.0  0.0  0.0  0.0  0.0   0.0
##  [4,]  0.0  0.0  0.0  0.1  0.0  0.0  0.0  0.0  0.0   0.0
##  [5,]  0.0  0.0  0.0  0.0  0.1  0.0  0.0  0.0  0.0   0.0
##  [6,]  0.0  0.0  0.0  0.0  0.0  0.1  0.0  0.0  0.0   0.0
##  [7,]  0.0  0.0  0.0  0.0  0.0  0.0  0.1  0.0  0.0   0.0
##  [8,]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.1  0.0   0.0
##  [9,]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.1   0.0
## [10,]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.1
```

```
XpRinX=t(X)%*%IRin%*%X
round(XpRinX,2)
```

```
##      1    2
## 1 0.06 0.00
## 2 0.00 0.07
```

```
XpRinZQ=t(X)%*%IRin%*%Z%*%Q
round(XpRinZQ,2)
```

```
##      A    B
## 1 0.02 0.04
## 2 0.03 0.05
```

```
XpRinZ=t(X)%*%IRin%*%Z
round(XpRinZ[,1:10],2)#Algunos animales
```

```
##     A1 A2   B1   B2   B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## 1 0.01  0 0.01 0.00 0.01 0.00 0.00 0.00 0.01   0.00
## 2 0.00  0 0.00 0.01 0.00 0.01 0.01 0.01 0.00   0.01
```

```
XpRinPs=t(X)%*%IRin%*%P%*%s
round(XpRinPs,2)#Algunos animales
```

```
##   Heterocigosis
## 1          0.01
## 2          0.02
```

```
XpRinPT=t(X)%*%IRin%*%P%*%T
round(XpRinPT[,1:10],2)#Algunos animales
```

```
##     A1   A2 B1   B2   B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## 1 0.00 0.01  0 0.00 0.01    0    0 0.00    0      0
## 2 0.01 0.00  0 0.01 0.01    0    0 0.01    0      0
```

```
QpZpRinX=t(Q)%*%t(Z)%*%IRin%*%X
round(QpZpRinX,2)
```

```
##      1    2
## A 0.02 0.03
## B 0.04 0.05
```

```
QpZpRinZQ=t(Q)%*%t(Z)%*%IRin%*%Z%*%Q
round(QpZpRinZQ,2)
```

```
##      A    B
## A 0.04 0.01
## B 0.01 0.07
```

```
QpZpRinZ=t(Q)%*%t(Z)%*%IRin%*%Z
round(QpZpRinZ[,1:10],2)#Algunos animales
```

```
##     A1 A2   B1   B2   B3 A1A2 B1B2 A1B2 B3A2 B3A1B2
## A 0.01  0 0.00 0.00 0.00 0.01 0.00 0.01 0.01   0.00
## B 0.00  0 0.01 0.01 0.01 0.00 0.01 0.01 0.01   0.01
```

```
QpZpRinPs=t(Q)%*%t(Z)%*%IRin%*%P%*%s
round(QpZpRinPs,2)
```

```
##   Heterocigosis
## A          0.02
## B          0.02
```

```
QpZpRinPT=t(Q)%*%t(Z)%*%IRin%*%P%*%T
round(QpZpRinPT[1:2,1:8],2)#Algunos animales
```

```
##     A1   A2 B1   B2   B3 A1A2 B1B2 A1B2
## A 0.01 0.01  0 0.01 0.01    0    0 0.00
## B 0.01 0.01  0 0.01 0.01    0    0 0.01
```

```
ZpRinX=t(Z)%*%IRin%*%X
round(ZpRinX,2)
```

```
##           1    2
## A1     0.01 0.00
## A2     0.00 0.00
## B1     0.01 0.00
## B2     0.00 0.01
## B3     0.01 0.00
## A1A2   0.00 0.01
## B1B2   0.00 0.01
## A1B2   0.00 0.01
## B3A2   0.01 0.00
## B3A1B2 0.00 0.01
```

```
ZpRinZQ=t(Z)%*%IRin%*%Z%*%Q
round(ZpRinZQ,2)
```

```
##            A    B
## A1      0.01 0.00
## A2      0.00 0.00
## B1      0.00 0.01
## B2      0.00 0.01
## B3      0.00 0.01
## A1A2    0.01 0.00
## B1B2    0.00 0.01
## A1B2    0.01 0.01
## B3A2    0.01 0.01
## B3A1B2 0.00 0.01
```

```
ZpRinZmasGin=t(Z)%*%IRin%*%Z+Gin
round(ZpRinZmasGin[1:10,1:7],2)#Algunos animales
```

```
##            A1    A2    B1    B2    B3  A1A2  B1B2
## A1       0.11  0.03  0.00  0.03  0.00 -0.05  0.00
## A2       0.03  0.10  0.00  0.00  0.03 -0.05  0.00
## B1       0.00  0.00  0.09  0.03  0.00  0.00 -0.05
## B2       0.03  0.00  0.03  0.11  0.00  0.00 -0.05
## B3       0.00  0.03  0.00  0.00  0.11  0.00  0.00
## A1A2    -0.05 -0.05  0.00  0.00  0.00  0.11  0.00
## B1B2     0.00  0.00 -0.05 -0.05  0.00  0.00  0.11
## A1B2    -0.05  0.00  0.00 -0.05  0.03  0.00  0.00
## B3A2     0.00 -0.05  0.00  0.00 -0.05  0.00  0.00
## B3A1B2   0.00  0.00  0.00  0.00 -0.05  0.00  0.00
```

```
ZpRinPs=t(Z)%*%IRin%*%P%*%s
round(ZpRinPs,3)
```

```
##         Heterocigosis
## A1              0.000
## A2              0.000
## B1              0.000
## B2              0.000
## B3              0.000
## A1A2            0.000
## B1B2            0.000
## A1B2            0.014
```

```
## B3A2            0.014
## B3A1B2          0.007
```

```
ZpRinPT=t(Z)%*%IRin%*%P%*%T
round(ZpRinPT[,3:10],3)#Algunos animales
```

```
##         B1    B2    B3 A1A2 B1B2  A1B2 B3A2 B3A1B2
## A1       0 0.000 0.000    0    0 0.000    0       0
## A2       0 0.000 0.000    0    0 0.000    0       0
## B1       0 0.000 0.000    0    0 0.000    0       0
## B2       0 0.000 0.000    0    0 0.000    0       0
## B3       0 0.000 0.000    0    0 0.000    0       0
## A1A2     0 0.000 0.000    0    0 0.000    0       0
## B1B2     0 0.000 0.000    0    0 0.000    0       0
## A1B2     0 0.014 0.000    0    0 0.000    0       0
## B3A2     0 0.000 0.014    0    0 0.000    0       0
## B3A1B2   0 0.000 0.007    0    0 0.007    0       0
```

```
spPpRinX=t(s)%*%t(P)%*%IRin%*%X
round(spPpRinX,3)
```

```
##                    1     2
## Heterocigosis 0.014 0.021
```

```
spPpRinZQ=t(s)%*%t(P)%*%IRin%*%Z%*%Q
round(spPpRinZQ[1,1:2],3)#Algunos animales
```

```
##     A     B
## 0.016 0.020
```

```
spPpRinZ=t(s)%*%t(P)%*%IRin%*%Z
round(spPpRinZ[,6:10],3)
```

```
##   A1A2   B1B2   A1B2   B3A2 B3A1B2
##  0.000  0.000  0.014  0.014  0.007
```

```
spPpRinPs=t(s)%*%t(P)%*%IRin%*%P%*%s
round(spPpRinPs,3)
```

```
##               Heterocigosis
## Heterocigosis         0.032
```

```
spPpRinPT=t(s)%*%t(P)%*%IRin%*%P%*%T
round(spPpRinPT[1,1:8],3)#Algunos animales


##    A1    A2    B1    B2    B3  A1A2  B1B2  A1B2
## 0.014 0.014 0.000 0.014 0.018 0.000 0.000 0.004
```

```
TpPpRinX=t(T)%*%t(P)%*%IRin%*%X
round(TpPpRinX,3)


##             1     2
## A1      0.000 0.014
## A2      0.014 0.000
## B1      0.000 0.000
## B2      0.000 0.014
## B3      0.014 0.007
## A1A2    0.000 0.000
## B1B2    0.000 0.000
## A1B2    0.000 0.007
## B3A2    0.000 0.000
## B3A1B2 0.000 0.000
```

```
TpPpRinZQ=t(T)%*%t(P)%*%IRin%*%Z%*%Q
round(TpPpRinZQ,3)


##             A     B
## A1      0.007 0.007
## A2      0.007 0.007
## B1      0.000 0.000
## B2      0.007 0.007
## B3      0.009 0.013
## A1A2    0.000 0.000
## B1B2    0.000 0.000
## A1B2    0.002 0.005
## B3A2    0.000 0.000
## B3A1B2 0.000 0.000
```

```
TpPpRinZ=t(T)%*%t(P)%*%IRin%*%Z
round(TpPpRinZ[,6:10],3)#Algunos animales


##        A1A2 B1B2  A1B2  B3A2 B3A1B2
## A1        0    0 0.014 0.000  0.000
```

```
## A2         0    0 0.000 0.014  0.000
## B1         0    0 0.000 0.000  0.000
## B2         0    0 0.014 0.000  0.000
## B3         0    0 0.000 0.014  0.007
## A1A2       0    0 0.000 0.000  0.000
## B1B2       0    0 0.000 0.000  0.000
## A1B2       0    0 0.000 0.000  0.007
## B3A2       0    0 0.000 0.000  0.000
## B3A1B2     0    0 0.000 0.000  0.000
```

```
TpPpRinPs=t(T)%*%t(P)%*%IRin%*%P%*%s
round(TpPpRinPs,3)
```

```
##         Heterocigosis
## A1              0.014
## A2              0.014
## B1              0.000
## B2              0.014
## B3              0.018
## A1A2            0.000
## B1B2            0.000
## A1B2            0.004
## B3A2            0.000
## B3A1B2          0.000
```

```
TpPpRinPTHin=t(T)%*%t(P)%*%IRin%*%P%*%T+IHin
round(TpPpRinPTHin[1:6,1:6],3)#Algunos animales
```

```
##          A1    A2  B1    B2    B3 A1A2
## A1    0.114 0.000 0.0 0.014 0.000  0.0
## A2    0.000 0.114 0.0 0.000 0.014  0.0
## B1    0.000 0.000 0.1 0.000 0.000  0.0
## B2    0.014 0.000 0.0 0.114 0.000  0.0
## B3    0.000 0.014 0.0 0.000 0.118  0.0
## A1A2  0.000 0.000 0.0 0.000 0.000  0.1
```

```
XpRiny=t(X)%*%IRin%*%y
round(XpRiny,3)
```

```
##   Peso
## 1   24
## 2   29
```

```
QpZpRiny=t(Q)%*%t(Z)%*%IRin%*%y
round(QpZpRiny,3)
```

```
##   Peso
## A   20
## B   33
```

```
ZpRiny=t(Z)%*%IRin%*%y
round(ZpRiny,3)
```

```
##        Peso
## A1      6.2
## A2      0.0
## B1      5.7
## B2      5.9
## B3      6.0
## A1A2    6.2
## B1B2    6.0
## A1B2    6.0
## B3A2    6.2
## B3A1B2  4.7
```

```
spPpRiny=t(s)%*%t(P)%*%IRin%*%y
round(spPpRiny,3)
```

```
##               Peso
## Heterocigosis   15
```

```
TpPpRiny=t(T)%*%t(P)%*%IRin%*%y
round(TpPpRiny,2)
```

```
##        Peso
## A1      6.0
## A2      6.2
## B1      0.0
## B2      6.0
## B3      8.5
## A1A2    0.0
## B1B2    0.0
## A1B2    2.4
## B3A2    0.0
## B3A1B2  0.0
```

```
Izq=rbind(
cbind(XpRinX,XpRinZQ,XpRinZ,XpRinPs,XpRinPT),
cbind(QpZpRinX,QpZpRinZQ,QpZpRinZ,QpZpRinPs,QpZpRinPT),
cbind(ZpRinX,ZpRinZQ,ZpRinZmasGin,ZpRinPs,ZpRinPT),
cbind(spPpRinX,spPpRinZQ,spPpRinZ,spPpRinPs,spPpRinPT),
cbind(TpPpRinX,TpPpRinZQ,TpPpRinZ,TpPpRinPs,TpPpRinPTHin))
```

```
Der=rbind(XpRiny,QpZpRiny,ZpRiny,spPpRiny,TpPpRiny)
round(Der,2)
```

```
##                Peso
## 1              24.1
## 2              28.8
## A              19.6
## B              33.3
## A1              6.2
## A2              0.0
## B1              5.7
## B2              5.9
## B3              6.0
## A1A2            6.2
## B1B2            6.0
## A1B2            6.0
## B3A2            6.2
## B3A1B2          4.7
## Heterocigosis 14.6
## A1              6.0
## A2              6.2
## B1              0.0
## B2              6.0
## B3              8.5
## A1A2            0.0
## B1B2            0.0
## A1B2            2.4
## B3A2            0.0
## B3A1B2          0.0
```

```
library(MASS)
Sol=round(ginv(Izq)%*%Der,1)
nombres=c("Sexo 1","Sexo 2","A","B",paste0(Base$id,"adi"),
          "heterosis",paste0(Base$id,"noadi"))
rownames(Sol)=nombres
```

Soluciones:

```
Efecfijo=as.matrix(Sol[1:2,1]);Efecfijo#efectos fijos
```

```
##        [,1]
## Sexo 1  219
## Sexo 2  200
```

```
GrupoGen=as.matrix(Sol[3:4,1]);GrupoGen
```

```
##   [,1]
## A  223
## B  197
```

```
Aditivo=as.matrix(Sol[5:14,1]);Aditivo
```

```
##           [,1]
## A1adi     -2.7
## A2adi      2.7
## B1adi     -1.0
## B2adi      4.2
## B3adi     -3.2
## A1A2adi    1.1
## B1B2adi    4.6
## A1B2adi   -0.8
## B3A2adi    1.3
## B3A1B2adi -9.8
```

```
heterosis=as.matrix(Sol[15,1]);heterosis
```

```
##      [,1]
## [1,] -8.9
```

```
NoAditiv=as.matrix(Sol[16:25,1]);NoAditiv
```

```
##             [,1]
## A1noadi      2.3
## A2noadi      1.6
## B1noadi      0.0
## B2noadi      2.3
```

```
## B3noadi     -2.3
## A1A2noadi    0.0
## B1B2noadi    0.0
## A1B2noadi   -3.9
## B3A2noadi    0.0
## B3A1B2noadi  0.0
```

7 CAPÍTULO SIETE

SELECCIÓN GENÓMICA

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

En este capítulo se presenta la selección genómica, una metodología para la predicción de valores genéticos que emplea genotipos de miles de marcadores moleculares distribuidos a través del genoma de los individuos. En la primera parte se introducen algunos conceptos de uso frecuente en esta área y los desarrollos preliminares que fueron pioneros en el uso de marcadores moleculares en evaluación genética. Luego, se discuten las características generales de la selección genómica para finalmente, presentar algunos de los modelos estadísticos empleados desde la aproximación frecuentista y la bayesiana.

7.1. Conceptos fundamentales

Marcador molecular: región específica del genoma, no necesariamente asociada con variación fenotípica, que se usa para señalar un locus particular.

Marcador molecular polimórfico: aquel que tiene más de un alelo. En el caso bialélico algunos autores imponen una restricción en la definición que corresponde a una frecuencia de al menos 1 % de la variante más rara.

Polimorfismo de nucleótido simple: variante localizada en la posición de un par de bases en el genoma.

Microsatélite: secuencias de ADN en las que un fragmento corto (dos a seis pares base) se repite de manera consecutiva.

Haplotipo: combinación de alelos de diferentes loci que se encuentran en el mismo cromosoma.

Diplotipo: combinación de un par de haplotipos, uno heredado el padre y otro de la madre.

Desequilibrio de ligamiento: asociación no aleatoria entre los genotipos de dos o más regiones del genoma (por ejemplo, genes o marcadores moleculares). Implica la no independencia de las variables aleatorias que representan los genotipos; por lo tanto, el genotipo en una región informa sobre los genotipos en las regiones con las que se encuentra en desequilibrio de ligamiento. Cuando existe independencia de los genotipos de diferentes regiones, se habla de equilibrio de ligamiento.

Término de segregación Mendeliana: también conocido como desvío de segregación Mendeliana, para un individuo i, con padres j y l, corresponde a la diferencia entre el valor genético medio de j y l y el valor genético de i. Proviene del proceso aleatorio que ocurre al generar una muestra de la mitad del material genético de cada parental en los gametos que originan al individuo.

7.2. Primeras aproximaciones al uso de marcadores moleculares en predicción de valores genéticos

En las décadas de los 70 y 80 del siglo XX, hubo adelantos en técnicas de biología molecular que permitieron genotipificar los animales para un conjunto pequeño de marcadores moleculares [41, 42]. El uso de la información derivada del análisis estadístico de estos genotipos dio origen a la denominada selección asistida por marcadores moleculares (*Marker Assisted Selection* – MAS en inglés, descrito por Soller [43]).

Uno de los primeros modelos que incorporaron marcadores moleculares en los modelos lineales mixtos usados para predecir valores genéticos fue propuesto por Fernando y Grossman [44]. En ese estudio se describía un modelo lineal mixto en el que además de los efectos poligénicos modelados mediante el pedigrí, se incorporaban los efectos genéticos aditivos de un marcador molecular separando los efectos del alelo paterno y el materno, los cuales se trataban como efectos aleatorios. También se describieron procedimientos para computar la matriz de covarianzas de estos efectos, así como su inversa y se discutieron extensiones para incluir más de un marcador.

Se tenían muchas expectativas por la incorporación de esta tecnología, por ejemplo, aumento en la confiabilidad de las predicciones de valores genéticos y reducción del intervalo generacional; no obstante, los impactos no fueron los esperados, con pocas excepciones, los resultados no fueron reproducibles [45] y el costo de validar los marcadores asociados a fenotipos de interés era muy alto [46].

7.3. Breve recuento de la llegada y evolución de la selección genómica

En un artículo de suma importancia, Meuwissen et al [47] propusieron el uso de una gran cantidad de datos genómicos para llevar a cabo la predicción de valores genéticos aditivos. Estos datos corresponden a números que dependen del genotipo observado en cada marcador o de diplotipos definidos a partir de haplotipos formados con marcadores adyacentes. Formalmente, son variables aleatorias discretas observables (ver la sección 9), y algunos las denominan codificaciones de los genotipos/haplotipos. Los marcadores están distribuidos a través de todo el genoma por lo que típicamente se tienen miles de estos. La metodología ha recibido varias denominaciones, entre ellas la ya mencionada selección genómica y otras como predicción genómica y predicción a través del genoma y representa un hito importante en la historia del mejoramiento genético [4].

Volviendo al trabajo pionero de Meuwissen et al [47], se propuso un modelo lineal mixto en el que las variables derivadas de los genotipos se consideran regresores con efectos aleatorios y se describieron diferentes aproximaciones para predecir los efectos aditivos de los marcadores: mínimos cuadrados, mejor predictor lineal insesgado y dos aproximaciones bayesianas denominadas bayes A y bayes B. Estos métodos se implementaron con datos simulados ya que para entonces no estaba disponible la tecnología de genotipificación a gran escala.

Años más tarde, VanRaden [48] propuso aproximaciones basadas en modelos lineales mixtos para obtener el MPLI de valores genéticos empleando marcadores moleculares. Allí se propusieron varios modelos, pero en particular, destacó uno en el que el modelo se parametriza en términos del valor genético aditivo de los individuos y se modifica la matriz de covarianzas del vector que los contiene empleando los datos genómicos. Esta aproximación implica varios pasos que se describen más adelante, razón por la cual es conocida como G-BLUP en varios pasos; G-BLUP proviene de su denominación en inglés “Genomic Best Linear Unbiased Predictor”. Llegado este punto, vale la pena mencionar que a aquellos modelos que consideran efectos de cada uno de los marcadores usando variables explicativas como el contenido alélico u otro tipo de variable definida a partir del genotipo del animal, se les denomina regresión a través del genoma o regresión de genoma completo.

Más adelante, con la finalidad de combinar información de pedigrí y la aportada por los marcadores moleculares en la matriz de covarianzas de los valores genéticos aditivos. Aguilar et al [49] desarrollaron un modelo al que llamaron el G-BLUP en un solo paso ssG-BLUP por sus siglas en inglés (*single step* G-BLUP). Esta denominación se dio debido a que no se requieren múltiples fases de análisis (pasos) como en la propuesta por VanRaden [48], sino que se lleva a cabo un solo análisis que toma el pedigrí, los genotipos y los datos fenotípicos para generar los MPLI de los valores genéticos. Una de las principales ventajas de este método es la capacidad de incluir animales sin genotipos que se encuentran en el pedigrí.

Gianola et al [50] acuñaron el término "Alfabeto Bayesiano", para referirse a una familia de modelos bayesianos de regresión a través del genoma, que comparten el mismo modelo de muestreo o verosimilitud (normal multivariado) y que cambian dependiendo de la distribución a priori que se asuma. Los primeros miembros son los ya mencionados bayes A y bayes B, luego se empezó a avanzar con versiones que buscaban de alguna manera mejorar a sus predecesoras o incorporar algunas presunciones sobre los efectos de los marcadores con la esperanza de encontrar mejores métodos de predicción.

7.4. Generalidades de la selección genómica

La selección genómica se basa en el desequilibrio de ligamiento entre los marcadores y los genes que controlan el rasgo o rasgos de interés; debe tenerse claro que los marcadores moleculares no necesariamente corresponden a las regiones que controlan un fenotipo determinado, así pues, estos se emplean como apuntadores o aproximaciones a los genes responsables de la variación fenotípica. En el caso de los SNP, estos pueden estar dentro de un gen o en regiones intergénicas y en un panel de estos marcadores, generalmente se encontrará una porción que no está asociada a un fenotipo dado, a no ser que el panel se haya diseñado exclusivamente para la población de interés con marcadores asociados, es decir, a la medida.

En el planteamiento original de Meuwissen et al [47], la implementación de un programa de selección genómica inicia con un conjunto de animales que cuentan con genotipos y registros fenotípicos. A este conjunto se le denomina población de referencia y se emplea para predecir los efectos de los marcadores moleculares, los cuales permiten construir una ecuación de predicción. Luego, se tiene un grupo de animales que son candidatos a selección y cuentan solamente con genotipos, estos genotipos generan las variables explicativas que se usan en la ecuación de predicción y permiten predecir los valores genéticos de estos individuos, los cuales se seleccionan a partir de dichas predicciones. El proceso se resume en la FIGURA NRO. 7.1.

Ahora bien, debido a que las frecuencias alélicas y los patrones de desequilibrio de ligamiento cambian por efecto de la selección, se requiere recalcular los efectos de los marcadores cada cierto número de generaciones debido a la caída en la confiabilidad, una simulación que ilustra esta situación puede encontrarse en Meuwissen et al [47].

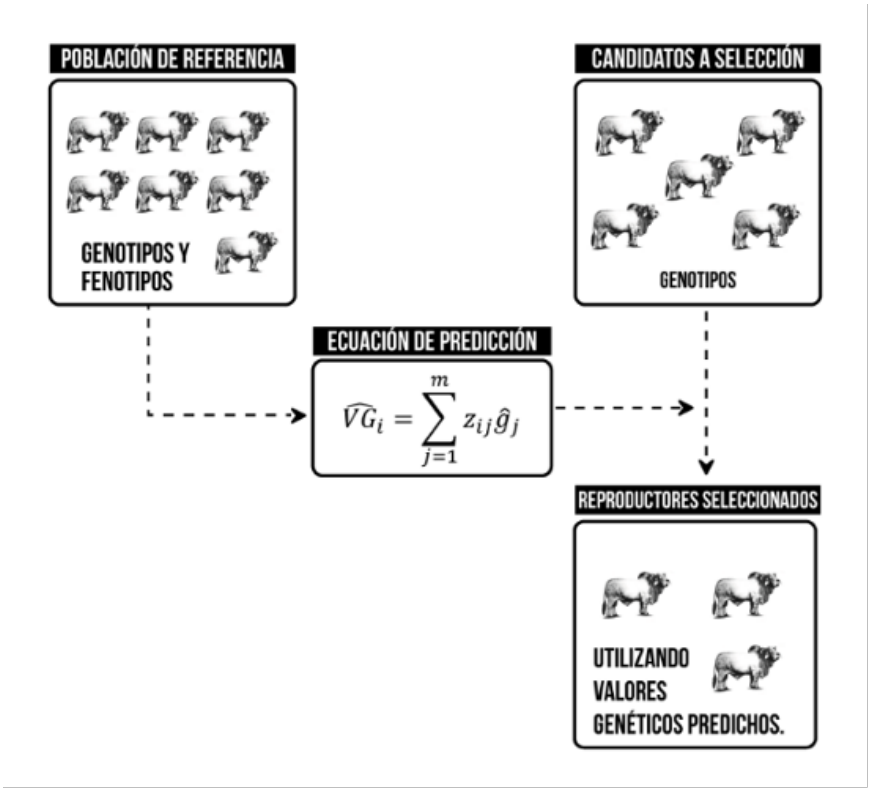

***Figura* 7.1:** Esquema de selección genómica.
Nota: Proceso de evaluaciones genómicas a partir de una población de referencia.
Fuente: elaboración propia (2024).

Son varios los factores que afectan la confiabilidad de la predicción genómica ([51, 52, 53], entre ellos:

1): el número de registros fenotípicos

2): el número de marcadores moleculares

3): heredabilidad de la característica

4): patrones de desequilibrio de ligamiento

5): relaciones de parentesco "capturadas" por los marcadores.

Ciertos trabajos han intentado estudiarlos desde el punto de vista analítico (es decir, usando un enfoque matemático); por ejemplo, Goddard [52] discute el impacto de factores como la heredabilidad y el tamaño de muestra.

7.5. Retos que se enfrentan en la selección genómica

Al considerar miles de marcadores moleculares, se tienen modelos con un gran número de parámetros (los cuales aumentan con el número de marcadores m). Además, en muchos casos el número de registros n es ampliamente superado por el número de marcadores considerados lo cual hace que el número de parámetros del modelo p, sea mayor que n; a esta condición se le llama "la maldición de la dimensión" debido a los retos que trae consigo, y en general, se denota como "p grande n pequeño", "$p > n$" o "$p \gg n$" para enfatizar que en ocasiones la diferencia es muy grande. Además, al tener miles de parámetros en el modelo, a los problemas correspondientes a estimarlos/predecirlos se les denomina "de gran dimensión".

Desde el punto de vista estadístico, la condición $p \gg n$ genera desafíos a nivel de los métodos empleados para inferir los parámetros del modelo, ya que la información contenida en la muestra no es suficiente para estimarlos/predecirlos de manera adecuada; técnicamente, esto se explica porque la verosimilitud no es identificable, lo cual, en términos simples implica que, no es posible aprender algunos parámetros del modelo a partir de la muestra. Este problema se presenta en los métodos frecuentistas y bayesianos, en este último caso, Gianola [54] presenta una formulación técnica del problema (ver Sección 15.1). Otros autores como León-Novelo y Casella [55] discuten el problema de identificabilidad en modelos bayesianos de regresión cuando $p \gg n$. La buena noticia para el caso particular de la selección genómica es que al tener como objetivo predecir el valor genético aditivo total, el problema de identificabilidad no impide que se puedan obtener predicciones confiables, así, estos modelos pueden ser máquinas predictivas útiles.

Otro de los retos que se encuentran tiene que ver con la demanda computacional. Al utilizar modelos mixtos para obtener el MPLI, las ecuaciones a resolver se hacen densas, es decir, muy pocas de las posiciones de la matriz de coeficientes (o del lado izquierdo de las ecuaciones) son nulas, caso contrario al de la evaluación genética basada en el modelo animal. Esto hace que se requiera un mayor número de operaciones para resolver el sistema de ecuaciones y que por ende aumenten los requerimientos de cómputo [56]. Si bien la computación ha avanzado bastante durante los últimos años, para el caso de conjuntos de datos con un alto número de registros productivos y de marcadores, la escalabilidad y la eficiencia computacional siguen siendo aspectos de importancia, que en muchos casos limitan la implementación de algunas metodologías.

Por otro lado, en especies con un corto intervalo generacional y un rápido desarrollo en las cuales las biotecnologías reproductivas como la inseminación artificial están poco diseminadas y el valor de los individuos no es muy elevado; por ejemplo, conejos o cuyes, la selección genómica puede tener un impacto limitado porque estas características no permiten que su uso lleve a cambios apreciables en el progreso genético; sumado a esto, aunque el costo de genotipificación ha disminuido vertiginosamente, en estos sistemas de producción puede llegar a ser prohibitivo.

7.6. Ventajas de la selección genómica

Al permitir llevar a cabo la selección de individuos a muy temprana edad y ejercer una mayor intensidad de selección, una de las grandes ventajas de la selección genómica es el incremento en la ganancia genética por generación y por unidad de tiempo. Sumado a esto, como se mostró en el capítulo 4, bajo el modelo animal unicarácter que considera efectos aditivos directos, el valor genético de un animal es una combinación de tres fuentes de información: el promedio de los valores genéticos de sus padres, los registros del individuo y sus descendientes. Por consiguiente, en el caso de animales que no poseen registros propios ni progenie, lo mejor que se puede hacer con el modelo animal es emplear el promedio de los valores genéticos de los padres y así, los hermanos completos tendrían la misma predicción, entonces se estaría ignorando el término de segregación mendeliana.

La selección genómica permite incorporar información propia del individuo que corresponde a su genotipo y así, tiene en cuenta la segregación mendeliana, esto causa que hermanos completos sin registros propios ni progenie tengan predicciones diferentes. Ahora, tal vez una de las más grandes ventajas de la selección genómica es el aumento en la confiabilidad de las predicciones, especialmente en animales jóvenes, en parte por el hecho de considerar el término de segregación mendeliana. Por ejemplo, Wiggans et al [57] reportaron cambios en la confiabilidad media de animales jóvenes de 30 % a 60 % en vacunos de leche, especie en la que VanRaden et al [58] reportó que la genotipificación de un animal aporta información equivalente a 20 hijas con registros. Además, también se ha reportado que el aumento en la confiabilidad es importante cuando se conocen genes con efectos mayores sobre el fenotipo de interés.

La FIGURA NRO. 7.2 (adaptada de Wiggans et al [57]) muestra la distribución empírica de la confiabilidad de las habilidades de transmisión predichas (la mitad de los valores genéticos predichos) para producción de leche en toros jóvenes de la raza Holstein al incorporar datos genómicos.

Ahora bien, es importante considerar que en el caso de individuos con valores genéticos predichos mediante el modelo animal tradicional y que tienen una alta confiabilidad; por ejemplo, toros de razas lecheras con un alto número de hijas bien distribuidas a través de los grupos contemporáneos, el uso de la selección genómica no ha de generar grandes cambios en la confiabilidad; así, el beneficio está en animales jóvenes o en animales con pocos descendientes y sin registros propios. Por otro lado, esta tecnología también tiene impactos positivos en la confiabilidad de valores genéticos predichos para rasgos cuya medición es costosa y en los que, por consiguiente, hay una proporción alta de animales sin registros fenotípicos, así como en el caso de características con una heredabilidad baja, pero en este último escenario deben considerarse los principios básicos del mejoramiento genético ya que la baja heredabilidad limita la respuesta a la selección y en estos casos se obtienen mayores respuestas manipulando el ambiente en que se desempeñan los animales.

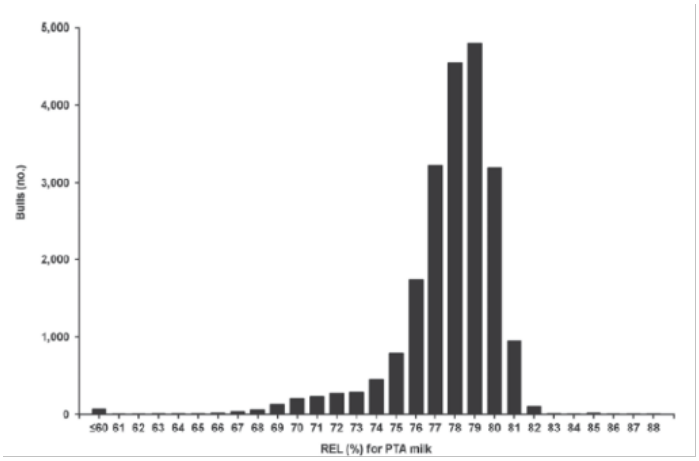

***Figura* 7.2:** Distribución de las confiabilidades de las habilidades de transmisión predichas en toros jóvenes de la raza Holstein para producción de leche en USA. Fuente: tomada de Wiggans et al [57].

7.7. El modelo de regresión a través del genoma

El siguiente es el modelo que contiene las codificaciones de genotipos de marcadores bialélicos como variables regresoras, esta es una versión del modelo de regresión a través del genoma frecuentemente utilizada debido a que los SNP (el tipo de marcadores que contienen los paneles comerciales para diferentes especies) son bialélicos.

$$y = X\beta + Zg + e$$

$$e \sim \mathcal{N}_{\mathrm{n}}(0, R)$$

$$g \sim \mathcal{F}_{\mathrm{g}}(\bullet; \theta)$$

$$e \perp g$$

Donde y,β, e y X son tal y como se definieron en el caso del modelo animal (capítulo 4), $g = (g_1, g_2, \ldots, g_m)$ es el vector aleatorio que contiene los coeficientes de regresión del fenotipo en las codificaciones de los genotipos y que corresponden a los efectos aditivos de cada marcador, Z es la matriz de diseño de g que contiene las codificaciones de los genotipos, la distribución de g varía a través de las diferentes versiones del modelo de regresión a través del genoma, por ejemplo, uno de los supuestos más frecuentes es $g \sim \mathcal{N}_{\mathrm{m}}(0, \sigma_g^2 I_m)$, donde σ_g^2 es la varianza de los efectos aditivos de los marcadores, m el número de marcadores e I_m la matriz de identidad de orden mxm. Por ahora se usa la notación general $\mathcal{F}_{\mathrm{g}}(\bullet; \theta)$ para indicar que se asigna una distribución de probabilidad

a los efectos aditivos de los marcadores que depende del parámetro (posiblemente multidimensional) θ, más adelante presentaremos algunas de las distribuciones que se asumen. La matriz de covarianzas de los errores es por lo general de la forma $R = \sigma^2 I_n$. Finalmente, el símbolo $\perp$ denota independencia probabilística.

Respecto a las codificaciones de los genotipos, las más usuales son el conteo del alelo de referencia, también conocido como dosis alélica, o una versión desplazada o "centrada" que corresponde a la dosis alélica menos uno. Así, al tratarse de individuos diploides, la dosis alélica toma valores de 0, 1 o 2. Será 2 en el caso del genotipo homocigoto para el alelo referencia, 1 para el heterocigoto y 0 para el homocigoto del otro alelo. Formalmente, para el individuo i y el marcador j, con A_j como alelo referencia, tenemos:

$$Z_{ij} = \begin{cases} 2, \text{si el genotipo es } A_jA_j \\ 1, \text{si el genotipo es } A_jB_j \\ 0, \text{si el genotipo es } B_jB_j \end{cases}$$

Para la dosis alélica, mientras que la versión "centrada" sería:

$$Z^*_{ij} = \begin{cases} 1, \text{si el genotipo es } A_jA_j \\ 0, \text{si el genotipo es } A_jB_j \\ -1, \text{si el genotipo es } B_jB_j \end{cases}$$

Ahora bien, bajo este modelo, el valor genético del i-ésimo individuo se obtiene mediante la siguiente ecuación:

$$VG_i = \sum_{j=1}^{p} z_{ij} g_j$$

Correspondiente al producto interno (ver capítulo 15.1) entre el vector de efectos aditivos y el vector que contiene los genotipos del i-ésimo animal, este último no es más que la i-ésima fila de la matriz Z. Nótese que para poder usar esta ecuación se requiere tener predicciones de los efectos de los marcadores, así, una vez se obtiene dicha predicción, denotada por $\hat{g}$, la ecuación de predicción del valor genético aditivo del i-ésimo individuo es:

$$\widehat{VG_i} = \sum_{j=1}^{p} z_{ij} \hat{g}_j$$

Que es la misma que se muestra en la FIGURA NRO. 7.1.

7.8. Modelos estadísticos empleados en selección genómica

En esta sección se presentan algunos de los modelos estadísticos empleados en selección genómica. Se incluyen aquellos que fueron pioneros y dieron lugar al desarrollo de algunas variaciones que buscaban superar una o más de sus deficiencias conceptuales. Cabe aclarar que el desarrollo de modelos y métodos estadísticos para mejorar las predicciones de valores genéticos usando marcadores moleculares es un área de investigación que sigue activa, lo que causa que cada día aparezcan nuevos modelos que incorporan más fuentes de información o presentan alguna modificación de los ya existentes, siempre buscando generar predicciones más confiables. Por lo tanto, se hace hincapié en los primeros métodos, es decir, en los que se encuentran con mayor frecuencia en las aplicaciones y en algunas modificaciones de los mismos; es importante recalcar que la omisión de algunos modelos, no es intencional.

7.8.1. El mejor predictor lineal insesgado genómico G-BLUP

En el artículo *métodos eficientes de cómputo para predicciones genéticas* de VanRaden [48], se encuentran algunos métodos de selección genómica basados en procedimientos, en los que las ecuaciones de modelos mixtos se modificaban para obtener las predicciones genómicas de manera eficiente (computacionalmente hablando). La idea central es modificar la matriz de covarianzas de los valores genéticos al incorporar los genotipos en su cómputo. Este método también se ha denominado G-BLUP en varios pasos, debido a que, como se verá más adelante, implica varias fases de análisis.

Se propusieron varias estrategias para predecir el valor genético incorporando marcadores moleculares, aquí nos enfocamos en aquellas de tipo lineal. Así, consideremos la matriz Z de dimensión nxm definida previamente. Las ecuaciones de modelos mixtos pueden emplear la matriz Z^TZ de dimensión mxm (cuando el modelo se parametriza en términos de efectos de cada marcador) o la matriz ZZ^T de dimensión nxn (cuando se parametriza en términos de los valores genéticos aditivos de cada animal evaluado). Sea p_j la frecuencia del alelo de referencia en el j-ésimo marcador, esta frecuencia debe ser estimada en la población base. Se van a discutir dos modelos equivalentes (misma esperanza y varianza del vector de registros), el primero se parametriza en términos de los valores genéticos aditivos de los individuos y emplea una matriz de parentesco modificada que usa los marcadores moleculares y se denomina "matriz de parentesco genómico". El investigador VanRaden [48] propuso tres formas de construirla, siendo la primera:

$$G_1 = \frac{ZZ^T}{2\sum_{j=1}^{m} p_j(1-p_j)}$$

La división por el término $2\sum_{j=1}^{m} p_j(1-p_j)$ es un escalamiento que busca que G_1 sea análoga a la matriz de parentesco A estudiada en el capítulo 4. La segunda

aproximación pondera los marcadores empleando la precisión, esto es, la inversa de la varianza de la variable aleatoria que cuenta el número de copias del alelo de referencia, entonces:

$$G_2 = ZDZ^T$$

Donde $D = diag\left(\frac{1}{2mp_j(1-p_j)}\right)$. La tercera forma de obtener esta matriz se basa en una regresión lineal multivariante de ZZ^T en la matriz A, de la forma:

$$ZZ^T = g_0 11^T + g_1 A + E$$

Donde 1 es un vector de dimensión nx1 cuyas entradas son todas iguales a 1, E es una matriz de errores de dimensión nxn, g_0 y g_1 son los coeficientes desconocidos que corresponden a intercepto y pendiente, los demás elementos ya fueron definidos. Una vez se obtienen los valores estimados de los coeficientes de la regresión (g_0,g_1), la matriz de parentesco genómico se computa como:

$$G_3 = \frac{ZZ^T - \widehat{g_0} 11^T}{\widehat{g_1}}$$

Cuando $n > m$, G_1 y G_2 son definidas no negativas por construcción; por lo tanto, en vista de los resultados presentados en el capítulo 8 (álgebra matricial), esto no es garantía de que sea invertible. La matriz será singular cuando hay individuos con genotipos iguales en los m marcadores. Por otro lado, en el caso $n < m$ la matriz siempre es singular. Para aliviar problemas de singularidad, cuando A es definida positiva, VanRaden [48] propuso el uso de una combinación lineal de la forma:

$$G_l^* = wG_l + (1-w)A$$

Donde G_l, con $l = 1, 2, 3$ es cualquiera de las matrices genómicas definidas arriba y w es un número real que toma valores entre cero y uno (sin incluir los extremos del intervalo). En vista del Teorema 8.1 y los resultados expuestos en el Capítulo 8 sobre matrices definidas positivas, G_l^* es definida positiva y por consiguiente invertible. Basado en consideraciones teóricas, VanRaden [48] propuso la siguiente forma de obtener el peso w:

$$w = \frac{0.05^2}{0.05^2 + \frac{0.125}{m}}$$

Nótese que cuando el número de marcadores m es muy grande, w toma valores cercanos a 1. Al reemplazar la matriz A en las ecuaciones de modelos mixtos para el modelo animal con efectos genéticos aditivos directos (capítulo 4), por cualquiera de las matrices de relaciones genéticas aditivas que se acaban de discutir y solucionar el sistema, se obtienen directamente las predicciones de valores genéticos de los animales evaluados, conocidos como valores genómicos predichos.

Ahora bien, el modelo lineal mixto propuesto por VanRaden [48] para obtener predicciones de los efectos de cada marcador es como sigue:

$$y = X\beta + Zu + e$$
$$u \sim \mathcal{N}_{\mathrm{m}}(0, \sigma_g^2 I_m)$$
$$e \sim \mathcal{N}_{\mathrm{n}}(0, \sigma_e^2 R)$$
$$u \perp e$$

Donde y es el vector de variables respuesta, que en este caso son valores genéticos deregresados o desviaciones productivas de la progenie, β es el vector de efectos fijos, u el vector de efectos de sustitución alélica de cada marcador, e el vector de errores, X y Z las matrices de diseño asociadas a β y u, respectivamente. La matriz R es diagonal y sus elementos no nulos tienen la forma:

$$R_{ii} = \frac{1}{R_{iP}} - 1$$

Donde R_{iP} es la confiabilidad del valor genético (tradicional) del individuo i basada en las progenies y excluyendo la información parental.

En VanRaden [48] se mencionan tres aproximaciones para obtener predicciones de los valores genéticos aditivos empleando los efectos de cada marcador. La primera se basa en obtener la solución de las ecuaciones de modelos mixtos correspondientes al modelo que se acaba de presentar, lo cual nos da el MPLI de los efectos de sustitución alélica mediante los cuales el predictor del vector de valores genéticos aditivos ($\widehat{a}$) se obtiene como $\widehat{a} = Z\widehat{u}$. Por otro lado, se pueden emplear las ecuaciones resultantes de la teoría de índices de selección, bajo las cuales:

$$\widehat{a}_{SI} = G_1 \left[G_1 + \frac{\sigma_e^2}{\sigma_a^2} R \right]^{-1} (y - X\widehat{\beta})$$

Cuando se emplea el mismo estimador de efectos fijos β, $\widehat{a}_{SI} = \widehat{a}$, esto es, los dos predictores son iguales. Una tercera manera de predecir los valores genéticos, que es equivalente la anterior, se puede obtener cuando G_1 es invertible y tiene la forma:

$$\widehat{a}_G = \left[R^{-1} + \frac{\sigma_e^2}{\sigma_a^2} G_1^{-1} \right]^{-1} R^{-1} (y - X\widehat{\beta})$$

En resumen, la implementación del G-BLUP puede hacerse de varias maneras e implica varios pasos, cuyo procedimiento se resume a continuación: 1) evaluación genética tradicional empleando el modelo animal; 2) extraer pseudo-registros (valores genéticos de-regresados o desvíos de producción de la progenie); 3) ajustar el modelo

mixto descrito en esta sección o cualquiera de los otros métodos para obtener predicciones de valores genómicos y, además, se podría incluir un cuarto paso en el que combinan los valores genómicos con el promedio de los padres y los valores genéticos tradicionales [59, 58]. Esta metodología ha sido empleada en evaluaciones genéticas nacionales de varias razas de bovinos de leche en Estados Unidos, como lo indicaron Wiggans et al [57] y VanRaden (Comunicación personal).

Por otro lado, se han reportado algunas desventajas (como en la gran mayoría de métodos o modelos estadísticos existentes) como la generación de algunos sesgos (las predicciones de animales jóvenes resultan infladas) y la pérdida de información [60], esto debido a que cada paso depende de muchos supuestos y existen varias aproximaciones para ejecutarlos sin tener claridad de cuál es la mejor. La consecuencia de la inflación de los valores genómicos predichos en toros jóvenes no probados es que les confiere una ventaja injusta sobre toros mayores con progenie [49]. Otra fuente de sesgo es el hecho de no tener en cuenta el efecto de la selección en las poblaciones que la experimentan. Además, en monogástricos y ganado de carne es frecuente que las pseudo-observaciones sean de baja calidad, lo cual va en detrimento de la calidad de las predicciones obtenidas [60, 49]. Estas consideraciones abrieron camino para desarrollar el modelo que se presenta a continuación.

7.8.2. El G-BLUP en un solo paso (ssG-BLUP)

La sigla ssG-BLUP proviene del inglés *single step* G-BLUP. Este modelo fue desarrollado por Aguilar et al [49] e involucra el uso de procedimientos de modelos lineales mixtos en los cuales la matriz de relaciones genéticas aditivas basada en el pedigrí se reemplaza por una que combina pedigrí e información genómica y fue desarrollada por Legarra et al [60]. En esta sección, la matriz de relaciones genéticas aditivas computada a partir de datos genotípicos se denota por G. Se considera un modelo para un solo rasgo que contempla solamente efectos genéticos aditivos directos. La matriz propuesta por Legarra et al [60] para combinar información de pedigrí e información genómica se basa en una partición del grupo de animales a ser evaluados en aquellos genotipificados (subíndice 1) y aquellos que no lo están (subíndice 2), así, podemos particionar la matriz de relaciones genéticas aditivas basada en el pedigrí y su inversa como sigue:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$$

De forma análoga, el vector de valores genéticos puede particionarse así:

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

además:

$$Var\left[a_2\right] = G$$

Ahora bien, invocando las propiedades de la distribución normal multivariada [9, 61], la distribución condicional de los valores genéticos de los individuos no genotipificados dados los de aquellos genotipificados es:

$$a_1|a_2 \sim \mathcal{N}_{\mathrm{n}}(A_{12}A_{22}^{-1}a_2, A_{11} - A_{12}A_{22}^{-1}A_{21})$$

Dada esta propiedad, la esperanza del vector de valores genéticos de los individuos no genotipificados puede escribirse explícitamente en función del vector de valores genéticos de aquellos que cuentan con genotipos mediante el modelo de regresión lineal:

$$E\left[a_1|a_2 = \delta\right] = A_{12}A_{22}^{-1}\delta + \epsilon$$
$$\epsilon \sim \mathcal{N}_{\mathrm{n}}(0, A_{11} - A_{12}A_{22}^{-1}A_{21})$$

Ahora, usando $Var\left[a_2\right] = G$, marginalmente tenemos:

$$Var\left[a_1\right] = A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} + A_{11} - A_{12}A_{22}^{-1}A_{21}$$

$$= A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}$$

Además, $Cov\left[a_1, a_2^T\right] = Cov\left[a_{12}A_{22}^{-1}a_2, a_2^T\right] = a_{12}A_{22}^{-1}G$, de aquí, la matriz que combina el pedigrí con los marcadores moleculares es de la forma:

$$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}$$

Por construcción, esta matriz es definida no negativa [60]. Sin embargo, al invertirla, se requiere la inversa de G y, en la práctica, no hay garantía de que esta sea invertible, esto es, no hay garantía de la existencia de G^{-1}. Empleando la propuesta de VanRaden [48] descrita previamente y utilizando la notación de esta sección, un procedimiento ad hoc es usar $wG + (1-w)A_{22}, w \in (0,1)$ en lugar de G. Como se discutió anteriormente, esta estrategia alivia los problemas de singularidad.

Una cualidad de esta matriz es su habilidad para proyectar relaciones genéticas de individuos con genotipos a través del pedigrí, esto permite identificar relaciones genéticas entre animales con información genealógica incompleta y considera esta información a la hora de predecir los valores genéticos de sus descendientes [60]. Por otro lado, la existencia de un algoritmo que permite obtener A_{22} sin necesidad de construir toda la matriz A (Colleau [62] facilitó la obtención de H).

Ahora bien, es posible resolver las ecuaciones de modelos mixtos sin la necesidad de invertir H como se discute en Legarra et al [60], pero si fuese posible encontrar una forma eficiente para invertirla, se podrían obtener predicciones de valores genéticos con

menor esfuerzo computacional. Los autores Aguilar et al [49] y Christensen y Lund [63] derivaron la siguiente expresión para H^{-1} independientemente :

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

La derivación de esta matriz permitió la implementación del ssG-BLUP en bases de datos de gran tamaño. A continuación, se mencionan algunas de las ventajas del método [49, 57]. Simplicidad, puesto que la incorporación de información genómica se realiza mediante la modificación de la matriz de relaciones genéticas; además, en el caso de tener la misma matriz para varios rasgos, la extensión al modelo de múltiples caracteres es simple. Por otro lado, se pueden incluir individuos no genotipificados y se pueden mejorar sus predicciones; además, con este método se generan pesos automáticos a las fuentes de información que componen la predicción del valor genético. Finalmente, si los animales son preseleccionados con base en sus genotipos, los beneficios del método aumentan puesto que se tiene en cuenta el sesgo por selección.

La primera implementación del ssG-BLUP con datos de campo fue la realizada por Aguilar et al [49] en el mismo trabajo en que se introdujo el modelo. Se usó una base de datos grande que contenía 10 466 066 registros de puntaje final provenientes de 6 232 548 vacas Holstein. Se encontraron confiabilidades ligeramente superiores a las obtenidas con el método de varios pasos.

7.8.3. Modelos bayesianos de regresión paramétrica: "Alfabeto Bayesiano"

El denominado "Alfabeto Bayesiano" es una familia de modelos lineales jerárquicos que comparten el mismo modelo de muestreo, normal multivariado, pero difieren en la distribución a priori [50, 54]. La forma general del modelo de regresión que se considera en este capítulo es como sigue:

$$y = 1\mu + Za + e$$

Donde $a = (a_1, a_2, \ldots, a_p)$ y $Z = \{z_{ij}\}_{nxp}$. Se asume que $\sigma_e^2 \sim \mathcal{N}_{\mathrm{n}}(0, \sigma_e^2 I)$ y típicamente a μ se le da una distribución a priori plana. La distribución de a cambia de un modelo a otro y en este capítulo se estudiarán algunos de los miembros de esta familia de modelos.

Nota: en algunas aplicaciones y en algunos de los trabajos citados en esta sección, se consideran fenotipos pre-corregidos, valores genéticos predichos con el modelo animal o desviaciones productivas de la progenie como variable respuesta, razón por la cual se considera solamente un intercepto en el componente no genético. Esta formulación

viene bien para el propósito de discutir las diferentes especificaciones de la distribución a priori de los efectos de los marcadores. Además, la extensión para incluir efectos como por ejemplo los de grupo contemporáneo, es relativamente sencilla.

Es importante considerar que en selección genómica los efectos aditivos se consideran aleatorios, pero en la teoría clásica de genética cuantitativa se tratan como efectos fijos [38]. Como se mencionó previamente, en el caso de considerar efectos de marcadores moleculares, las variables z_{ij} pueden corresponder al número de copias del alelo de referencia (dosis alélica) o a este valor menos uno.

El problema p grande n pequeño ya era conocido por los genetistas cuantitativos de animales, puesto que en sistemas productivos como la ganadería de leche es común que el número de animales a evaluar exceda el número de registros fenotípicos, en parte porque los machos jamás tendrán observaciones para fenotipos como intervalo entre partos o características de la leche (producción y composición) y porque ciertos rasgos toman mucho tiempo en ser observados y en consecuencia los animales jóvenes no tienen registros. Diferentes aproximaciones han sido empleadas para lidiar con esta condición; por ejemplo, técnicas que buscan seleccionar subconjuntos de variables que tengan un desempeño admisible o reducir la dimensión al emplear variables explicativas sintéticas derivadas del conjunto original como la regresión en componentes principales. En particular, se emplean los modelos bayesianos porque el uso de una distribución a priori permite llevar a cabo inferencias cuando $p > n$ [55].

Una solución para el problema de alta dimensión es imponer restricciones sobre la magnitud de los estimadores, a estos se les conoce como estimadores de encogimiento o acortamiento. Existen diferentes formas de llevar a cabo dicho encogimiento, se puede hacer hacia una media, superficies de regresión, o una constante dada que por lo general es cero [64]. El encogimiento hacia cero se basa en la idea de que muchas de las variables regresoras no tienen efecto, lo cual equivale a que el respectivo coeficiente de regresión sea cero. En el contexto de la selección genómica, esto quiere decir que se asume que muchos de los marcadores no tienen un efecto sobre el rasgo de interés. En los casos en que muchos de los parámetros estimados son cero, se dice que se tiene un estimador ralo.

En el caso particular de modelos de regresión lineal bayesianos, como los que se estudian en esta sección, el encogimiento se genera de manera automática y varía según la distribución a priori [54]. Así, los miembros del alfabeto bayesiano ejercen diferentes formas de encogimiento. Más adelante se discute el encogimiento bajo algunos miembros del alfabeto, pero por ahora, es relevante mencionar que el G-BLUP puede verse como una regresión de cresta ("ridge" en inglés) bayesiana y en este caso, cuando se trabaja con marcadores bialélicos como los SNP, el grado de encogimiento hacia cero depende de $2p_j(1 - p_j)$, donde p_j es la frecuencia del alelo de referencia del j-ésimo marcador [54]. Ahora bien, la función $p_j(1 - p_j)$ se maximiza cuando $p_j = 0.5$, así, a medida que las frecuencias de los dos alelos son más similares el grado de encogimiento es mayor.

En un modelo de regresión lineal bayesiano, cuando $p > n$, la verosimilitud (ver sección 15.1) no es identificable y por consiguiente algunos parámetros del modelo no

son estimables, además, la distribución a priori tiende a ser muy influyente [55, 54], es decir, la distribución posterior es afectada en gran medida por la a priori relegando el peso de los datos, una característica no deseable en un análisis bayesiano. En Leon y Casella [55] se estudió el comportamiento de un modelo de regresión bayesiana con una a priori normal multivariada para los coeficientes de regresión, se enfocaron en el caso en que el número de parámetros tiende a infinito, encontrando condiciones bajo las cuales la distribución a priori no domina la inferencia. Estos autores concluyeron que, a no ser que las varianzas a priori de los coeficientes de regresión disminuyan a una tasa $\frac{1}{p}$, en el límite, su distribución posterior es la misma que la a priori, lo cual claramente es una propiedad indeseable puesto que no hay grado alguno de aprendizaje.

Ahora es importante considerar una consecuencia de la no identificabilidad de la verosimilitud. Si bien este texto se enfoca en la predicción de valores genéticos y este capítulo en particular discute el uso de datos genómicos en dicho proceso, vale la pena que el lector esté al tanto de la siguiente situación que es consecuencia de la condición "p grande n pequeño" y repercute en los denominados estudios de asociación genómica a través del genoma (GWAS). La no identificabilidad tiene como consecuencia que los efectos individuales de algunos marcadores no se pueden estimar correctamente, sino que contienen información sobre los efectos de otros marcadores. Así, este fenómeno debe considerarse en el caso del GWAS puesto que claramente puede conducir a inferencias inapropiadas que a su vez llevan a conclusiones erróneas.

En el ámbito bayesiano, la definición de identificabilidad no es la misma que se presenta en el complemento 9, de hecho, no existe una única definición. En la sección 15.1 se presentan algunos detalles de la no identificabilidad de algunos parámetros en el caso $p > n$ en el ámbito bayesiano siguiendo el procedimiento mostrado por Gianola [54].

A continuación, se discuten brevemente algunos de los modelos de regresión lineal bayesianos empleados en predicción genómica. Los que aparecieron en el trabajo pionero de Meuwissen et al [47] son bayes A y bayes B, siendo el primero un caso particular del segundo. Otros modelos como bayes $C-\pi$ y $D-\pi$ [65] corresponden a extensiones de bayes A y B que intentan superar algunas de sus limitantes conceptuales.

Bayes A: este es un modelo bayesiano jerárquico con la siguiente distribución a priori:

$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e^2)$$

$$a_j|\sigma_{a_j}^2 \overset{\text{iid}}{\sim} \mathcal{N}\left(0, Diag(\sigma_{a_j}^2)\right)$$

$$\sigma_{a_j}^2|\upsilon, S^2 \overset{\text{iid}}{\sim} \chi^{-2}(\upsilon, S^2)$$

$$\forall j = 1, 2, \ldots, p$$

Donde $\chi^{-2}(a, b)$ representa la distribución chi-cuadrada escalada invertida con a grados de libertad y parámetro de escala b. La distribución marginal de los efectos de los marcadores es la misma y corresponde a una t de student escalada con υ grados de

libertad y parámetro de escala S^2, la varianza de esta distribución es $(S^2)^2 \frac{\upsilon}{\upsilon-2}$ [61]. Por lo tanto, a priori, los marcadores tienen la misma varianza marginal. Bajo este modelo, el nivel de encogimiento es mayor para aquellos marcadores con efecto cercano a cero que para aquellos con efectos grandes. Además, este también aumenta a medida que los grados de libertad υ crecen [54].

Como se dijo antes, siempre que se tenga $p > n$ la distribución a priori es muy influyente en la posterior, reduciendo así el peso de los datos. En particular, al asignar varianzas condicionales diferentes a cada marcador, bayes A presenta el inconveniente de que el nivel de encogimiento es altamente dependiente de S^2 [65], una característica que muestra la influencia de la distribución a priori en este modelo particular. Además, el nivel de aprendizaje bayesiano sobre las varianzas de cada marcador es pobre tal como fue reportado por Gianola et al [50]. Estos autores llegaron a tal conclusión tras observar que la distribución condicional completa de σ^2_{aj} tiene solamente un grado de libertad más que la a priori, sin importar cual sea el tamaño de muestra o el número de marcadores. Por otro lado, el coeficiente de variación posterior de estos parámetros no tiende a cero a medida que el número de datos aumenta y la ganancia relativa de información (el cambio de entropía una vez se observan los datos) es negligible. Finalmente, también encontraron que la divergencia de Kullback-Leibler entre la condicional completa de σ^2_{aj} y su a priori es cercana a cero, esto quiere decir que las dos distribuciones son muy similares. Se empleó la distribución condicional completa porque la marginal posterior no tiene forma cerrada (es decir, no se conoce su forma matemática). En conclusión, formular una varianza a priori condicional distinta para cada marcador no resulta útil porque no hay aprendizaje sobre estos parámetros y el modelo también presenta falencias al inferir los efectos de los marcadores [54].

Bayes B: Meuwissen et al [47] formularon este modelo con el ánimo de tener en cuenta el hecho de que varios loci podrían ser no segregantes. Bajo el modelo clásico de genética cuantitativa, en el cual los efectos a_j son fijos, los loci que no segregan no presentan varianza genética puesto que esta proviene de la variación en los genotipos [38, 50]. La idea general de este modelo es que existe una proporción de marcadores que no tienen efecto sobre le fenotipo de interés. La distribución a priori se especifica así:

$$a_j|\sigma^2_{a_j} \overset{\text{indep.}}{\sim} \begin{cases} \text{Degenerada en 0, si } \sigma^2_{a_j} = 0 \\ \mathcal{N}(0, \sigma^2_{a_j}), \text{ si } \sigma^2_{a_j} > 0 \end{cases}$$

$$\sigma^2_{a_j}|\pi \overset{\text{indep.}}{\sim} \begin{cases} \text{Degenerada en 0, con probabilidad } \pi \\ \chi^{-2}(\upsilon, S^2) \text{ con probabilidad } 1-\pi \end{cases}$$

$$\forall j = 1, 2, \ldots, p$$

Así, la distribución conjunta de σ^2_{aj} y a_j es:

$$a_j, \sigma^2_{a_j}|\pi \overset{\text{indep.}}{\sim} \begin{cases} \text{Degenerada en 0, con probabilidad } \pi \\ \mathcal{N}(0, \sigma^2_{a_j}) \text{x} \chi^{-2}(\upsilon, S^2), \text{ con probabilidad } 1-\pi \end{cases}$$
$$\forall j = 1, 2, \ldots, p$$

Nótese que si $\pi = 0$, bayes B es equivalente a bayes A. Similar a lo ocurrido con bayes A, tras integrar $f(a_j, \sigma^2_{a_j}|\pi)$ con respecto a $\sigma^2_{a_j}$ se obtiene la siguiente distribución marginal de a_j:

$$a_j|S^2, \upsilon, \pi \overset{\text{iid}}{\sim} \begin{cases} \text{Degenerada en 0, con probabilidad} \pi \\ \text{t}(0, S^2, \upsilon) \text{ con probabilidad } 1-\pi \end{cases}$$
$$\forall j = 1, 2, \ldots, p$$

Donde $\text{t}(0, S^2, \upsilon)$ representa una distribución t centrada y escalada con parámetro de escala S^2 y υ grados de libertad. Así podemos notar que la distribución a priori marginal asignada a los efectos aditivos de los marcadores es una mixtura de una distribución degenerada en 0 (punto de masa en 0) y una t centrada y escalada (con los parámetros antes descritos) con probabilidades de mezcla π y $1-\pi$. La esperanza de esta distribución es cero y la varianza es:

$$Var\left[a_j|S^2, \upsilon, \pi\right] = (1-\pi)(S^2)^2 \frac{\upsilon}{\upsilon - 2}$$

Por ende, al igual que en bayes A, se está signando la misma distribución a priori marginal a los efectos de los marcadores. Las mismas falencias que se mencionaron para bayes A, también aplican para el modelo bayes B. Por otro lado, Gianola et al [50] señalaron una falencia conceptual en la formulación de bayes B. Bajo un marco de trabajo bayesiano, el hecho de que la varianza del efecto de un marcador sea cero indica que su efecto se conoce con total certeza, mas esto no significa que el efecto sea nulo. Por lo tanto, se recomienda especificar bayes B como una mixtura que incluya punto de masa en 0, pero a nivel de los efectos de los marcadores, no de sus varianzas.

Bayes C y bayes D: los autores Habier et al [65] propusieron dos modelos jerárquicos que buscaban controlar la alta influencia de la distribución a priori en bayes A y B. Sin embargo, antes de presentar los modelos, vale la pena resaltar que tal y como se presenta en el capítulo 15.2, cuando el número de marcadores es mayor al número de registros fenotípicos para la característica evaluada, es inevitable que la distribución a priori tenga una alta influencia sobre la posterior.

Uno de estos modelos fue denominado bayes $C-\pi$, este asigna la misma varianza a priori condicional a los efectos de los marcadores y trata la probabilidad de mezcla como uno de los parámetros desconocidos del modelo. La distribución a priori es:

$$a_j|\pi, \sigma^2_g \overset{\text{iid}}{\sim} \begin{cases} \text{Degenerada en 0, con probabilidad } \pi \\ \mathcal{N}(0, \sigma^2_g), \text{ con probabilidad } 1-\pi \end{cases}$$

$$\forall j = 1, 2, \ldots, p$$
$$\sigma_g^2 | \upsilon, S^2 \sim \chi^{-2}(\upsilon, S^2)$$
$$\pi \sim \text{Uniforme}(0, 1)$$

El otro modelo fue denominado bayes $D - \pi$, bajo este se consideran varianzas a priori condicionales diferentes para el efecto de cada marcador, pero se asigna una distribución Gamma$(1, 1)$ o equivalentemente $Exponencial(1)$ al parámetro S^2. La distribución de la varianza del error es la misma que en el caso de bayes A y B.

Empleando datos simulados, Habier et al [65] consideraron variables como el costo computacional y la incertidumbre sobre el número de QTL que afectan el fenotipo estudiado y concluyeron que bayes $C - \pi$ es un modelo atractivo para evaluaciones de rutina. De hecho, este es uno de los modelos bayesianos más empleados en la predicción de valores de cría genómicos en animales y plantas. Finalmente, vale la pena mencionar que algunos autores denominan bayes C y bayes D a las versiones de estos modelos que tratan la probabilidad de mezcla como un parámetro conocido.

Bayes L: en 1996, Tibshirani [66] desarrolló el LASSO, denominado así por sus siglas en inglés de *Least Abolsute Shrinkage and Selection Operator*, operador minimal absoluto de encogimiento y selección, como un método para estimar los coeficientes de un modelo de regresión lineal múltiple ejerciendo tanto encogimiento hacia 0 como selección de variables, esto último quiere decir que los valores estimados de los coeficientes de regresión de algunas variables son cero, por ende, las variables seleccionadas son aquellas cuyos coeficientes no son nulos. El término "absoluto" viene del hecho de que se añade una penalización tipo L_1 (la cual involucra la suma de los valores absolutos de los coeficientes) a la función objetivo de los mínimos cuadrados ordinarios. Por lo tanto, para un modelo lineal con componente sistemático $E\left[y\right] = X\beta$ el estimador LASSO corresponde a:

$$\widehat{\beta}_{LASSO} = \underset{\beta}{\text{argmin}} \left[(y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

Donde $\beta = \{\beta_j\}_{px1}$, $\lambda > 0$ es un parámetro de sintonización, es decir, aquel que no se estima directamente por el método, sino que debe elegirse bajo algún otro criterio. Mayores detalles sobre el proceso de sintonización van más allá del ámbito de este texto, estos son ampliamente utilizados en áreas como el aprendizaje de máquina y aprendizaje estadístico, el lector interesado puede referirse a los textos de Bishop [67] y Hastie et al [68].

Para Tibshirani [66], el estimador LASSO puede interpretarse como la moda de la distribución posterior de los coeficientes de un modelo de regresión bajo distribuciones a priori Laplace (o doble exponencial) independientes, siendo así un estimador tipo máximo a posteriori. Siguiendo esta idea, Park y Casella [69] propusieron un tratamiento completamente bayesiano del problema, las densidades a priori son:

$$f(u|\sigma_e^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt[2]{\sigma_e^2}} \exp^{\frac{-\lambda|u_j|}{\sqrt[2]{\sigma_e^2}}}$$

$$f(\sigma_e^2) = \frac{1}{\sigma_e^2}$$

La distribución impropia $\frac{1}{\sigma_e^2}$ puede ser reemplazada por una Gamma inversa para tener una a priori conjugada; además, Park y Casella [69] demostraron que esta formulación condicionada en σ_e^2 es necesaria para tener una posterior unimodal. Para simplificar la inferencia bajo este modelo, los autores usaron la técnica de aumento de datos y propusieron una distribución a priori equivalente:

$$u|\sigma_e^2, \tau^2 \sim \mathcal{N}_{\mathrm{p}}(0, \sigma_e^2 D_\tau)$$

$$\tau_1^2, \ldots, \tau_p^2 \overset{\text{iid}}{\sim} \exp\left(\frac{\lambda^2}{2}\right)$$

$$\sigma_e^2 \sim \mathrm{F}(\sigma_e^2)$$

Donde $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)$, $D_\tau = diag(\tau_1^2, \ldots, \tau_p^2)$. La equivalencia viene dada porque al integrar la densidad conjunta de u y τ^2 dado σ_e^2 con respecto a τ^2, se obtiene el mismo modelo presentado originalmente.

Una característica del LASSO bayesiano es que no genera ceros exactos como su contraparte frecuentista, así, lo que muchos hacen es imponer un umbral de manera que efectos de marcadores cuyo valor absoluto es inferior a dicho umbral se declaran nulos. Finalmente, este modelo ejerce un encogimiento hacia cero más fuerte que la regresión de cresta debido a que hay mayor densidad alrededor de este valor [54].

Existen otros miembros del alfabeto como por ejemplo bayes R [70], el cual emplea una mixtura de distribuciones normales no centradas (su media no es cero). Por otro lado, en una serie de tres trabajos de Martínez et al [71, 72, 73] introdujeron dos clases de modelos gráficos empleados para estimar matrices de covarianza cuando $p > n$, estos se denominan modelos Gaussianos de grafos de covarianza (MGGCov) y modelos Gaussianos de grafos de concentración (MGGCon). Los primeros estiman directamente la matriz de covarianzas mientras que los segundos estiman su inversa, es decir, la matriz de concentración (de allí su nombre) o de precisión. Estos modelos heredan su nombre del hecho de que el patrón de ceros de la matriz de covarianzas (MGGCov) o la matriz de concentración (MGGCon) se codifica empleando un grafo.

Brevemente, un grafo G es una colección de dos conjuntos, un conjunto de nodos o vértices V y un conjunto de aristas A, los elementos de A son pares que indican los vértices (elementos de V) que están unidos mediante una arista. La FIGURA NRO. 7.3 muestra el grafo con conjunto de vértices $V = (1, 2, 3, 4)$ y conjunto de aristas $A = ((1, 2), (1, 3), (2, 4), (3, 4))$. Este tipo de grafos se denominan no dirigidos porque las aristas no tienen dirección, a diferencia de aquellos en los que la dirección se indica mediante una cabeza de flecha y que se conocen como grafos dirigidos. Por ejemplo, un pedigrí es un grafo dirigido, las flechas siempre van de padres a hijos.

En el caso de MGGCov y MGGCov, el grafo codifica el patrón de ceros de la matriz correspondiente así: los vértices o nodos representan las variables aleatorias (efectos de marcadores en el caso de selección genómica), aquellos que comparten

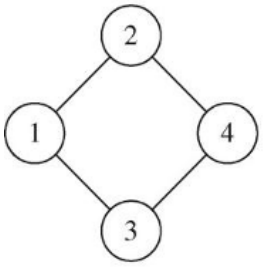

***Figura* 7.3:** Ejemplo de un grafo no dirigido $G = (V, A)$, con $V = (1, 2, 3, 4)$, y $A = ((1, 2), (1, 3), (2, 4), (3, 4))$.
Fuente: elaboración propia (2024).

una arista están correlacionados marginalmente (MGGCov) o condicionalmente dadas todas las otras variables aleatorias consideradas (MGGCon), mientras que aquellos que no comparten una arista tienen covarianza nula y la matriz respectiva (de covarianzas o de concentración) tiene ceros en las posiciones correspondientes. Así, si el grafo en la FIGURA NRO. 7.3 se emplea para codificar el patrón de ceros de la matriz de covarianzas, los siguientes pares de variables aleatorias tienen covarianza nula y de allí, correlación nula (complemento 9): 1 y 4, 2 y 3; la matriz correspondiente es de la forma:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & \sigma_{24} \\ \sigma_{13} & 0 & \sigma_3^2 & \sigma_{43} \\ 0 & \sigma_{24} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

Por su parte, cuando el grafo codifica las entradas nulas de la matriz de concentración o de precisión, debe recordarse que estas corresponden a varianzas y covarianzas condicionales, dadas todas las demás variables. En el ejemplo, dadas 2 y 3, 1 y 4 tienen covarianza condicional nula y de allí correlación parcial nula, dadas 1 y 4, 2 y 3 tienen covarianza condicional nula y de allí correlación parcial nula. En este caso la matriz de concentración o de precisión es de la forma:

$$\Omega = \begin{bmatrix} \omega_1^2 & \omega_{12} & \omega_{13} & 0 \\ \omega_{12} & \omega_2^2 & 0 & \omega_{24} \\ \omega_{13} & 0 & \omega_3^2 & \omega_{43} \\ 0 & \omega_{24} & \omega_{43} & \omega_4^2 \end{bmatrix}$$

Un cero en Ω no siempre corresponde a un cero en Σ, excepto cuando el grafo tiene ciertas propiedades. Mayores detalles pueden consultarse en Lauritzen [74] y Martínez et al [71, 72, 73]. En particular, estos modelos permiten tener en cuenta la correlación existente entre efectos de marcadores (recordar que bajo normalidad, no correlación e independencia son equivalentes), localizados incluso en diferentes cromosomas (puesto que la correlación no solo se debe a cercanía espacial) en un modelo de regresión a través del genoma. Se propusieron tres modelos: bayes G [72], que permite inferir la matriz de concentración de los efectos aditivos de los marcadores cuando el grafo que la gobierna se define de antemano empleando principios de genética, bayes G-Cov [71], que permite estimar la matriz de covarianza bajo el mismo escenario y bayes G-Sel [73] que permite estimar la matriz de concentración sin necesidad de especificar previamente el grafo, en este caso se habla de un problema de selección de modelo gráfico.

Finalmente, cabe mencionar que, las extensiones de algunos miembros del alfabeto bayesiano al caso de respuestas no Gaussianas mediante el uso de modelos lineales generalizados bayesianos; por ejemplo, bayes TA, TB y TC-π [75] corresponden a las versiones de bayes A, B y C-π para el modelo umbral, un modelo utilizado en la evaluación de fenotipos medidos en una escala ordinal como la facilidad de parto. El lector interesado en una introducción los modelos umbral puede referirse a Lynch y Walsh [76] y Mrode [21]. También presentamos en el capítulo 15.2.2 ejercicios que le permitirán tener una mejor comprensión de la predicción genómica.

SEGUNDA PARTE

CAPÍTULOS COMPLEMENTARIOS

8 CAPÍTULO OCHO

Algunos conceptos de álgebra matricial

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

El lenguaje para describir los métodos y modelos utilizados en evaluación genética implica una gran cantidad de elementos de álgebra matricial. Por lo tanto, resulta relevante hacer un breve repaso de algunos de estos conceptos y de la notación requerida en este campo. Así, esta sección pretende brindar una fuente de consulta muy puntual en la que el lector pueda revisar conceptos que se encontrarán a lo largo del texto, para quienes deseen profundizar más en este tema se recomiendan las obras de Harville [77] y Searle [78].

Una matriz puede definirse como un arreglo de escalares en filas y columnas. Cuando hay una sola fila o columna se habla de un vector. En este texto solo consideraremos las matrices reales, es decir, matrices cuyos elementos son números reales. Normalmente se denotan con letras mayúsculas, así, al escribir A_{nxm} nos referimos a una matriz con n filas y m columnas. Las entradas de una matriz se identifican de manera única por su posición (fila-columna), así, a_{ij} es la entrada de A ubicada en la fila i, columna j.

8.0.1. Operaciones básicas

Suma: se suma elemento a elemento; por lo tanto $A + B$ está definida si A y B tienen las mismas dimensiones y además, la operación es conmutativa $A + B = B + A$. Entonces, si:

$$A = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 6 & 2 \\ 3 & 2 & 4 \end{bmatrix}$$

$$B = \begin{bmatrix} 3 & 0 & 4 \\ 1 & 7 & 6 \\ 4 & 7 & 10 \end{bmatrix}$$

$$A + B = \begin{bmatrix} 7 & 1 & 7 \\ 2 & 13 & 8 \\ 7 & 9 & 14 \end{bmatrix} = B + A$$

Multiplicación por escalar: se multiplica el escalar por cada elemento de la matriz.

Si el escalar $c = 2$ lo multiplicamos con la matriz A, se tendría:

$$c * A = 2 * \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} c * a_{11} & c * a_{12} & c * a_{13} \\ c * a_{21} & c * a_{22} & c * a_{23} \\ c * a_{31} & c * a_{32} & c * a_{33} \end{bmatrix} = \begin{bmatrix} 8 & 2 & 6 \\ 2 & 12 & 4 \\ 6 & 4 & 8 \end{bmatrix}$$

Multiplicación matricial: a diferencia de la multiplicación de números reales, la multiplicación de matrices no es conmutativa, por lo tanto, si tenemos las matrices A y D y escribimos AD, se dice que A premultiplica a D, o que D postmultiplica a A. Ahora bien, el producto AD está definido cuando el número de columnas de A es igual al número de filas de D, en este caso, se dice que las matrices son conformables para el producto AD. Nótese que el producto AD puede estar definido, pero el producto DA no, a manera de ejemplo:

Sean A_{nxm} y D_{mxp}, $m \neq n \neq p$, entonces el producto AD está definido, pero DA no lo está. Aun si, AD y DA están definidos, esto no implica $AD = DA$.

$$A_{nxm} D_{mxp} = E_{nxp}$$

$$A_{nxm} = \{a_{ij}\}, D_{mxp} = \{d_{ij}\}, E_{nxp} = \{e_{ij}\}$$

$$e_{ij} = \sum_{k=1}^{m} a_{ik} d_{kj}$$

Suponga que:

$$D = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 10 \end{bmatrix}$$

Entonces, el producto AD se puede computar porque A tiene tres columnas y D tiene tres filas, además, como A tiene tres filas y D tiene dos columnas, la dimensión de la matriz resultante E será 3×2. Entonces, si definimos como $E = \{e_{ij}\}$ a la matriz resultante del producto AD, es decir, $E = AD$, sus entradas se obtienen así:

La multiplicación de A y D sería:

$$E = A * D = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} * \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix}$$

$$e_{11} = 4 * 1 + 1 * 2 + 3 * 3 = 15$$

$$e_{12} = 4 * 0 + 1 * 1 + 3 * 10 = 31$$

$$e_{21} = 1 * 1 + 6 * 2 + 2 * 3 = 19$$

$$e_{22} = 1 * 0 + 6 * 1 + 2 * 10 = 26$$

$$e_{31} = 3 * 1 + 2 * 2 + 4 * 3 = 19$$

$$e_{32} = 3 * 0 + 2 * 1 + 4 * 10 = 42$$

Por lo tanto:

$$A_{3x3} * D_{3x2} = E_{3x2} = \begin{bmatrix} 15 & 31 \\ 19 & 26 \\ 19 & 42 \end{bmatrix}$$

8.0.2. Dependencia e independencia lineal

Dependencia lineal: los vectores $v_1, v_2, \ldots, v_m$ en $\mathbb{R}^n$ son linealmente dependientes (LD) si existen escalares $\lambda_1, \lambda_2, \ldots, \lambda_m$ no todos nulos, tales que:

$$\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_m v_m = \sum_{i=1}^{m} \lambda_i v_i = 0_{nx1}$$

Si dichos escalares no existen, entonces los vectores $v_1, v_2, \ldots, v_m$ son linealmente independientes (LI), Por ejemplo:

$$v_1 = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix}$$

Se tendría que:

$$(2 * v_1) - v_2 = \left(2 * \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix} \right) - \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Por lo tanto, v_1 y v_2 son LD

Por otro lado v_1 y v_3 son LI

$$v_2 = \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix}$$

$$v_3 = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

Resultado: si $n < m$, cualquier conjunto de m vectores n-dimensionales es LD. Ahora bien, en muchos casos resulta provechoso ver las columnas y las filas de una matriz como vectores. En este caso, podemos hablar de rango fila y rango columna. Rango fila (columna) de una matriz: número de filas (columnas) LI.

Resultado: El rango fila es igual al rango columna.

8.0.3. Formas cuadráticas

Sea A una matriz nxn y x un vector en $\mathbb{R}^n$. Una función de x de la forma $x^T Ax$ se conoce como una forma cuadrática. La forma cuadrática $x^T Ax$ se puede escribir como:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$$

$$= \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}}^{n} a_{ij} x_i x_j$$

Veamos el caso de la matriz A:

$$A = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 6 & 2 \\ 3 & 2 & 4 \end{bmatrix}$$

La deducción de la forma cuadrática de la matriz A, considerando $x = (w, y, z)$, sería:

Las diagonales: $4w^2 + 6y^2 + 4z^2$

Fuera de la diagonal: $(1 + 1)wy + (3 + 3)wz + (2 + 2)yz$

La forma cuadrática sería:

$4w^2 + 6y^2 + 4z^2 + (1 + 1)wy + (3 + 3)wz + (2 + 2)yz$

Nota: la definición de forma cuadrática presentada aquí es aquella dada por Harville [77]. Otras definiciones imponen la condición de simetría sobre la matriz A. La conexión viene dada por el siguiente corolario de Harville [77], que lo desarrollaremos más adelante y que reza: Para cualquier forma cuadrática $x^T Ax$ existe una única matriz simétrica S tal que $x^T Ax = x^T Sx$ para todo x.

8.0.4. Algunos tipos de matrices de importancia

Matriz de rango columna (o fila) completo: una matriz tal que todas sus columnas (o filas) son LI. Por ejemplo, la matriz A definida anteriormente es de rango completo, así:

$$rango(A) = 3$$

Matriz diagonal: sea M una matriz cuadrada con dimensiones nxn, se dice que M es diagonal si $m_{ij} = 0 \forall 1 \leq i, j \leq n, i \neq j$. La siguiente matriz es diagonal:

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Matriz identidad o idéntica: es una matriz diagonal tal que todos los elementos de la diagonal son 1. Por ejemplo:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

La matriz identidad es el elemento neutro de la multiplicación matricial, esto es, para todas las matrices reales A y B tales que los productos AI y BI están definidos $AI = A, BI = B$.

Matrices singulares y no singulares: una matriz cuadrada de rango completo se llama no singular. Una matriz cuadrada de rango incompleto se llama singular, y a su determinante es igual a 0 (la definición de determinante la presentamos más adelante). Por ejemplo, la matriz A de $rango(A) = 3$ y $det(A) = 34$ es no singular y la matriz B de rango $rango(B) = 2$ y $det(B) = 0$ es singular.

Matriz invertible: A es invertible si existe una matriz A^{-1} (el $^{-1}$ es la notación para la inversa), tal que $A^{-1} * A = A * A^{-1} = I$. Si existe, la inversa es única, por ejemplo:

$$A = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 6 & 2 \\ 3 & 2 & 4 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 0.59 & 0.06 & -0.47 \\ 0.06 & 0.21 & -0.15 \\ -0.47 & -0.15 & 0.68 \end{bmatrix}$$

Entonces:

$$A * A^{-1} = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 6 & 2 \\ 3 & 2 & 4 \end{bmatrix} * \begin{bmatrix} 0.59 & 0.06 & -0.47 \\ 0.06 & 0.21 & -0.15 \\ -0.47 & -0.15 & 0.68 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Resultado: si A es invertible, entonces A es no singular.

Matriz transpuesta: consideremos la matriz $B_{mxn} = \{b_{ij}\}$, definimos la transpuesta de B, como $B^T_{nxm} = \{b_{ji}\}$. La transpuesta se denota con B^T o B'. En el caso de la matriz B indicada anteriormente, su transpuesta es:

$$B^T = \begin{bmatrix} 3 & 1 & 4 \\ 0 & 7 & 7 \\ 4 & 6 & 10 \end{bmatrix}$$

Matriz simétrica: una matriz $H_{nn} = \{h_{ij}\}$ se define como simétrica, si $h_{ij} = h_{ji}; 1 \leq i, j \leq n, i \neq j$. Si existe la inversa de una matriz simétrica, esta también es simétrica, por ejemplo:

$$H = \begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.65 \\ 0.25 & 0.65 & 1.25 \end{bmatrix}$$

$$H^{-1} = \begin{bmatrix} 1.35 & -0.75 & 0.12 \\ -0.75 & 1.93 & -0.85 \\ 0.12 & -0.85 & 1.22 \end{bmatrix}$$

Matriz definida positiva: la matriz A_{nxn} es definida positiva si: $x^T Ax > 0 \forall x \neq 0$. Cuando es una matriz diagonal, la matriz es definida positiva, si los valores de la diagonal son positivos.

Matriz definida no negativa: la matriz J_{nxn} es definida no negativa si $x^T Jx \geq 0 \forall x \neq 0$

Matriz semi-definida positiva: una matriz que es definida no negativa pero no es definida positiva se define como semi-definida positiva.

De manera análoga, podemos definir matrices definidas no positivas, definidas negativas y semi-definidas negativas, pero no son de mayor relevancia para los fines de este curso.

Las matrices definidas positivas son siempre invertibles (y por lo tanto son de rango completo) y su inversa también es definida positiva.

Determinante de una matriz: existen varias definiciones equivalentes y múltiples interpretaciones. Aquí simplemente diremos, que el determinante es una función que a cada matriz cuadrada le asigna un número real y se representa como $|A|$ o $det(A)$. Uno de los aspectos importantes del determinante es que este nos ayuda a conocer ciertas propiedades de la matriz, como veremos a continuación:

1) una matriz A es invertible (y por lo tanto no singular), si y solo si el determinante de A es diferente de cero. De aquí se sigue que el determinante de A nos informa sobre la existencia de su inversa y sobre su rango, ya que en este caso también sería de rango completo.

2) si A es simétrica y definida positiva, entonces su determinante es positivo.

3) el determinante de una matriz simétrica semi-definida positiva es cero.

Para las matrices A y B descritas anteriormente, tenemos:

$$|A| = det(A) = 34$$

$$|B| = det(B) = 0$$

En estadística, resultan de importancia las matrices simétricas definidas positivas (como se verá más adelante al estudiar la matriz de covarianza). En resumen, este tipo de matrices se caracterizan porque:

1) su determinante es mayor que cero

2) son de rango completo (no singular)

3) su inversa existe, es simétrica y también es definida positiva.

Inversa generalizada.

La matriz inversa generalizada de una matriz B_{nxm} es B^{-}_{mxn} (el símbolo $^{-}$ denota esta inversa) y es una matriz que satisface $BB^{-}B = B$.

Toda matriz tiene por lo menos una inversa generalizada. Cuando la matriz es no singular, como en el caso de A, la inversa generalizada es única y corresponde a A^{-1}, cuando la matriz es singular como es el caso de B y E existen infinitas inversas generalizadas.

La siguiente es una inversa generalizada de la matriz B:

$$B^{-} = \begin{bmatrix} 0.12 & -0.08 & 0.04 \\ -0.13 & 0.13 & 0 \\ 0.08 & -0.02 & 0.05 \end{bmatrix}$$

$$BB^{-}B = \begin{bmatrix} 3 & 0 & 4 \\ 1 & 7 & 6 \\ 4 & 7 & 10 \end{bmatrix} = B$$

La siguiente es una inversa generalizada de la matriz E:

$$E^{-} = \begin{bmatrix} -0.03 & 0.14 & -0.07 \\ 0.02 & -0.07 & 0.05 \end{bmatrix}$$

$$EE^{-}E = \begin{bmatrix} 15 & 31 \\ 19 & 26 \\ 19 & 42 \end{bmatrix}$$

Cerramos esta sección con un resultado que se empleó al estudiar los modelos lineales mixtos y justifica porqué, cualquier función lineal del vector de efectos aleatorios, es predecible.

Teorema 8.1. Sea F una matriz real simétrica y G una matriz definida positiva de las mismas dimensiones que F. Entonces, $F + G$ es una matriz definida positiva.

La matriz de covarianzas: sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias. Definimos su matriz de covarianzas como:

$$\Sigma := \begin{bmatrix} Var\left[Y_1\right] & Cov\left[Y_1, Y_2\right] & \cdots & Cov\left[Y_1, Y_n\right] \\ Cov\left[Y_2, Y_1\right] & Var\left[Y_2\right] & \cdots & Cov\left[Y_2, Y_n\right] \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left[Y_n, Y_1\right] & Cov\left[Y_n, Y_2\right] & \cdots & Var\left[Y_n\right] \end{bmatrix}$$

Para simplificar la notación, podemos escribirla así:

$$\Sigma := \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Recordemos que la covarianza es simétrica. $Cov\left[Y_i, Y_j\right] = Cov\left[Y_j, Y_i\right]; 1 \leq i, j \leq n, i \neq j$, por lo tanto Σ es una matriz simétrica. Además, se puede demostrar que la matriz de covarianzas es definida no negativa, sin embargo, en muchos problemas en estadística, el interés se centra en matrices de covarianza definidas positivas.

La matriz Σ también puede encontrarse con el nombre de matriz de varianzas (elementos de la diagonal) y covarianzas (elementos fuera de la diagonal). Recordemos que para una variable aleatoria unidimensional Y, $Cov\left[Y, Y\right] = Var\left[Y\right]$, así, al usar el nombre matriz de covarianzas no se está omitiendo el hecho de que esta también contiene las varianzas de las variables aleatorias en cuestión. Con la matriz Σ se pueden conocer las correlaciones (ρ) entre las variables como se indica en el Capítulo 9.

8.0.5. Ejercicios en R-project

Suma de matrices:

```
A=matrix(nrow=3,ncol=3,byrow=TRUE,c(
4, 1, 3,
1, 6, 2,
3, 2, 4))
A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
det(A)
```

```
## [1] 34
```

```
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 8.70 4.41 0.89
##
## $vectors
##       [,1]  [,2]  [,3]
## [1,] -0.51  0.55  0.66
## [2,] -0.63 -0.76  0.16
## [3,] -0.59  0.33 -0.74
```

```
B=matrix(nrow=3,ncol=3,byrow=TRUE,c(
3, 0, 4,
1, 7, 6,
4, 7,10))
B
```

```
##      [,1] [,2] [,3]
## [1,]    3    0    4
## [2,]    1    7    6
## [3,]    4    7   10
```

```
A+B
```

```
##      [,1] [,2] [,3]
## [1,]    7    1    7
## [2,]    2   13    8
## [3,]    7    9   14
```

Multiplicación con un escalar:

```
c_11=matrix(nrow=1,ncol=1,c(2))
c_11%x%A
```

```
##      [,1] [,2] [,3]
```

```
## [1,]    8    2    6
## [2,]    2   12    4
## [3,]    6    4    8
```

Multiplicación de matrices:

```
D=matrix(nrow=3,ncol=2,byrow=TRUE,c(
1, 0,
2, 1,
3,10))
D
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    2    1
## [3,]    3   10
```

```
E=A%*%D
E
```

```
##      [,1] [,2]
## [1,]   15   31
## [2,]   19   26
## [3,]   19   42
```

Dependencia lineal:

```
v_1=matrix(nrow=3,ncol=1,byrow=TRUE,c(
0,
2,
3))
v_1
```

```
##      [,1]
## [1,]    0
## [2,]    2
## [3,]    3
```

```
v_2=matrix(nrow=3,ncol=1,byrow=TRUE,c(
0,
4,
```

```
6))
v_2
```

```
##       [,1]
## [1,]    0
## [2,]    4
## [3,]    6
```

```
(c_11%x%v_1)-v_2
```

```
##       [,1]
## [1,]    0
## [2,]    0
## [3,]    0
```

```
v_3=matrix(nrow=3,ncol=1,byrow=TRUE,c(
3,
1,
2))
v_3
```

```
##       [,1]
## [1,]    3
## [2,]    1
## [3,]    2
```

Forma cuadrática:

```
x=matrix(nrow=3,ncol=1,byrow=TRUE,c(
8.7,
4.4,
0.9))
x
```

```
##       [,1]
## [1,]  8.7
## [2,]  4.4
## [3,]  0.9
```

```
tx=t(x)
tx


##      [,1] [,2] [,3]
## [1,]  8.7  4.4  0.9
```

```
Resultado=t(x)%*%A%*%x
Resultado


##      [,1]
## [1,]  562
```

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

```
tx[1,1]%*%A[1,1]%*%x[1,1]+
tx[1,2]%*%A[1,2]%*%x[1,1]+
tx[1,3]%*%A[1,3]%*%x[1,1]+

tx[1,1]%*%A[2,1]%*%x[2,1]+
tx[1,2]%*%A[2,2]%*%x[2,1]+
tx[1,3]%*%A[2,3]%*%x[2,1]+

tx[1,1]%*%A[3,1]%*%x[3,1]+
tx[1,2]%*%A[3,2]%*%x[3,1]+
tx[1,3]%*%A[3,3]%*%x[3,1]


##      [,1]
## [1,]  562
```

$$= \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{1 \le i,j \le ni \ne j}^{n} a_{ij} x_i x_j$$

```
tx[1,1]^2%*%A[1,1]+
tx[1,2]^2%*%A[2,2]+
tx[1,3]^2%*%A[3,3]+
tx[1,2]%*%A[1,2]%*%x[1,1]+
```

```
tx[1,3]%*%A[1,3]%*%x[1,1]+
tx[1,1]%*%A[2,1]%*%x[2,1]+
tx[1,3]%*%A[2,3]%*%x[2,1]+
tx[1,1]%*%A[3,1]%*%x[3,1]+
tx[1,2]%*%A[3,2]%*%x[3,1]
```

```
##      [,1]
## [1,]  562
```

Matriz de rango completo:

```
A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
nrow(A)
```

```
## [1] 3
```

```
ncol(A)
```

```
## [1] 3
```

```
rango_A=as.matrix(qr(A)$rank)
rango_A
```

```
##      [,1]
## [1,]    3
```

Matriz diagonal:

```
M=matrix(nrow=3,ncol=3,0);diag(M)=c(1,2,3)
M
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
```

```
## [2,]    0    2    0
## [3,]    0    0    3
```

Matriz identidad:

```
I=matrix(nrow=3,ncol=3,0);diag(I)=1
I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
I%*%A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
A%*%I
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
I%*%B
```

```
##      [,1] [,2] [,3]
## [1,]    3    0    4
## [2,]    1    7    6
## [3,]    4    7   10
```

```
B%*%I
```

```
##      [,1] [,2] [,3]
## [1,]    3    0    4
## [2,]    1    7    6
## [3,]    4    7   10
```

Matrices singulares y no singulares:

```
#matriz no singular
A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
rango_A#rango completo
```

```
##      [,1]
## [1,]    3
```

```
det(A)#determinante diferente de 0
```

```
## [1] 34
```

```
#matriz singular
B
```

```
##      [,1] [,2] [,3]
## [1,]    3    0    4
## [2,]    1    7    6
## [3,]    4    7   10
```

```
ncol(B)
```

```
## [1] 3
```

```
nrow(B)
```

```
## [1] 3
```

```
det(B)#determinante igual que 0
```

```
## [1] 0
```

```
rango_B=qr(B)$rank
rango_B#rango incompleto


## [1] 2
```

Matrices invertibles. Utilizaremos la librería ≪MASS≫ [18] para calcular la inversa:

```
library("MASS")
Ainv=solve(A)
Ainv


##         [,1]   [,2]  [,3]
## [1,]  0.588  0.059 -0.47
## [2,]  0.059  0.206 -0.15
## [3,] -0.471 -0.147  0.68
```

```
Ainv%*%A


##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1


A%*%Ainv


##      [,1]      [,2] [,3]
## [1,]    1  0.0e+00    0
## [2,]    0  1.0e+00    0
## [3,]    0 -1.1e-16    1
```

Transpuesta de una matriz:

```
Btrans=t(B)
Btrans


##      [,1] [,2] [,3]
## [1,]    3    1    4
## [2,]    0    7    7
## [3,]    4    6   10
```

Matriz simétrica:

```
H=matrix(nrow=3, ncol=3,c(1.00,0.50,0.25,
                          .5,1.00,0.65,
                          .25,0.65,1.25), byrow=TRUE)
H
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.50 0.25
## [2,] 0.50 1.00 0.65
## [3,] 0.25 0.65 1.25
```

```
Hinv=solve(H)
Hinv
```

```
##       [,1]  [,2]  [,3]
## [1,]  1.35 -0.75  0.12
## [2,] -0.75  1.93 -0.85
## [3,]  0.12 -0.85  1.22
```

Matriz definida positiva:

```
J=matrix(nrow=3,ncol=3,c(
  1,0, 0,
  0,2, 0,
  0,0, 3
), byrow=TRUE)
J
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    2    0
## [3,]    0    0    3
```

Matriz definida positiva:

```
A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4
```

```
A_lambdas=eigen(A)$values
A_lambdas


## [1] 8.70 4.41 0.89
```

```
A_defiposi=t(x)%*%A%*%x
A_defiposi


##      [,1]
## [1,]  562
```

```
#En el caso de J
J_lambdas=eigen(J)$values
```

Determinante de una matriz:

```
A


##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    1    6    2
## [3,]    3    2    4


detA=det(A)
detA


## [1] 34
```

Inversa generalizada:

```
B


##      [,1] [,2] [,3]
## [1,]    3    0    4
## [2,]    1    7    6
## [3,]    4    7   10


detB=det(B)
detB# por consiguente no tiene inversa
```

```
## [1] 0


#Tendría inversa generalizada
Binvgen=ginv(B)
Binvgen


##         [,1]   [,2]    [,3]
## [1,]  0.122 -0.080  0.0411
## [2,] -0.130  0.126 -0.0033
## [3,]  0.076 -0.023  0.0525
```

```
B%*%Binvgen%*%B


##      [,1]    [,2] [,3]
## [1,]    3 3.1e-15    4
## [2,]    1 7.0e+00    6
## [3,]    4 7.0e+00   10
```

```
E


##      [,1] [,2]
## [1,]   15   31
## [2,]   19   26
## [3,]   19   42


#Tendría inversa generalizada
Einvgen=ginv(E)
Einvgen


##         [,1]   [,2]   [,3]
## [1,] -0.026  0.142 -0.069
## [2,]  0.022 -0.066  0.048
```

```
confirmado=E%*%Einvgen%*%E
confirmado


##      [,1] [,2]
## [1,]   15   31
## [2,]   19   26
## [3,]   19   42
```

Matriz de varianzas y covarianzas y matriz de correlaciones:

```
Sigma=matrix(nrow=3,ncol=3, c(100,30,-5,
                              30,10,5,
                              -5,5,6),byrow=TRUE)
Sigma
```

```
##      [,1] [,2] [,3]
## [1,]  100   30   -5
## [2,]   30   10    5
## [3,]   -5    5    6
```

```
rho=cov2cor(Sigma)
rho
```

```
##       [,1] [,2]  [,3]
## [1,]  1.00 0.95 -0.20
## [2,]  0.95 1.00  0.65
## [3,] -0.20 0.65  1.00
```

9 CAPÍTULO NUEVE

Bases de probabilidad

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

A través de este texto se hará un uso extensivo de modelos matemáticos, en particular, una clase de ellos denominados modelos estadísticos, los que se definen como una familia de modelos matemáticos que buscan representar el proceso que generó los datos y constan de al menos dos componentes: una función matemática que expresa la relación entre la *esperanza* (u otro parámetro de localización) de la variable respuesta y un conjunto de variables explicativas (componente sistemático) y una *distribución de probabilidad* que caracteriza la *variación aleatoria* de la respuesta (componente estocástico).

Nótese en el párrafo anterior, que las palabras que se encuentran en letra cursiva son términos desconocidos para la mayoría de las personas que se están iniciando en el estudio del mejoramiento genético y áreas afines e incluso, para algunas que cuentan con algún grado de conocimiento en la materia. Además, son conceptos que se tratan de una manera superficial en la mayoría de los cursos de introducción a la estadística que se ofrecen en las carreras del sector agropecuario.

Estas definiciones hacen parte de la rama de la matemática conocida como probabilidad, la cual estudia fenómenos aleatorios y se encarga de dar un tratamiento matemático al azar. Ahora bien, es importante reiterar que la probabilidad es una rama de la matemática que no hace parte de la estadística, por lo tanto, se trata de dos disciplinas diferentes, pero estrechamente ligadas tanto en la teoría como en la

práctica. La estadística estudia la variación aleatoria utilizando los axiomas de la teoría de la probabilidad; esta definición deja aún más clara la relación entre la estadística y la probabilidad, pero a la vez indica que son áreas separadas.

En este capítulo se presenta una introducción muy general a la teoría de la probabilidad, que busca brindar algunos conceptos clave que permitan al lector comprender la notación y las propiedades de los modelos estadísticos usados en evaluación genética. De ninguna manera se pretende hacer una presentación totalmente rigurosa y profunda del tema; además, con el ánimo de enfatizar en la simplicidad y sacrificando el rigor, algunas definiciones no son completamente formales, esta situación se aclara donde se requiera precisarlo. Iniciamos con la definición de algunos conceptos fundamentales de la teoría de la probabilidad.

Experimento aleatorio: un experimento es aleatorio, si su resultado no puede ser determinado de antemano. Al emplear la palabra "experimento", muchos piensan en una situación que requiere un ensayo muy elaborado como los que se realizan en investigación agropecuaria; no obstante, en este contexto se pueden citar ejemplos mucho más sencillos que pueden encontrarse a diario en el campo de la producción animal:

1) determinar el peso de un individuo

2) medir la producción de leche de una cabra

3) determinar el perímetro del cráneo de un perro Rottweiler

4) genotipificar un búfalo.

Espacio muestral (S): conjunto de todos los posibles resultados de un experimento aleatorio.

Ahora bien, muchos conceptos de la teoría básica de probabilidad requieren la definición de sigma-álgebra, pero un tratamiento detallado de dicho concepto va más allá del ámbito de este libro; por lo tanto, algunas de las definiciones que implican una sigma-álgebra no serán totalmente rigurosas (matemáticamente hablando). Aquí simplemente diremos que una sigma-álgebra es una colección de subconjuntos de S que satisface ciertas condiciones. La importancia de mencionar la sigma-álgebra viene dada por la siguiente definición.

Eventos: son los elementos de una sigma-álgebra. Sin embargo, una definición más pragmática que funciona para efectos de este texto es la siguiente (siguiendo a Casella y Berger [61]): un evento es un subconjunto del espacio muestral (incluyéndolo). Teóricamente, para poder definir evento como cualquier subconjunto de S, se debe elegir un tipo particular de sigma-álgebra, pero estos detalles se omiten. La importancia de esta definición es que los eventos son la clase de subconjuntos de S a los que podemos asignar medidas de probabilidad, un concepto que se introduce más adelante.

Algunos eventos de interés son:

1) el evento $A \cup B$ (unión de A y B) ocurre, si y solo si, A ocurre, B ocurre o los dos ocurren

2) el evento $A \cap B$ (intersección de A y B) ocurre, si y solo si, A y B ocurren

3) el evento A^C (complemento de A) ocurre, si y solo si, A no ocurre

4) el evento $A \setminus B$ (diferencia de conjuntos) ocurre, si y solo si, A ocurre pero B no.

Eventos mutuamente excluyentes o disjuntos: dos eventos A y B son mutuamente excluyentes o disjuntos si estos no pueden suceder simultáneamente, es decir, sucede A o sucede B pero nunca los dos a la vez, formalmente escribimos:

$$A \cap B = \varnothing$$

Donde $\varnothing$ es el conjunto vacío.

La siguiente es una definición central en la teoría de la probabilidad, pues dio lugar a la axiomatización de la disciplina, esto permite darle un tratamiento matemático riguroso, independientemente de la interpretación que se le dé a la probabilidad.

Medida de probabilidad: sea S un espacio muestral y F una sigma-álgebra, una función $P(\bullet)$ con dominio F que satisface:

1) $P(A) \geq 0 \forall A \in F$

2) $P(S) = 1$

3) Si $A_1, A_2, ... \in F$ son disjuntos.

$$\Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Se conoce como una medida de probabilidad. Además, estas tres propiedades se conocen como axiomas de Kolmogorov, en honor a quien los creó, el matemático ruso Andréi Nikoláyevich Kolmogórov (1903-1987). Este desarrollo axiomatizó el concepto de probabilidad, de allí su gran importancia.

La tripla (S, F, P) se conoce como espacio de probabilidad y es un constructo matemático que nos dice cuáles son los resultados posibles del experimento aleatorio (S), a qué subconjuntos del mismo les podemos asignar una medida de probabilidad (elementos de F) y cómo es tal medida (P).

Una vez presentados los conceptos de evento y medida de probabilidad, podemos hablar de evento seguro, aquel tal que $P(A) = 1$, y evento imposible, aquel tal que $P(A) = 0$. Además, para un par de eventos A y B, se tiene:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Ejemplo de medida de probabilidad en genética : considere un locus bialélico (alelos denotados como I y R) bajo equilibrio de Hardy-Weimberg, en el genoma de un organismo diploide. El experimento aleatorio consiste en seleccionar un animal al azar y determinar su genotipo para este locus. Definamos los eventos.

$A :=$ "el genotipo es II"

$B :=$ "el genotipo es IR"

$C :=$ "el genotipo es RR".

Entonces $S = \{A, B, C\}$. En este ejemplo definimos $F =$ partes de S (todos los posibles subconjuntos de S), esto es:

$$F = (\emptyset, A, B, C, A \cup B, B \cup C, A \cup C, S)$$

Nótese que el individuo solo puede tener uno de los tres genotipos posibles, así, los eventos A, B y C son mutuamente excluyentes. Además, la medida de probabilidad se construye de la manera usual empleando el supuesto de equilibrio de Hardy-Weimberg. Así, definimos p como la probabilidad de observar el alelo I y corresponde a la frecuencia relativa de este alelo en la población, luego, se obtienen las frecuencias de cada genotipo como p^2 para II, $2p(1-p)$ para IR y $(1-p)^2$ para RR, es decir, $p^2 = P(A), 2p(1-p) = P(B), (1-p)^2 = P(C)$.

Ahora, procedemos a examinar los axiomas de Kolmogorov para esta media de probabilidad, asi:

$$P(S) = p^2 + 2p(1-p) + (1-p)^2$$
$$(p + (1-p))^2 = 1$$

Además, en efecto, $P(E) \geq 0 \; \forall \, E \in F$.

Es importante considerar la siguiente propiedad: para todo evento A, $P(A^C) = 1 - P(A)$, esto nos permite comprobar el axioma restante. Al calcular $P(A \cup B)$, se tiene en cuenta que el evento $A \cup B$ ocurre si y solo si el evento C no ocurre, esto es:

$$\begin{aligned} P(A \cup B) &= 1 - P(C) \\ &= 1 - (1-p)^2 \\ &= p^2 + 2p(1-p) + (1-p)^2 - (1-p)^2 \\ &= p^2 + 2p(1-p) \\ &= P(A) + P(B) \end{aligned}$$

Siguiendo argumentos similares, se tiene que:

$$P(B \cup C) = 2p(1-p) + (1-p)^2 = P(B) + P(C)$$
$$P(A \cup C) = p^2 + (1-p)^2 = P(A) + P(C)$$

Además, como $P(\varnothing) = 0$ y para cualquier evento E, $E \cap \varnothing = \varnothing$, se llega a que:

$$\begin{aligned} P(E \cup \varnothing) &= P(E) + P(\varnothing) - P(\varnothing) \\ &= P(E) \end{aligned}$$

Finalmente:

$$\begin{aligned} P(A \cup B \cup C \cup \varnothing) &= P(S) \\ &= 1 \\ &= p^2 + 2p(1-p) + (1-p)^2 + 0 \\ &= P(A) + P(B) + P(C) + P(\varnothing) \end{aligned}$$

Por lo tanto, se satisface el tercer axioma.

Ahora que se ha dado la definición de medida de probabilidad, podemos discutir la de independencia probabilística. De manera informal, dos eventos son independientes cuando la ocurrencia de uno no afecta la del otro. A nivel probabilístico, esto implica que:

$$P(A \cap B) = P(A)P(B)$$

Es decir, la probabilidad de que los dos eventos ocurran es igual al producto de las probabilidades marginales.

Para continuar con esta breve introducción a la teoría básica de la probabilidad, discutimos el concepto de probabilidad condicional. Si sabemos que el evento B ha

ocurrido, ¿cuál es la probabilidad de que ocurra A?, la respuesta a esta pregunta está dada por la probabilidad del evento A dado que se observó el evento B, que se simboliza como $P(A|B)$ y se lee "probabilidad de A dado B". Entonces, si B aporta información sobre A, la observación "B ha ocurrido", proporciona información sobre la probabilidad de que A ocurra. La pregunta que surge en este punto es: ¿cómo computar esta probabilidad?, a continuación, se presenta un resultado que nos permite calcularla.

Regla del producto. La probabilidad del evento $A \cap B$ puede escribirse de como sigue:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

Nótese que a partir de la regla del producto podemos encontrar una expresión para la probabilidad de A dado B y la probabilidad de B dado A.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Es importante acotar que $P(A|B)$ y $P(B|A)$ están definidas siempre y cuando $P(B) \neq 0$ y $P(A) \neq 0$, respectivamente.

Esta regla, junto con el teorema que se presenta enseguida son ampliamente utilizados y nos llevarán a un resultado de gran importancia, no solamente en mejoramiento genético, sino en la estadística en general: el teorema o regla de bayes.

Teorema o ley de la probabilidad total. Si $B_1, B_2, \ldots, B_n$ son eventos que forman una partición del espacio muestral y A es otro evento del mismo espacio de probabilidad, entonces:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Nótese que para cada i, $P(A|B_i)P(B_i) = P(A \cap B_i)$, por lo tanto, este teorema nos dice que la probabilidad del evento A, puede calcularse como la suma de las probabilidades de sus intersecciones con los eventos $B_1, B_2, \ldots, B_n$. Es importante notar que, en el teorema se menciona que estos eventos forman una partición del espacio muestral, esto quiere decir, no solamente que la unión de estos eventos es igual al espacio muestral, sino que son disjuntos por pares.

Ahora exploramos una propiedad para el caso de eventos independientes. La intuición nos dice que si A y B son independientes, entonces el hecho de que ocurra B

no provee información sobre la probabilidad de ocurrencia de A y de manera similar, la ocurrencia de A no provee información alguna sobre la probabilidad de ocurrencia de B. Por lo tanto, si A y B son independientes, se deberían satisfacer las siguientes identidades: $P(A|B) = P(A)$ y $P(B|A) = P(B)$. Estas son fácilmente establecidas siguiendo la propiedad de eventos independientes que se mencionó previamente. Así, si $P(B) \neq 0$, $P(A) \neq 0$ y además A y B son independientes, entonces:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(A)P(B)}{P(B)} (\because \text{independencia})$$

$$= P(A)$$

Similarmente:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$= \frac{P(A)P(B)}{P(A)}$$

$$= P(B)$$

Teorema o regla de bayes. Sean A y B eventos asociados a un mismo espacio de probabilidad, entonces la probabilidad condicional de A dado B se calcula como sigue:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Para llegar a este resultado, simplemente se aplica la fórmula que ya se había presentado:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Posteriormente, se reescribe el numerador usando la regla del producto. Por otro lado, existe una versión equivalente que involucra la regla o teorema de la probabilidad total, pero que requiere modificar el enunciado. Sean A_1, A_2, ..., A_n eventos que forman una partición del espacio muestral y B es otro evento del mismo espacio de probabilidad, entonces, para cualquier $i = 1, 2, \ldots, n$:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Nótese que la regla de la probabilidad total se aplicó en el denominador.

Variables aleatorias : un concepto central en probabilidad y en sus múltiples aplicaciones es el de variable aleatoria, puesto que es el tipo de variables estudiadas por esta disciplina. En el campo del mejoramiento genético animal las variables aleatorias están siempre presentes. El fenotipo se modela como una variable aleatoria observable, mientras que los valores genéticos de los individuos (capítulos 4 hasta la sección 6.2) o los efectos aditivos de los marcadores moleculares (capítulo 7) se consideran variables aleatorias no observables. En libros introductorios a la genética cuantitativa como el libro de Falconer y Mackay [38], se hace un tratamiento de los diferentes tipos de efectos genéticos sin necesidad de lidiar con conceptos y propiedades de variables aleatorias, sin embargo, al estudiar los modelos estadísticos utilizados en evaluación genética, es necesario tener una comprensión mínima de las variables aleatorias y algunas funciones asociadas a las mismas como la función de distribución acumulativa, y las funciones de densidad de probabilidad (caso contínuo) y de masa de probabilidad (caso discreto). Iniciamos con una definición informal del tipo de variables aleatorias que nos interesan en este texto.

Variable aleatoria real: una variable aleatoria real es una función que va del espacio muestral a los números reales.

Así, la variable aleatoria real tiene como dominio el espacio muestral y como rango los números reales (típicamente un subconjunto). Ahora nos enfocamos en una función que caracteriza las variables aleatorias, de allí su importancia.

Función de distribución acumulativa. sea X una variable aleatoria real, su función de distribución acumulativa (FDA) se define como:

$$F_X(x) = P(X \leq x), \text{para todo } x \in R$$

Una puntualización importante en términos de notación es la siguiente: las letras mayúsculas indican la variable aleatoria, mientras que letras minúsculas indican un valor particular de la misma; un valor particular que toma una variable aleatoria se conoce como una realización o valor realizado. Así, en la anterior definición, se puede escribir de manera equivalente:

$$F_X(a) = P(X \leq a), \text{para todo } a \in R$$

Nótese que esta función está definida para todos los números reales. En esta notación, el subíndice X de la letra F representa la variable aleatoria que se está

considerando, puesto que usualmente se usa la letra F para denotar la FDA de cualquier variable aleatoria. Esta función goza de varias propiedades, pero algunas son avanzadas para el alcance de este texto, por lo tanto, nos restringimos a mencionar que es una función no decreciente, y que además de la probabilidad que se presenta en su definición, nos permite calcular probabilidades de la siguiente manera:

$$P(m < X \leq n) = F_X(n) - F_X(m)$$

Nótese que el intervalo es abierto a izquierda y cerrado a derecha, este detalle se tratará más adelante. Otra propiedad muy útil de esta función es que su naturaleza nos permite establecer el tipo de variable aleatoria, pero antes, debemos describir cuáles son los tipos de variable aleatoria que podemos encontrar. En primer lugar, una variable discreta es aquella que toma valores en un conjunto finito o en un conjunto infinito numerable, por otro lado, una variable continua toma valores en un intervalo de números reales (o simplemente intervalo real) que es un conjunto infinito no numerable. Por ejemplo, el tamaño de camada en perros es un número entero positivo que puede variar desde 1 hasta algún umbral superior como 18, así, esta variable puede tomar valores en un conjunto de tamaño 18, es una variable discreta. Por otro lado, la altura a la cruz de un perro tomará valores en un intervalo de números reales de la forma $[a, b]$; por ejemplo, en la raza Rottweiler, la alzada de los machos a la cruz va de 61 a 68 cm, así esta variable pertenece al intervalo $[61, 68]$, el cual, como cualquier intervalo real, tiene infinitos elementos y no es numerable, así, tenemos una variable continua. Una vez hecha esta salvedad, podemos definir tres tipos de variables aleatorias observando directamente la naturaleza de su FDA así:

1): variable aleatoria discreta. Si la FDA de X es escalonada (tiene saltos), entonces X es una variable aleatoria discreta.

2): variable aleatoria continua.Si la FDA de X es continua, entonces X es una variable aleatoria continua.

3): variable aleatoria mixta. Si la FDA de X se puede expresar como una combinación lineal de una función continua y una función escalonada, entonces X es una variable aleatoria mixta.

Nos enfocamos en los dos primeros tipos. La Figura FIGURA NRO. 9.1 muestra la FDA para una variable aleatoria discreta y la FIGURA NRO. 9.2 la de una variable aleatoria continua.

Las letras a, b y c indican la magnitud de cada salto y esta corresponde a la probabilidad de que la variable aleatoria tome el valor que corresponde en el eje X. De aquí se pueden calcular algunas probabilidades como las siguientes:

$$P(X = 2) = a$$
$$P(X = 1) = b$$

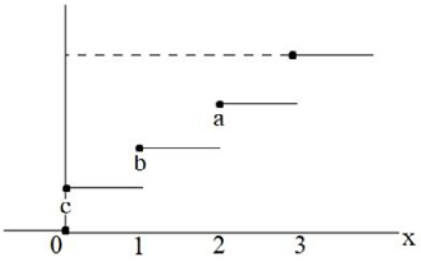

***Figura* 9.1:** Función de distribución acumulativa de una variable aleatoria discreta que toma valores en el conjunto {0,1,2,3}.
Fuente: elaboración propia (2024).

$$P(X = 0) = c$$

$$P(X \leq 2) = a + b + c$$

$$P(1 \leq X \leq 2) = a + b$$

$$P(X < 2) = b + c$$

La propiedad de que la magnitud del salto en cada uno de los posibles valores de la variable aleatoria sea la probabilidad de que esta tome dicho valor es muy conveniente pues le concede una interpretación a esta magnitud. Esto se traduce en una característica de la función de masa de probabilidad, concepto que se estudiará más adelante. Además, así puede verse la razón por la cual

$$P(m < X \leq n) = F_X(n) - F_X(m)$$

Pues, al sustraer $F_X(m)$ se incluye la probabilidad de que la variable aleatoria tome este valor y, por lo tanto, el intervalo no incluye su límite inferior, es decir, es abierto a izquierda.

Ahora volvemos nuestra atención hacia el caso continuo (FIGURA NRO. 9.2). El siguiente es un típico ejemplo de cómo luce la FDA de este tipo de variable aleatoria, aunque vale mencionar que no siempre se tendrá esta forma sigmoidea.

Haciendo a un lado las explicaciones técnicas, la probabilidad de que estas variables aleatorias tomen un valor puntual es cero, esto afecta de cierta forma la interpretabilidad, pero trae una propiedad que se escribe de la siguiente manera:

$$\begin{aligned} F_X(n) - F_X(m) &= P(m < X \leq n) \\ &= P(m \leq X \leq n) \\ &= P(m \leq X < n) \end{aligned}$$

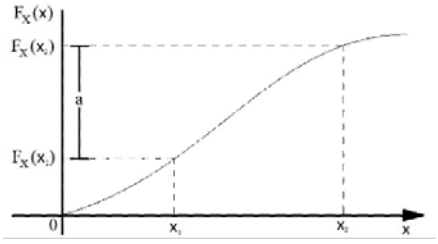

***Figura* 9.2:** Función de distribución acumulativa de una variable aleatoria continua que toma valores en los reales positivos.
Fuente: elaboración propia (2024).

$$= P(m < X < n)$$

Lo cual es ventajoso, porque al calcular probabilidades en intervalos no debemos preocuparnos por los valores en la frontera. En la FIGURA NRO. 9.3 la magnitud a representa la diferencia entre $F_X(x_2)$ y $F_X(x_1)$ y, por lo tanto, es la probabilidad de que la variable tome valores entre x_1 y x_2 y como se acaba de discutir, el intervalo puede ser cerrado, abierto o mixto pues esto no cambia la probabilidad en el caso continuo.

Ejemplo de construcción de una variable aleatoria discreta en genética: continuemos con el ejemplo de un locus bialélico para un organismo diploide. Definimos la variable aleatoria discreta X como el conteo de alelos I en el genotipo observado. Es decir:

$$X = \begin{cases} 2, \text{ si } A \text{ ocurre} \\ 1, \text{ si } B \text{ ocurre} \\ 0, \text{ si } C \text{ ocurre} \end{cases} = \begin{cases} 2, \text{ si el genotipo es } II \\ 1, \text{ si el genotipo es } IR \\ 0, \text{ si el genotipo es } RR \end{cases}$$

Nótese que X es una función que va de S (espacio muestral) a los números reales, en este caso particular al subconjunto $\{0, 1, 2\}$. A esta variable aleatoria se le conoce como la dosis génica o dosis alélica. La FIGURA NRO. 9.3 presenta la FDA para esta variable aleatoria.

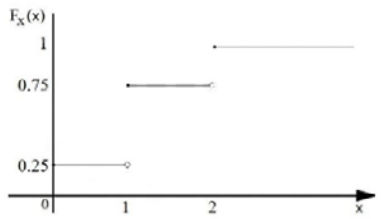

***Figura* 9.3:** Función de distribución acumulativa de la dosis génica cuando $p = \frac{1}{2}$.
Fuente: elaboración propia (2024).

A continuación, se presentan dos conceptos previamente mencionados, la función de masa de probabilidad (caso discreto) y función de densidad de probabilidad (caso continuo), las cuales están ligadas a la FDA.

Función de masa de probabilidad (FMP). Sea X una variable aleatoria discreta. Definimos su FMP como:

$$f_X(x) = P(X = x) \, \forall \, x \in R$$

Continuando con el ejemplo de genética, asumiendo que el locus en cuestión está bajo equilibrio de Hardy-Weinberg (EHW), la FMP de la dosis génica puede escribirse como sigue:

$$f_X(x) = P(X = x) = \begin{cases} p^2, \text{ si } x = 2 \\ 2p(1-p), \text{ si } x = 1 \\ (1-p)^2, \text{ si } x = 0 \\ 0, \text{ en otro caso} \end{cases}$$

Como se mencionó antes, una propiedad de la que goza la FMP es que, al graficarla, su altura indica la probabilidad de que la variable aleatoria tome el valor respectivo. La FIGURA NRO. 9.4 indica la FMP para la dosis génica.

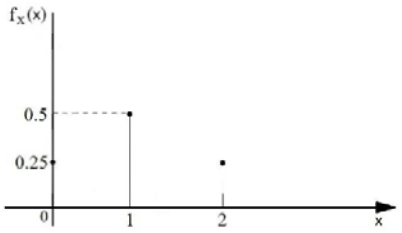

***Figura* 9.4:** Función de masa de probabilidad para la dosis génica con $p = \frac{1}{2}$. Las líneas verticales se incluyen meramente para facilitar la visualización de la gráfica. Fuente: elaboración propia (2024).

La FMP está definida para todos los reales, pero solo tomará valores positivos en los puntos donde la FDA presenta saltos, y su valor en cada uno de dichos puntos corresponde a la magnitud del salto. Así, al sumar los valores no nulos de la FMP desde el valor más pequeño que puede tomar la variable aleatoria hasta un punto a cuya FMP no es nula, o el valor más pequeño que a, y más cercano a a, cuya FMP no es nula, tenemos $F_X(a)$. Sumado a esto, al observar la gráfica de una FMP, las regiones con mayor altura corresponden a aquellas en las que la variable aleatoria toma valores con mayor frecuencia.

Función de densidad de probabilidad (FDP): sea X una variable aleatoria continua. La FDP de X es la función que satisface:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

Es importante recalcar que la FDP no siempre existe y que aquellas variables aleatorias para las que existe se denominan absolutamente continuas. La mayoría de las distribuciones continuas que se emplean en bioestadística, y en particular, en genética cuantitativa y mejoramiento genético, son de este tipo. Por ejemplo, en selección genómica, un supuesto usual es asumir que los efectos aditivos de los marcadores son variables aleatorias independientes, cada una de las cuales siguen una distribución normal o Gaussiana con media 0 y varianza σ_m^2, entonces su densidad existe y tiene la forma:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{\frac{-1}{2\sigma_m^2}x^2}, x \in \mathbb{R}$$

Si se grafica esta función, la altura de la misma en un punto dado no puede interpretarse como la probabilidad de que la variable aleatoria tome este valor, de hecho, es posible que la altura de la función sea mayor que 1, lo cual no representa ninguna falencia ya que esto no está proporcionando una medida de probabilidad. La FIGURA NRO. 9.5 muestra la gráfica de la FDP de una variable aleatoria absolutamente continua que toma valores en los números reales, se muestra el área bajo la curva entre los puntos x_1 y x_2 que corresponde a la probabilidad de que la variable tome valores en este intervalo.

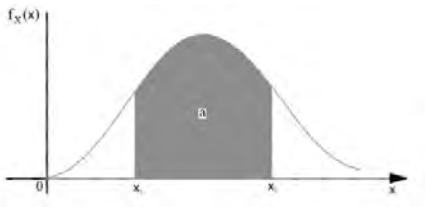

***Figura* 9.5:** FDP de una variable aleatoria absolutamente continua que toma valores en los números reales.
Fuente: elaboración propia (2024).

La FDP nos permite calcular la probabilidad de que la variable aleatoria tome valores en un intervalo dado. Consideremos la FIGURA NRO. 9.5, el área bajo la curva entre los puntos x_1 y x_2, representada por la letra a, corresponde a:

$$P(x_1 < X \leq x_2) = P(x_1 \leq X \leq x_2)$$

$$= P(x_1 < X < x_2)$$

$$= P(x_1 \leq X < x_2)$$

Nuevamente, esto ilustra una ventaja que tienen las variables aleatorias continuas, la probabilidad no varía si el intervalo incluye o no los extremos. Al observar la FIGURA NRO. 9.4 y la FIGURA NRO. 9.5, se advierte la relación que existe entre la FDA y la FDP puesto que a es el área bajo la curva entre x_1 y x_2 de la FDP, pero también es la diferencia entre $F_X(x_2)$ y $F_X(x_1)$ en la FDA. Esta relación siempre se tiene y viene dada por un resultado conocido como el teorema fundamental del cálculo, esto es, para $x_2 > x_1$:

$$F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx$$

Además, el teorema fundamental del cálculo también nos dice que la densidad es la derivada de la FDA. Se emplean integrales porque estas permiten calcular el área bajo la curva de una función (siempre y cuando esta sea integrable, que es el caso de la FDP). El lector interesado en mayores detalles sobre integración y el teorema fundamental del cálculo puede consultar el libro de cálculo de Stewart [79] para una presentación amigable del tema, o el libro de cálculo de Spivak [80] para un tratamiento más detallado. Similar al caso de la FMP, al observar la gráfica de la FDP, regiones en las que la función es mayor son aquellas en las que la variable aleatoria toma valores con mayor frecuencia.

Al discutir los saltos en la FDA en el caso discreto, se mencionaron los valores posibles que una variable aleatoria puede tomar; similarmente, al discutir la FDA en la FIGURA NRO. 9.5, se menciona que se trata de una variable aleatoria que toma valores en los reales. Así, cuando se piensa en ejemplos del sector pecuario como tamaño de camada, dosis génica, o alzada a la cruz, se puede notar que cada variable aleatoria toma valores en un conjunto dado. Cuando se presentaron las variables aleatorias discretas y continuas, se habló sobre la naturaleza de los conjuntos en los que toman valores: finitos, infinitos numerables y no numerables. Ahora nos fijaremos en estos conjuntos con mayor detalle, por ejemplo: podemos tener dos variables aleatorias que toman valores en conjuntos finitos y por consiguiente son discretas, pero estos conjuntos pueden ser completamente diferentes. Por lo tanto, a continuación, se introduce el concepto de conjunto soporte o simplemente soporte.

Conjunto soporte: corresponde al conjunto de posibles valores de una variable aleatoria, dicho de otra forma, los valores que esta puede llegar a tomar. Formalmente, tenemos esta definición: Si X es una variable aleatoria discreta con FMP (caso discreto) o FDP (caso absolutamente continuo) $f_X(x)$ definimos el soporte o conjunto soporte de $f_X(x)$ como:

$$\mathfrak{X} = \{x : f_X(x) > 0\}$$

En el caso discreto, debido a la interpretación que se puede hacer de la FMP, podemos escribir $\mathfrak{X} = \{x : f_X(x) = P(X = x) > 0\}$. Continuando con el ejemplo, tenemos que el soporte de la distribución de la dosis génica es $\mathfrak{X} = \{0, 1, 2\}$.

Características generales de las FMP y FDP. Las FMP y las FDP satisfacen las siguientes condiciones:

1) $f_X(x) \geq 0 \forall x \in R$

2) Si X es una variable aleatoria discreta: $\sum_{x \in \mathfrak{X}} f_X(x) = 1$

3) Si X es una variable aleatoria continua $\int_{-\infty}^{\infty} f_X(x) = \int_{\mathfrak{X}} f_X(x) = 1$.

La igualdad $\int_{-\infty}^{\infty} f_X(x) = \int_{\mathfrak{X}} f_X(x)$ se tiene al considerar el hecho de que fuera del conjunto soporte $\mathfrak{X}$ la FDP es nula, e invocar propiedades de las integrales. De nuevo, el lector que requiera detalles sobre cálculo integral puede consultar textos como Stewart [79] y Spivak [80].

Media, valor esperado, esperanza o esperanza matemática de una variable aleatoria: cuando se piensa en el valor medio de una variable se tiene la noción de un número que está entre el valor mínimo y el máximo que esta puede tomar, esto es tal vez lo más general que se puede pensar. Por otro lado, si se considera la frecuencia con que la variable toma valores a través del conjunto soporte, se podría intuir que el valor esperado o media de la variable estará influenciado por la densidad o la masa de estas regiones. Así, teniendo en cuenta los conceptos de FMP y FDP se podría postular que el valor esperado debe combinar el soporte de la variable aleatoria y la FMP para el caso discreto o FDP para el caso continuo. Además, al hablar de media, el lector puede pensar en el estadístico (un estadístico es cualquier función de la muestra):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Que corresponde a la media muestral o empírica y también es un estimador la media, esperanza o esperanza matemática, que es un parámetro poblacional. La siguiente es la definición de media.

Sea X una variable aleatoria, definimos su media, valor esperado, esperanza o esperanza matemática como:

$$E\left[X\right] = \mu_X = \begin{cases} \int_{-\infty}^{\infty} x f_X(x) dx, \text{ si } X \text{ es continua} \\ \sum_{x \in \mathfrak{X}} x f_X(x) = \sum_{x \in \mathfrak{X}} x P(X = x), \text{ si } X \text{ es discreta} \end{cases}$$

Ahora bien, una función de una variable aleatoria es a su vez una variable aleatoria, así, se define la esperanza de una función $g(X)$ de una variable aleatoria como:

$$E\left[g(X)\right] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx, \text{ caso continuo} \\ \sum_{x \in \mathfrak{X}} g(x) f_X(x) = \sum_{x \in \mathfrak{X}} g(x) P(X = x), \text{ caso discreto} \end{cases}$$

Nota 1: la esperanza no siempre existe, sin embargo, aquí no se lidia con variables aleatorias que presenten esta característica.

Continuando con el ejemplo, la media de la dosis génica es:

$$E\left[X\right] = \sum_{x=0}^{2} x f_X(x)$$

$$= 2 \bullet p^2 + 1 \bullet 2p(1-p) + 0 \bullet (1-p)^2$$

$$= 2p^2 + 2p(1-p)$$

$$= 2p(p + 1 - p)$$

$$= 2p$$

Nota 2: en la segunda línea del procedimiento anterior se empleó explícitamente el símbolo "•" para denotar producto, esto para diferenciar el valor de la variable aleatoria y su probabilidad.

Algunas propiedades de la esperanza que son de utilidad son las siguientes: sea X una variable aleatoria cuya esperanza existe y k una constante. Además, sean g y h funciones de valor real tales que $g(X)$ y $h(X)$ son variables aleatorias cuyas esperanzas existen, entonces:

1) $E\left[kX\right] = kE\left[X\right]$

2) $E\left[g(X) + h(X)\right] = E\left[g(X)\right] + E\left[h(X)\right]$

3) $E\left[k\right] = k$

4) Si $P(X \geq k) = 1 \Rightarrow E\left[X\right] \geq k$

5) Si $g(x) > h(x) \forall x \Rightarrow E\left[g(X)\right] > E\left[h(X)\right]$

Las dos primeras propiedades hacen que la esperanza sea catalogada como un operador lineal, porque las constantes pueden salir del operador quedando como

multiplicadores y el operador aplicado a la suma es la suma del operador aplicado a cada sumando, que en este caso, indica que la esperanza de la suma es la suma de las esperanzas.

Algunas de estas propiedades se emplean cuando se deriva la media de la variable respuesta en un modelo estadístico, a esta se le denomina el primer momento del modelo.

Varianza: la media se conoce como una medida de tendencia central, pese a que matemáticamente hablando no es una medida, por lo tanto, puede ser más apropiado hablar de un parámetro de tendencia central. Si ya se cuenta con un parámetro de tendencia central, conviene tener uno de dispersión alrededor del primero, es decir, que resuma información sobre las desviaciones de la variable respecto a la media. Este parámetro se conoce varianza o segundo momento alrededor de la media y se define como la esperanza de la función $(X-\mu_X)^2$, es decir:

$$Var\left[X\right] = E\left[(X-\mu_X)^2\right]$$

Como en el caso de la media, la varianza no siempre existe, pero aquí no se tratarán estos casos. Nótese además que la varianza es no negativa, a continuación, se enuncian otras de sus propiedades.

Sea X una variable aleatoria cuya varianza existe y b una constante entonces:

1) $Var\left[X\right] = E\left[X^2\right] - \mu_X^2$

2) $Var\left[bX\right] = b^2 Var\left[X\right]$

3) $Var\left[X+b\right] = Var\left[X\right]$

4) $Var\left[b\right] = 0$.

La primera propiedad resulta útil porque en muchos casos facilita el cómputo de la varianza. En el ejemplo de la dosis génica la varianza se obtiene como sigue.

En primer lugar, calculamos $E\left[X^2\right]$:

$$E\left[X^2\right] = \sum_{x=0}^{2} x^2 f_X(x)$$

$$2^2 \bullet p^2 + 1^2 \bullet 2p(1-p) + 0^2 \bullet (1-p)^2$$

$$= (2p)^2 + 2p(1-p)$$

$$= 2p(2p+1-p)$$

$$= 2p(p+1)$$

Previamente se calculó μ_X, entonces:

$$\begin{aligned} Var\left[X\right] &= E\left[X^2\right] - \mu_X^2 \\ &= 2p(p+1) - (2p)^2 \\ &= 2p^2 + 2p - 4p^2 \\ &= 2p - 2p^2 \\ &= 2p(1-p) \end{aligned}$$

Como p toma valores en el intervalo $\left[0, 1\right]$, podemos ver que la varianza es no negativa, con valor de cero si $p = 1$ o $p = 0$, además, debido a que la función $p(1-p)$ se maximiza en $p = \frac{1}{2}$, la varianza es máxima cuando la frecuencia del alelo de referencia es $\frac{1}{2}$, lo cual quiere decir que los dos alelos tienen la misma frecuencia y bajo EHW esto implica que los genotipos II, IR, RR se encuentran en frecuencias $\frac{1}{4}$, $\frac{1}{2}$ y $\frac{1}{4}$, respectivamente.

Múltiples variables aleatorias (vectores y matrices aleatorias): es posible definir variables aleatorias multidimensionales, estas pueden arreglarse en vectores o en matrices aleatorias, siendo el primer caso el más común. Ejemplos del uso de matrices aleatorias en genética pueden encontrarse en los artículos de Martínez et al [81, 82, 83], trabajos en los que se postulan modelos bayesianos que consideran matrices aleatorias en selección genómica y estimación de composición racial, empleando marcadores moleculares.

Así pues, podemos definir vectores aleatorios observables como el que contiene la variable respuesta de un modelo animal unicaracter observada en n individuos. En algunos casos no importa si el vector es fila o columna, pero a la hora de escribir los modelos en forma matricial se suele definir como un vector columna, por lo tanto:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

También se pueden definir la FDA, FMP y FDP para múltiples variables aleatorias, los detalles técnicos se omiten.

Un apunte sobre notación: en álgebra matricial (ver Complemento 1) se emplean letras minúsculas para representar vectores y letras mayúsculas para las matrices; por otro lado, en probabilidad, en el caso univariado se usan letras mayúsculas para representar variables aleatorias y letras minúsculas para representar valores realizados de las mismas. Así, al introducir vectores aleatorios algunos autores siguen empleando

letras minúsculas para representarlos, mientras que otros usan letras mayúsculas. Esta situación no representa un problema mientras la notación sea clara.

De regreso al tema de esta sección, al considerar múltiples variables aleatorias aparecen conceptos como distribución conjunta, distribuciones condicionales y distribuciones marginales. Por ejemplo, si X y Z son variables aleatorias absolutamente continuas podemos escribir $f_{(X,Z)}(x,z)$ (densidad conjunta), $f_{X|Z}(x|z), f_{Z|X}(z|x)$ (densidades condicionales) y $f_X(x), f_Z(z)$ (densidades marginales). En el caso de tres o más variables aleatorias se puede hablar de distribuciones conjuntas, marginales univariadas, marginales conjuntas, marginales conjuntas condicionales y marginales univariadas condicionales con sus respectivas FDP o FMP según el caso. La regla del producto puede extenderse al caso de una FDA o FMP, lo que a su vez permite escribir $f_{(X|Z)}(x|z)$ y $f_{(Z|X)}(z|x)$, así:

$$f_{(X|Z)}(x|z) = \frac{f_{(X,Z)}(x,z)}{f_Z(z)}$$

$$f_{(Z|X)}(z|x) = \frac{f_{(X,Z)}(x,z)}{f_X(x)}$$

Consideremos el caso de tres variables aleatorias absolutamente continuas X, Z, y W, los siguientes son algunos ejemplos:

1) Densidad conjunta: $f(X,Z,W)(x,z,w)$

2) Densidades marginales univariadas: $f_X(x)$, $f_Z(z)$ y $f_W(w)$

3) Densidades marginales conjuntas: $f_{X,Z}(x,z)$, $f_{X,W}(x,w)$ y $f_{Z,W}(z,w)$

4) Densidades marginales univariadas condicionales completas: $f_{X|Z,W}(x|z,w)$, $f_{Z|X,W}(z|x,w)$ y $f_{W|X,Z}(w|x,z)$

5) Densidades marginales univariadas condicionales incompletas (no se incluyen todos los casos posibles): $f_{X|Z}(x|z)$, $f_{X|W}(x|w)$, $f_{W|Z}(w|z)$ y $f_{Z|X}(z|x)$

6) Densidades marginales conjuntas condicionales: $f_{X,W|Z}(x,w|z)$, $f_{X,Z|W}(x,z|w)$, $f_{Z,W|X}(z,w|x)$.

Como puede notarse, las posibilidades aumentan con el número de variables consideradas.

Independencia de variables aleatorias: dos variables aleatorias son independientes si $\forall x \in R$, $\forall z \in R$, los eventos $X \leq x$ y $Z \leq z$ son independientes. Esto implica que la FDA conjunta $F_{X,Z}(x,z)$ puede escribirse así:

$$F_{X,Z}(x,z) = F_X(x)F_Z(z)$$

Esto es, la FDA conjunta es igual al producto de las marginales. Además, si X y Z son independientes, la PMF o PDF conjunta también corresponde al producto de las marginales, es decir:

$$f_{X,Z}(x,z) = f_X(x)f_Z(z) \forall x \in R, \forall z \in R$$

Similar al caso de probabilidades de eventos, bajo independencia, se esperaría que la densidad condicional de X dado Z sea igual a la marginal de X y que la densidad condicional de Z dado X sea igual a la marginal de Z. Esto se establece fácilmente procediendo como sigue:

$$\begin{aligned} f_{X|Z}(x|z) &= \frac{f_{X,Z}(x,z)}{f_Z(z)} \\ &= \frac{f_X(x)f_Z(z)}{f_Z(z)} \\ &= f_X(x) \end{aligned}$$

De manera similar $f_{Z|X}(z|x) = f_Z(z)$. Además, bajo independencia, también se tiene la siguiente propiedad $E\left[XZ\right] = E\left[X\right]E\left[Z\right] = \mu_X\mu_Z$.

Tal y como se escribió la regla del producto en términos de FDP y FMP, el siguiente es el Teorema de bayes para variables aleatorias continuas y discretas, respectivamente:

$$f_{X|Z}(x|z) = \frac{f_{Z|X}(z|x)f_X(x)}{\int_{\boldsymbol{x}} f_{Z|X}(z|x)f_X(x)dx}$$

$$f_{X|Z}(x|z) = \frac{f_{Z|X}(z|x)f_X(x)}{\sum_{x\in\boldsymbol{x}} f_{Z|X}(z|x)f_X(x)}$$

Cuando tenemos p variables aleatorias Y_1, Y_2, ... y Y_p, arregladas en un vector aleatorio Y, podemos definir el vector p-dimensional $E\left[Y\right]$, así:

$$E\left[Y\right] = \begin{bmatrix} E\left[Y_1\right] \\ E\left[Y_2\right] \\ \vdots \\ E\left[Y_p\right] \end{bmatrix}$$

Es decir, un vector cuyas entradas corresponden a las esperanzas de variables aleatorias unidimensionales.

Ahora que se ha introducido el concepto de variables aleatorias multidimensionales, se puede abordar un tema de gran importancia en estadística y, en particular, en el campo de la mejora genética animal. En genética cuantitativa, es bien sabido que cuando se hace selección para un fenotipo dado, se generan cambios en la media de otros que no hacían parte del programa de mejoramiento genético, en algunos casos estos cambian en la dirección deseada y por ende, se tiene una ganancia extra con la que, a primera vista, no se contaba, no obstante, también se tienen escenarios en los que el cambio ocurre en la dirección no deseada y se tiene un efecto deletéreo con el que, de nuevo, al parecer no se contaba; esto se debe a un fenómeno conocido como respuesta correlacionada, que se explica por la existencia de correlación genética entre estos rasgos, esta correlación se presenta si existe covarianza genética. Ahora bien, en la presentación del fenómeno de respuesta correlacionada se habló de ganancia o pérdida con las que "al parecer no se contaba", esto es debido a que un programa de mejoramiento genético bien diseñado debe basarse en el conocimiento de los parámetros genéticos para un grupo de fenotipos relevantes en el sistema productivo y entre estos, se encuentran las correlaciones genéticas, las cuales permitirán predecir la respuesta a la selección, no solo para algunos fenotipos en los que se enfoque el programa, sino para todos aquellos que impactan la explotación, esto permite saber de antemano que rasgos se verán afectados negativamente y establecer así el programa de selección apropiado.

Volviendo a la teoría básica de probabilidad, vamos a presentar los conceptos de *covarianza* y *correlación*. La covarianza es un parámetro que informa sobre la asociación lineal entre dos variables aleatorias, es un indicador de variación conjunta. Sean X y Z dos variables aleatorias, definimos su covarianza (segundo momento producto respecto a la media) como:

$$Cov\left[X,Z\right] = E\left[(X - E\left[X\right])(Z - E\left[Z\right])\right]$$

$$= E\left[(X - \mu_X)(Z - \mu_Z)\right]$$

Sean X,Y, Z y W variables aleatorias y a, b, c y k, constantes; las siguientes son algunas propiedades de la covarianza:

1) $Cov\left[X,Z\right] = Cov\left[Z,X\right]$ (simetría)

2) $Cov\left[X,Z\right] = E\left[XZ\right] - \mu_X\mu_Z$

3) $Cov\left[X,X\right] = Var\left[X\right]$

4) $Cov\left[X,k\right] = 0$

5) $Cov\left[bX,kZ\right] = bkCov\left[X,Z\right]$

6) $Cov\left[X+b,Z+k\right] = Cov\left[X,Z\right]$

7) $Cov\left[aX+bY,cZ+kW\right] = acCov\left[X,Z\right] + akCov\left[X,W\right] + bcCov\left[Y,Z\right] + +bkCov\left[Y,W\right]$ (bilinealidad).

Con el concepto de covarianza se puede introducir otra propiedad de la varianza:

$$Var\left[X+Z\right] = Var\left[X\right] + Var\left[Z\right] + 2Cov\left[X,Z\right]$$

Similarmente,

$$Var\left[X-Z\right] = Var\left[X\right] + Var\left[Z\right] - 2Cov\left[X,Z\right]$$

Llegado a este punto, el lector puede preguntarse sobre el comportamiento de la covarianza cuando las variables aleatorias consideradas son independientes, a continuación, se discute este escenario.

Independencia y covarianza: denotamos la independencia de las variables aleatorias X y Z como $X \perp Z$. Si X y Z son variables aleatorias independientes entonces:

$$\begin{aligned} Cov\left[X,Z\right] &= E\left[XZ\right] - \mu_X\mu_Z \\ &= E\left[X\right]E\left[Z\right] - \mu_X\mu_Z (\because X \perp Z) \\ &= \mu_X\mu_Z - \mu_X\mu_Z = 0 \end{aligned}$$

Podemos escribir este resultado como $X \perp Z \Rightarrow Cov\left[X,Z\right] = 0$. Esto es: independencia probabilística implica covarianza nula, ¿qué ocurre en el otro sentido?, en general, no es cierto que covarianza nula implique independencia, es decir:

$$Cov\left[X,Z\right] = 0 \nRightarrow X \perp Z$$

Hay que recordar que la covarianza informa sobre asociación lineal, así, puede haber casos de variables que tienen una asociación no lineal que las hace dependientes y cuya covarianza es nula. No obstante, existe una condición bajo la cual la covarianza nula es equivalente a independencia. Si X y Z siguen una distribución normal (Gaussiana) bivariada, entonces:

$$Cov\left[X,Z\right] = 0 \iff X \perp Z$$

La varianza se expresa en unidades al cuadrado, la covarianza por su parte se expresa en el producto de las unidades en las que se expresan las dos variables implicadas. En muchos casos no es conveniente que el parámetro de interés dependa de las unidades, por ello, resulta ventajoso emplear el coeficiente de variación (la razón

entre la desviación estándar y la media) para expresar variabilidad puesto que no tiene unidades y su escala no varía al cambiar las unidades. Existe un parámetro ligado a la covarianza que también informa sobre asociación lineal entre pares de variables, pero que no depende de las unidades, de hecho, siempre toma valores entre -1 y 1, se trata del coeficiente de correlación de Pearson, que se define así:

$$r_{(X,Z)} = \frac{Cov\left[X,Z\right]}{\sqrt{Var\left[X\right]Var\left[Z\right]}}$$
$$Var\left[X\right] > 0, Var\left[Z\right] > 0$$

Nótese que:

1) La covarianza le da el signo a la correlación (porque el denominador siempre es positivo)

2) Si la covarianza es nula, entonces la correlación es nula y viceversa, entonces podemos escribir $Cov\left[X,Z\right] = 0 \iff r_{(X,Z)} = 0$.

Al generalizar al caso de p variables aleatorias, se necesita hacer un arreglo de las covarianzas entre cada uno de los $\binom{p}{2}$ (p combinado 2) pares de variables aleatorias, además, considerando que la covarianza de una variable aleatoria consigo misma es igual a su varianza, también conviene incluir las p varianzas correspondientes. Este objetivo se logra arreglando estos parámetros en una matriz cuadrada conocida como la matriz de covarianzas y denotada mediante la letra griega Σ. En esta, tanto filas como columnas, corresponden a cada una de las variables aleatorias consideradas, así, tenemos una matriz de dimensión pxp que en la primera entrada de la fila uno contiene la varianza de Y_1 y en las demás entradas de esta fila contiene las covarianzas de Y_1 con $Y_2, Y_3, \ldots$ yY_p, al proceder de forma análoga con las demás filas se construye una matriz con la siguiente forma:

$$\Sigma := \begin{pmatrix} Var\left[Y_1\right] & Cov\left[Y_1,Y_2\right] & \ldots & Cov\left[Y_1,Y_p\right] \\ Cov\left[Y_2,Y_1\right] & Var\left[Y_2\right] & \ldots & Cov\left[Y_2,Y_p\right] \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left[Y_p,Y_1\right] & Cov\left[Y_p,Y_2\right] & \ldots & Var\left[Y_p\right] \end{pmatrix}$$

$$:= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_p^2 \end{pmatrix}$$

Además, por la propiedad de simetría de la covarianza, es decir,

$$Cov\left[Y_i,Y_j\right] = Cov\left[Y_j,Y_i\right], \forall i, \forall j, i = 1, 2 \ldots, p, j = 1, 2, \ldots, p, i \neq j$$

La matriz Σ es simétrica y podemos escribirla así:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ & \sigma_2^2 & \dots & \sigma_{2p} \\ Sim & & \ddots & \vdots \\ & & & \sigma_p^2 \end{pmatrix}$$

Esta matriz es definida no negativa por construcción, es decir, $x^T \Sigma x \geq 0 \forall x \neq 0$, lo cual es análogo a que la varianza de una variable aleatoria unidimensional sea no negativa y así, como en este caso, lo usual es enfocarse en varianza positiva (es decir, se descarta que sea nula). En muchas aplicaciones resulta de interés una matriz de covarianzas definida positiva, es decir: $x^T \Sigma x > 0 \forall x \neq 0$.

De manera similar, podemos arreglar las correlaciones en una matriz así:

$$C = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ & 1 & \dots & r_{2p} \\ & & \ddots & \vdots \\ & sim & & 1 \end{pmatrix}$$

Además, la matriz de correlación y la matriz de covarianzas guardan la siguiente relación:

$$C = \left(diag \left[\Sigma \right]^{\frac{-1}{2}} \right) \Sigma \left(diag \left[\Sigma \right]^{\frac{-1}{2}} \right)$$

Donde $diag \left[\Sigma \right]^{\frac{-1}{2}}$ es una matriz diagonal, cuyos elementos de la diagonal principal son las raíces cuadradas del inverso multiplicativo de los elementos diagonales de la matriz de covarianzas (es decir, la raíz del inverso multiplicativo de las varianzas), esto es:

$$diag \left[\Sigma \right]^{-\frac{1}{2}} = \begin{pmatrix} (\sigma_1^2)^{-\frac{1}{2}} & 0 & \dots & 0 \\ & (\sigma_2^2)^{-\frac{1}{2}} & \dots & 0 \\ Sim & & \ddots & (\sigma_p^2)^{-\frac{1}{2}} \end{pmatrix}$$

10 CAPÍTULO DIEZ

Breve introducción a la estadística bayesiana

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

10.1. Interpretaciones de probabilidad

En el Capítulo 9 se presenta una definición axiomática de probabilidad. Los axiomas de Kolmogorov nos permiten dar un tratamiento matemático a las probabilidades y crear una gran cantidad de teoría que conduce a muchas aplicaciones de utilidad. Sin embargo, en este capítulo nos enfocaremos en una de las dos interpretaciones de probabilidad, la cual genera toda una escuela de pensamiento conocida como estadística bayesiana porque el Teorema de bayes juega un papel central.

Históricamente, la probabilidad ha estado ligada a la evaluación de la incertidumbre, como en los juegos de azar. Una interpretación de la probabilidad de ocurrencia del evento A, es su frecuencia relativa, es decir, la razón entre el número de veces que ocurrió y el total de experimentos, esto implica que el experimento aleatorio se llevó a cabo varias veces. Por otro lado, se tiene la interpretación bajo la cual la probabilidad es subjetiva y representa el grado de incertidumbre y las creencias sobre el evento. Así, cuando se dice que la probabilidad de cara de una moneda es 0.5, la interpretación frecuentista es que al lanzar la moneda muchas veces, en la mitad de los lanzamientos se obtendrían caras.

Por otro lado, si la persona apuesta un centavo a que sale cara, la interpretación subjetiva le diría que si sale cara gana un centavo, mientras que si sale sello pierde un centavo, se tiene un juego justo. Consideremos ahora la probabilidad de que llueva mañana, asignar una probabilidad de 0.98 indica que se tiene mucha certeza sobre la presentación de lluvia (deberíamos llevar un paraguas), en tanto que, asignar probabilidad de 0.02 indica alta certeza sobre la no ocurrencia del evento. Finalmente, si asignamos una probabilidad de 0.5 (o valores cercanos), se tiene una alta incertidumbre.

Existen escenarios en los que la interpretación subjetiva es más atractiva o lógica. Considere la probabilidad de que un perro particular pase una prueba de trabajo en una ocasión específica; aquí no tiene sentido invocar un escenario abstracto en el que el perro toma la prueba muchas veces bajo condiciones similares o interpretar la probabilidad como la frecuencia de perros que pasan la prueba, se tiene interés en el resultado de la prueba de ese perro en ese día. La aproximación bayesiana a la estadística se basa en la interpretación subjetiva. Es importante recordar que, por fortuna, las probabilidades se tratan de la misma manera bajo cualquiera de sus interpretaciones.

10.2. Generalidades

El bayesiano siempre trata los parámetros desconocidos del modelo como variables aleatorias, esto le permite expresar el grado de incertidumbre sobre los mismos, por lo tanto, en inferencia bayesiana se asigna una distribución de probabilidad a los parámetros del modelo.

En adelante, denotaremos por θ al parámetro desconocido p-dimensional sobre el cual se quieren obtener inferencias, X es un vector aleatorio n-dimensional que contiene la variable respuesta, mientras que x los datos observados, es decir, una realización de X.

El procedimiento básico es atractivo porque sigue un razonamiento lógico. Se inicia con algún grado de conocimiento (puede ser nulo) sobre los parámetros de interés; y se postula un modelo que relaciona los datos con estos parámetros, entonces, al observar los datos tenemos información sobre los parámetros que se puede usar para actualizar el conocimiento sobre los mismos. Surge la pregunta acerca de la forma de combinar el conocimiento previo (aún si es nulo) y los datos. Autores como Cox [84, 85] y Savage [86] mostraron que una forma óptima de actualizar el conocimiento sobre un parámetro es empleando el Teorema de bayes, de allí el nombre de esta escuela.

Ahora bien, tratar los parámetros de un modelo estadístico como variables aleatorias no es algo nuevo en mejoramiento genético, pues los valores de cría se consideran variables aleatorias y existe una teoría para formular una distribución normal multivariada (lo que sustenta el modelo animal). En el Capítulo 3 se presentan

los modelos lineales mixtos desde una perspectiva frecuentista, bajo la cual el modelo recibe esta denominación porque tiene efectos fijos y aleatorios, entonces tenemos un juego de parámetros que se tratan como constantes desconocidas (fijos) y otros como variables aleatorias (aleatorios). En el caso bayesiano, todos los parámetros son aleatorios y por ende hablar de modelos mixtos es un poco difuso.

Otro ejemplo que motiva el uso de un marco de trabajo bayesiano en genética es la estimación de frecuencias alélicas. Wright [87] derivó los principios que indican que estas frecuencias presentan variación aleatoria, lo que justifica el uso de un modelo bayesiano para estimarlas. Entonces, se les asigna una distribución de probabilidad en lugar de tratarlas como constantes desconocidas. Las ventajas de la aproximación bayesiana son:

1) Permite el análisis rutinario de datos complejos

2) Se ajusta bien al análisis de datos de gran dimensión

3) Interpretación adecuada de intervalos (conjuntos) de confianza

4) Permite manejar la incertidumbre de una forma estructurada y sencilla, por ello algunos acuñan la expresión “la incertidumbre es gratis”.

Ejemplos en los que un abordaje bayesiano permite tener en cuenta la incertidumbre con relativa facilidad pueden encontrarse en Martínez et al [82, 83]. En el caso de Martínez et al [82] desarrollaron un modelo bayesiano jerárquico (término que se precisa más adelante en este capítulo) para hacer predicción genómica (Capítulo 7) en varias poblaciones. El abordaje bayesiano permitió considerar marcadores con genotipos perdidos haciendo la imputación sobre la marcha (de manera automática) y teniendo en cuenta la incertidumbre asociada de una forma relativamente simple. Por su parte, Martínez et al [83] propusieron un modelo bayesiano que permitió tener en cuenta la incertidumbre sobre las estimaciones de frecuencias alélicas en la inferencia de la composición racial de animales híbridos empleando marcadores moleculares.

Otra de las ventajas del uso de una aproximación bayesiana puede encontrarse en Martínez et al [88], quienes empleando la teoría de la decisión bayesiana (tema que se discute en la sección 10.9), derivaron estimadores alternativos de las frecuencias alélicas que son siempre más precisos que el estimador usual que suele estudiarse en genética básica y que bajo ciertas condiciones es el estimador máximo-verosímil.

Ahora nos enfocamos en un resultado de gran importancia en estadística bayesiana conocido como el Teorema de de Finetti, pero antes de presentarlo, se introducen conceptos necesarios para su comprensión:

Definición. La secuencia de variables aleatorias $X_1, X_2, \dots, X_n$ es finitamente intercambiable si para cada una de las n! permutaciones posibles $(X_{K_1}, X_{K_2}, \dots, X_{K_n})$ se tiene :

$$(X_{K_1}, X_{K_2}, \dots, X_{K_n}) \overset{d}{=} (X_1, X_2, \dots, X_n)$$

Donde, $\overset{d}{=}$ significa igualdad en distribución. Es decir, el orden de las variables aleatorias no altera su distribución conjunta.

Definición. La secuencia de variables aleatorias $X_1, X_2, \ldots, X_n, \ldots$ es infinitamente intercambiable si cada secuencia finita $(X_1, X_2, \ldots, X_m)$, es intercambiable, $m \geq 1$.

Las variables idéntica e independientemente distribuidas (iid) son intercambiables, pero la intercambiabilidad no implica independencia, como lo indica el siguiente ejemplo. Considere la FMP conjunta:

$$f(x_1, x_2) = \frac{e^{\theta(x_1+x_2)+\phi x_1 x_2}}{1 + 2e^{\theta} + e^{2\theta+\phi}},$$

$$x_1 \in \{0, 1\}, x_2 \in \{0, 1\}$$

Las variables aleatorias X_1 y X_2 son intercambiables porque la FMP conjunta se mantiene invariante a su orden y lo mismo ocurrirá con la FDA, pero no son independientes porque su FMP conjunta no puede escribirse como el producto de dos funciones, una que depende solo de X_1 y otra que depende solo de X_2.

Teorema de de Finetti: si $X_1, X_2, \ldots, X_n, \ldots$ es una secuencia de variables aleatorias infinitamente intercambiables, entonces, existe una variable aleatoria θ con función de distribución Π y soporte Ω, tal que condicionando en $\theta, X_1, X_2, \ldots, X_m$ son iid $\forall m \geq 1$. El resultado también se tiene en la otra dirección, esta doble implicación corresponde al Teorema de de Finetti. Ahora bien, otra forma de ver el Teorema es la siguiente:

$$X_1, X_2, \ldots, X_m | \theta \text{ son } iid \text{ y } \theta \sim \Pi(\theta) \iff X_1, X_2, \ldots, X_m \text{ son intercambiables } \forall m \geq 1.$$

Por lo tanto, si $X_1, X_2, \ldots, X_m, \ldots$ son infinitamente intercambiables, entonces, $\forall m \geq 1$:

$$\int_{\Omega} \left(\prod_{i=1}^{m} f(x_i|\theta) \right) \pi(\theta) d\theta = f(X_1, X_2, \ldots, X_m)$$

¿Qué nos dice esto? Que, al asumir intercambiabilidad, existe una variable aleatoria θ que permite construir un modelo en el que se pueden asumir variables iid condicionadas en θ, dicho de otra manera, el asumir intercambiabilidad de las variables observables nos lleva al planteamiento de una verosimilitud iid y la existencia de la distribución a priori.

Llegado este punto se hace un apunte sobre notación. En el Capítulo 9 se empleaban subíndices en la letra f para denotar la variable aleatoria y distribución a la que correspondía esta FDP o FMP, por ejemplo, $f_X(x)$ o $f_Y(y)$. Para facilitar la notación, en este capítulo se optará por una notación menos explícita, pero equivalente y clara, así, para X e Y variables aleatorias tenemos:

$$\begin{aligned} f(x) &\coloneqq f_X(x) \\ f(y) &\coloneqq f_Y(y) \\ f(x,y) &\coloneqq f_{X,Y}(x,y) \\ f(x|y) &\coloneqq f_{X|Y}(x|y) \\ f(y|x) &\coloneqq f_{Y|X}(y|x) \end{aligned}$$

Similarmente, se emplea la letra F para denotar la FDA, y el argumento de la función indica la variable aleatoria y la naturaleza de la distribución, por ejemplo, si X es una variable aleatoria con distribución normal y Z es una variable aleatoria con distribución Poisson, $F(x)$ es la FDA de una distribución normal mientras que $F(z)$ es la FDA de una distribución Poisson.

Continuando con la filosofía bayesiana, definimos la distribución de los parámetros antes de observar los datos como la distribución a priori o previa $\Pi(\theta)$, esta representa el conocimiento preliminar que, como se ha enfatizado varias veces, puede ser nulo. Cuando se tiene una variable aleatoria absolutamente continua o una variable discreta, la respectiva FDP o FMP se denota como $\pi(\theta)$, al soporte de esta distribución se le conoce como espacio paramétrico y se simboliza con la letra griega Ω.

La función de verosimilitud o modelo de muestreo es la distribución conjunta de los datos dado θ y se nota como $f(\mathbf{x}|\theta)$. Utilizamos la regla o Teorema de bayes para "actualizar" nuestro conocimiento sobre θ que se expresa mediante la distribución posterior, esta corresponde a la distribución de θ dados los datos y se escribe como $\Pi(\theta|\mathbf{x})$, con FDP (caso absolutamente continuo) o FMP (caso discreto) $\pi(\theta|\mathbf{x})$. Por lo tanto, para una variable aleatoria absolutamente continua se tiene:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int\limits_{\Omega} f(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

La distribución marginal $m(\mathbf{x}) = \int\limits_{\Omega} f(\mathbf{x}|\theta)\pi(\theta)d\theta$ no depende de θ porque se está integrando respecto a esta variable, por lo tanto, podemos escribir:

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

Donde, el símbolo "$\propto$" significa proporcional, entonces, la ecuación anterior se lee "la posterior es proporcional al producto de la verosimilitud y la a priori", estos son tres elementos centrales en la estadística bayesiana. Cabe notar que, al ser la verosimilitud la densidad conjunta de los datos dados los parámetros vista como función de los parámetros, existe notación que enfatiza este hecho, por ejemplo, podemos escribir:

$$L(\theta; \mathbf{x}) \coloneqq f(\mathbf{x}|\theta)$$

Antes de continuar con el desarrollo del tema principal de este capítulo, se discute una propiedad de las distribuciones que resulta de utilidad para reconocer la naturaleza de la distribución posterior.

Sea $f(x)$ una FDP con soporte $\mathfrak{X}$, sabemos que $\int_{\mathfrak{X}} f(x)dx = 1$, esto nos permite ver que una función integrable y no negativa g, puede convertirse fácilmente en una densidad puesto que:

$$\int_{\mathfrak{X}} g(x)dx = k, k > 0$$

Por lo tanto, podemos crear una función que satisface los requerimientos para ser FDP (Capítulo 9), así:

$$f(x) = k^{-1}g(x).$$

Veamos que en efecto $f(x)$ integra a 1:

$$\int_{\mathfrak{X}} f(x)dx = \int_{\mathfrak{X}} k^{-1}g(x)dx$$

$$= k^{-1}\int_{\mathfrak{X}} g(x)dx$$

$$k^{-1}k = 1$$

Además, como $k > 0$, y g es no negativa, f es no negativa; en consecuencia, tenemos una función de densidad. A k^{-1} se le denomina constante normalizadora, puesto que permite que la densidad integre a 1. Por otro lado, la función g se conoce como núcleo. De esta manera, ciertas integrales complejas pueden resolverse fácilmente si el integrando corresponde al núcleo de una densidad y se está integrando sobre el conjunto soporte. En el caso discreto, se puede seguir un procedimiento análogo que muestra la misma descomposición para la FMP.

En el EJEMPLO 10.2 se presentan dos casos, el de una variable aleatoria continua con distribución normal y otro de una variable aleatoria discreta con distribución bernoulli.

Si $X \sim N(\mu, \sigma^2)$, su densidad es como sigue:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, -\infty < x < \infty$$

Entonces, $k^{-1} = \frac{1}{\sqrt{2\pi\sigma^2}}$, la parte de la densidad que no depende de la variable aleatoria, es decir, la constante normalizadora, mientras que $g(x) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, por lo tanto, como la densidad integra a uno (probarlo para esta densidad particular va más allá del ámbito de este libro), sabemos que:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \sqrt{2\pi\sigma^2}$$

Al integrar el núcleo sobre el soporte de la variable, se obtiene el inverso multiplicativo de la constante normalizadora.

Ahora, sea H una variable aleatoria con distribución bernoulli con parámetro p, es decir, $H \sim \text{Bernoulli}(p)$, entonces su FMP es de la forma:

$$f(h) = p^h(1-p)^{1-h}, h \in \mathcal{H} = \{0, 1\}$$

Al sumar sobre los posibles valores de h, se obtiene:

$$\sum_{h \in \mathcal{H}} p^h(1-p)^{1-h} = p^1(1-p)^{1-1} + p^0(1-p)^{1-0}$$

$$= p + 1 - p = 1$$

En este caso $k^{-1} = 1, g(p) = p^h(1-p)^{1-h}$.

Vale la pena recalcar que el reconocimiento del núcleo resulta bastante útil para identificar una distribución, en estadística bayesiana se obtiene el producto de la verosimilitud y la densidad a priori para obtener la posterior no normalizada, es decir, la posterior sin la constante normalizadora, esto en vista de que el denominador $m(x)$ en el Teorema de bayes no se considera, así, reconociendo el núcleo, se puede establecer si la posterior pertenece a alguna de las familias conocidas o si se trata de una distribución "no estándar".

Debido al razonamiento anterior, en los procedimientos algebraicos para encontrar la distribución posterior suelen omitirse las constantes normalizadoras de la a priori y la función de verosimilitud. El truco de enfocarse solo en el núcleo de la densidad para reconocer la distribución y así evitar la manipulación de las constantes normalizadoras se ilustrará más adelante.

10.3. Distribuciones a priori

La siguiente es terminología relacionada a la distribución a priori.

Definición: el proceso de definición de la distribución a priori se conoce como elicitación de la a priori.

Definición: los parámetros de la distribución a priori se conocen como hiperparámetros. Una vez se ha definido la familia a la que pertenece la distribución a priori, por ejemplo, la normal o la beta, el proceso de definición de sus parámetros se conoce como elicitación de hiperparámetros.

Una ventaja que tiene la inferencia bayesiana es que permite usar funciones que no corresponden a densidades propiamente dichas como densidades a priori. Estas se conocen como a priori impropias y se definen así.

Definición: la distribución a priori se dice impropia si una función no negativa $\pi(\theta)$ que satisface $\int_{R^p} \pi(\theta)d\theta = \infty$ juega el papel de FDP. En el caso discreto tendríamos una función no negativa cuya sumatoria sobre el espacio paramétrico es igual a infinito y que hace las veces de FMP. Por lo tanto, una a priori impropia corresponde a una función que se emplea como instrumento de cálculo, pero que no corresponde a una distribución. Cuando se emplean este tipo de a priori, se debe chequear que la posterior sea propia, este es el riesgo que se corre. En muchos casos estas funciones se usan por conveniencia, por ejemplo, por simplicidad o para asegurar alguna propiedad de la posterior, un ejemplo puntual se presenta en el Capítulo 7 al discutir un modelo denominado Lasso bayesiano.

Definición: una distribución a priori se dice subjetiva cuando esta se define a partir de información preliminar que se tiene sobre el parámetro de interés. Esta información puede tener varios orígenes como datos recolectados previamente, literatura, conocimiento o experiencia del equipo de trabajo, o las creencias del investigador sobre el fenómeno o sistema estudiado. Existen varias técnicas para poder representar esta información mediante una distribución de probabilidad.

Un ejemplo del uso de información disponible en la literatura para elicitar los hiperparámetros y definir así una distribución a priori subjetiva puede encontrarse en Martínez et al [83]. Allí, se emplearon las estimaciones de frecuencias alélicas para las razas Brahman y Angus en Estados Unidos para definir la distribución a priori de estos parámetros. Cabe destacar que se debe ser cuidadoso con el uso de información disponible en la literatura, lo ideal es que esta haya sido obtenida bajo condiciones similares a aquellas en las que se realizará el estudio o se obtuvieron los datos, de manera que tenga sentido usarla. Como ilustración, en el artículo de Martínez et al [83] se trabajó con una población multirracial Angus-Brahman en Estados Unidos y el estudio que sirvió de referencia para definir la a priori utilizó genotipos de animales de estas razas en este país entre los que se contaba con reproductores que tenían buena representatividad según su número de progenies.

Definición: la distribución a priori se conoce como objetiva, no subjetiva o predeterminada cuando esta se define a partir de algún principio general y no a través de información preliminar, por ejemplo, la a priori de referencia de Bernardo corresponde a aquella que, para una verosimilitud dada, máxima la divergencia de Kullback-Leibler entre a la priori y la posterior. La divergencia de Kullback-Leibler cuantifica que tan diferentes son dos distribuciones y cumple la propiedad de que la divergencia entre una distribución con ella misma es cero. Para mayores detalles ver Ghosh [89].

Los primeros usos de este tipo de distribuciones a priori no subjetivas se atribuyen a fundadores de la estadística bayesiana como Bayes y Laplace, quienes sugirieron usar la distribución uniforme para inferir una proporción Binomial [89].

Otro ejemplo de este tipo de distribuciones a priori es la a priori de Jeffreys, a la cual se puede llegar bajo diferentes principios matemáticos, uno de ellos es invarianza a transformaciones invertibles del parámetro. En este caso la densidad a priori es de la forma:

$$\pi_J \propto |I(\theta)|^{\frac{1}{2}}$$

Donde $I(\theta)$ es la matriz de información de Fisher. Entonces, la densidad es proporcional a la raíz cuadrada del determinante de la matriz de información de Fisher.

EJEMPLO 10.3 En el caso de una verosimilitud Bernoulli(θ) con una muestra de tamaño n obtenida de forma independiente la información es $\frac{n}{\theta(1-\theta)}$, por lo tanto, la densidad a priori de Jeffreys satisface $\pi_J(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, esto es, una Beta($\frac{1}{2}, \frac{1}{2}$).

Definición: una distribución a priori que expresa información puntual o muy específica sobre el parámetro se conoce como informativa; por ejemplo, cuando se elige una distribución con una media y una varianza que reflejan información o creencias sobre el parámetro. Cuando se tienen reportes preliminares sobre valores estimados del parámetro y su dispersión, esta información se puede incluir a través de una distribución a priori cuya media y varianza coinciden con estos reportes. Por otro lado, cuando la a priori expresa información muy vaga o general sobre el parámetro, se conoce no informativa, plana o difusa.

Definición: sea P_δ una familia de distribuciones indexada por el parámetro δ, $L(\theta; x)$ una función de verosimilitud y $\Pi(\theta) \in P_\delta$, una distribución a priori $\Pi(\theta)$, se dice conjugada para $L(\theta; x)$, si la distribución posterior correspondiente $\Pi(\theta|x)$ también pertenece a P_δ.

La búsqueda de una a priori conjugada puede servir como principio para definir una a priori objetiva. Para muchas verosimilitudes se conocen las a priori que son conjugadas, por ejemplo, la distribución Beta es conjugada para la verosimilitud Binomial o su caso especial con $n = 1$, la bernoulli. Para distribuciones que hacen parte de un grupo llamado familia exponencial, se cuenta con un resultado general que sirve para encontrar una distribución a priori conjugada que, además, goza de otras

propiedades, este se conoce como el Teorema de Diaconis-Ylvisaker [90]. Entonces, este Teorema nos brinda un principio para obtener distribuciones a priori objetivas.

Existen relaciones entre las anteriores definiciones y no todas inducen grupos mutuamente excluyentes, por ejemplo, las a priori impropias pueden usarse como no informativas, una distribución a priori objetiva es no informativa mientras que una subjetiva es informativa.

10.4. Ejemplo de la obtención de la distribución posterior

EJEMPLO 10.4: este ejemplo consta de dos partes, en las dos se tienen a priori conjugadas. En la primera parte consideramos una verosimilitud bernoulli y una a priori Beta, entonces, el modelo se escribe así:

$$X_1, X_2, \dots, X_n|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

$$\theta \sim \text{Beta}(\alpha, \beta)$$

Sea $x := (x_1, x_2, \dots, x_n)$ el vector de datos observados. La densidad a priori es:

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \alpha > 0, \beta > 0$$

Mientras que bajo los supuestos de distribución idéntica e independiente la verosimilitud es de la forma:

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{\sum_{i=1}^{n}(1-x_i)}$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Nótese que se escribió solamente el núcleo de la a priori, mientras que en el caso de la verosimilitud la constante normalizadora es 1 y por ende se tiene igualdad en lugar de proporcionalidad. Ahora se obtiene el producto de la a priori y la verosimilitud (separadas por el signo $\times$ en la primera línea) y tras algunas manipulaciones algebraicas sencillas se encuentra el núcleo de la posterior al enfocarnos en las expresiones que contengan a θ.

$$\pi(\theta|\mathbf{x}) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

$$= \theta^{\sum_{i=1}^{n} x_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^{n} x_i + \beta - 1}$$

Se reconoce esta expresión como el núcleo de una Beta $\left(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta\right)$, es decir:

$$\theta|\mathbf{x} \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta\right)$$

Ahora, la segunda parte del ejemplo muestra un caso que requiere mayor manipulación algebraica y usar un truco llamado "completar el cuadrado". Consideramos $X_1, X_2, \ldots, X_n|\theta \overset{iid}{\sim} \text{N}(\theta, \sigma^2)$, nuevamente, sea $\mathbf{x} := (x_1, x_2, \ldots, x_n)$ el vector de datos observados. Asignamos la siguiente distribución a priori:

$$\theta \sim \text{N}(\mu, \tau^2)$$

Donde, los hiperparámetros μ y τ^2 son conocidos. Nótese que se considera un problema en el que tenemos una distribución normal con varianza conocida y se quiere inferir la media θ.

Nuevamente, bajo el supuesto de independencia y distribución idéntica, la verosimilitud se obtiene así:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$\propto exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

$$= exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \theta)^2\right)$$

En la última igualdad se sumó y se restó la media aritmética de la muestra:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Recordemos que nos interesa la media, entonces, se busca simplificar el último exponencial para obtener solo los términos que involucran θ, para ello, manipulamos la sumatoria en el exponente:

$$\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \theta)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - \theta)\sum_{i=1}^{n}(x_i - \overline{x}) + \sum_{i=1}^{n}(\overline{x} - \theta)^2$$

Pero,

$$\sum_{i=1}^{n}(x_i-\overline{x})=\sum_{i=1}^{n}x_i-n\overline{x}$$

$$=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i$$

$$=0$$

Resultado que se obtiene porque $n\overline{x}=\sum_{i=1}^{n}x_i$. De aquí se tiene que:

$$\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\theta)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\theta)^2$$

Este resultado permite particionar el exponencial de interés, así:

$$exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right)$$

$$\propto exp\left(-\frac{n}{(2\sigma^2}(\overline{x}-\theta)^2\right)$$

En cuanto a la a priori, consideramos solo el núcleo y así tenemos:

$$\pi(\theta)\propto exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right)$$

Por lo tanto,

$$\pi(\theta|\mathbf{x})\propto exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right)\times exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right)$$

$$=exp\left(-\frac{1}{2}\left(\frac{n(\overline{x}-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right)\right)$$

Nos enfocamos en el exponente para dejar solamente los términos que involucran a θ. Tras expandir los dos cuadrados y factorizar en términos de θ^2 y θ se tiene:

$$\frac{n(\overline{x}-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}=\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)-2\theta\left(\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)+\frac{n\overline{x}^2}{\sigma^2}+\frac{\mu^2}{\tau^2}$$

Ahora se aplica una técnica que en matemáticas se conoce como “completar el cuadrado”. El razonamiento es el siguiente: nótese que en la anterior expresión se tienen términos que involucran θ^2 y θ, la idea es sumar y restar, dividir y multiplicar

términos, de manera que se llegue a una expresión equivalente en la que se tenga una función cuadrática de la forma $(\theta - k)^2$, con k una expresión que no contenga a θ. Aplicando este procedimiento se obtiene que el exponente de la densidad posterior es:

$$\left(\frac{n}{\sigma^2}+\frac{2}{\tau^2}\right)\left(\theta-\frac{\left(\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2+\frac{n\overline{x}^2}{\sigma^2}+\frac{\mu^2}{\tau^2}-\frac{\left(\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)^2}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$

Al dejar solamente los términos que dependen de θ, se tiene que:

$$\pi(\theta|\mathbf{x}) \propto exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta-\frac{\left(\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right).$$

Para simplificar esta expresión usamos la siguiente identidad:

$$\frac{\left(\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} = (1-B)\overline{x}+B\mu$$

Donde $B = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2} = \frac{\sigma^2}{\sigma^2+\tau^2 n}$, y de aquí se tiene que:

$$\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1} = \frac{\sigma^2}{n}(1-B)$$

Entonces:

$$\pi(\theta|\mathbf{x}) \propto exp\left(-\frac{1}{2}\left(\frac{\sigma^2}{n}(1-B)\right)^{-1}(\theta-\left((1-B)\overline{x}+B\mu\right))^2\right)$$

Por lo tanto, tenemos el núcleo de una distribución normal con media $(1-B)\overline{x}+B\mu$ y varianza $\frac{\sigma^2}{n}(1-B)$, esto es:

$$\theta|\mathbf{x} \sim \mathrm{N}\left((1-B)\overline{x}+B\mu, \frac{\sigma^2}{n}(1-B)\right).$$

10.5. Ejemplo del uso de una a priori impropia

EJEMPLO 10.5: para el problema de inferir el parámetro θ (probabilidad de éxito) de una distribución bernoulli empleando una muestra de tamaño n obtenida de manera independiente (primera parte del EJEMPLO 10.4), vamos a emplear la distribución a priori impropia de Haldane, la cual se obtiene al asignar valores de cero a los dos parámetros de una Beta, es decir tenemos una $"Beta(0,0)"$. Se usan las comillas porque formalmente no es una distribución Beta puesto que sus parámetros deben ser mayores a cero. En virtud del EJEMPLO 10.4, sabemos que:

$$\theta|\mathbf{x} \sim \text{Beta}\left(\sum_{i=1}^{n} x_i, n - \sum_{i=1}^{n} x_i\right)$$

Entonces, debemos hacernos la siguiente pregunta: ¿Bajo qué condiciones se tiene una distribución propia? Nótese que cuando todos los resultados son aciertos, es decir, $x_i = 1 \forall i = 1, 2, \ldots, n$ tenemos que $\sum_{i=1}^{n} x_i = n$ y, por lo tanto, el segundo parámetro será nulo, lo cual hace que la posterior sea impropia y no se puedan hacer inferencias válidas. Por otro lado, cuando todos los resultados son fracasos, $x_i = 0 \forall i = 1, 2, \ldots, n$, se tiene que $\sum_{i=1}^{n} x_i = 0$ y por consiguiente el primer parámetro es nulo, lo que, nuevamente, hace que la distribución sea impropia. En conclusión, cuando se usa la distribución a priori impropia de Haldane, las dos clases posibles de la variable respuesta deben ser observadas en la muestra para que la posterior sea propia y se puedan hacer inferencias válidas.

Hasta el momento, hemos estudiado el uso del Teorema de bayes para actualizar el conocimiento que se tiene sobre los parámetros del modelo una vez se observan los datos y se han ilustrado las manipulaciones algebraicas que se realizan para encontrar la distribución posterior mediante dos ejemplos. Las conclusiones que se obtienen al realizar un análisis bayesiano se basan en la distribución posterior, por lo tanto, esta se puede ver como el instrumento central de inferencia. En inferencia estadística se distinguen dos problemas: prueba de hipótesis y estimación, la estimación puede ser puntual o por intervalo. Debido a la naturaleza de este texto, resultan de importancia los problemas de estimación puntual. Así, dejamos a un lado la prueba de hipótesis y la estimación por intervalo.

10.6. Relación entre la media posterior, la media a priori y la media muestral, y encogimiento hacia la media a priori

Cuando se piensa en un estimador puntual basado en la distribución posterior, algún parámetro de tendencia central parece ser una opción viable, entonces, se puede usar la media (si existe), la mediana o la moda posterior (si la distribución es unimodal) como estimador puntual del parámetro de interés.

En la primera parte del EJEMPLO 10.4 se encontró que la distribución posterior de la probabilidad de éxito de la distribución bernoulli era Beta($\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta$) y, por lo tanto, la media posterior es:

$$E[\theta|\mathbf{x}] = \frac{\sum_{i=1}^{n} x_i + \alpha)}{\sum_{i=1}^{n} x_i + \alpha + n - \sum_{i=1}^{n} x_i + \beta}$$

$$= \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha + \beta + n}$$

Esta expresión puede manipularse para encontrar una combinación convexa de la media a priori $\frac{\alpha}{\alpha+\beta}$ y la media muestral $\overline{x}$, para ello procedemos como sigue:

$$\frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha + \beta + n} = \frac{n\overline{x} + \alpha}{\alpha + \beta + n}$$

$$= \frac{n\overline{x}}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta + n}$$

$$= \frac{n}{\alpha + \beta + n}\overline{x} + \frac{\alpha + \beta}{\alpha + \beta + n}\left(\frac{\alpha}{\alpha + \beta}\right).$$

Cuando n aumenta, el componente que proviene de los datos tiene más peso que el que proviene de la distribución a priori porque sus coeficientes tienden a 1 y 0, respectivamente, esta es una propiedad deseable.

Por otro lado, en la segunda parte del EJEMPLO 10.4, la media posterior tiene la forma:

$$E[\theta|\mathrm{x}] = (1 - B)\overline{x} + B\mu$$

De nuevo tenemos una combinación convexa entre la media muestral y la media a priori y, por ende, la media posterior tomará valores entre $min\{\overline{x}, \mu\}$ y $max\{\overline{x}, \mu\}$ y si $\overline{x} = \mu = \phi$, entonces $E[\theta|\mathrm{x}] = \phi$. Recordemos que:

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2 n} \in (0, 1)$$

Por lo tanto, a medida que el tamaño de muestra crece, B tiende a cero y la media posterior se acerca a $\overline{x}$, es decir, al igual que en la primera parte del EJEMPLO 10.4, los datos tienen más peso en el estimador puntual del parámetro de interés θ, claramente, esta es una propiedad atractiva (a medida que aumenta el tamaño de muestra se da más peso a la información proveniente de los datos).

Continuando con este ejemplo, reescribimos la media posterior para ilustrar otra propiedad conocida como encogimiento:

$$E[\theta|\mathrm{x}] = (1-B)\overline{x} + B\mu$$

$$= \overline{x} - B\overline{x} + B\mu$$

$$= \overline{x} - B(\overline{x} - \mu)$$

Empleando esta expresión podemos encontrar que:

Si $\overline{x} < \mu \Longrightarrow B(\overline{x} - \mu) < 0$ (recordemos que $B \in (0,1)$) y en este caso la media posterior excede a la media muestral por la cantidad $B(\overline{x} - \mu)$, es decir, a la media muestral le sumamos la cantidad $B(\overline{x} - \mu)$ para obtener la media posterior. Entonces partiendo de la media muestral, nos desplazamos hacia la media a priori.

Si $\overline{x} > \mu \Longrightarrow B(\overline{x} - \mu) > 0$ y en este caso la media posterior es inferior a la media muestral por la cantidad $B(\overline{x}-\mu)$, es decir, a la media muestral le sustraemos la cantidad $B(\overline{x} - \mu)$ para obtener la media posterior. Nuevamente, nos desplazamos en dirección a la media a priori.

Si $\overline{x} = \mu \Longrightarrow B(\overline{x} - \mu) = 0 \Longrightarrow E[\theta|\mathrm{x}] = \overline{x} = \mu$, la media muestral, la media a priori y la media posterior son iguales, la media muestral no se desplaza para obtener la media posterior.

En resumen, cuando la media muestral y la media a priori son diferentes, al computar la media posterior, la media muestral se desplaza $B(\overline{x} - \mu)$ unidades en dirección a la media a priori. Es decir, se acerca a la media a priori, a esto se le conoce como encogimiento (*Shrinkage* en inglés), en este caso, encogimiento hacia la media a priori. Este fenómeno se emplea a menudo en la elicitación de hiperparámetros, especialmente en problemas en los que se tiene un gran número de parámetros como en los modelos de selección genómica que se estudiaron en el Capítulo 7. La premisa es que muchos efectos pueden ser nulos y así, se asigna una distribución a priori que genere encogimiento hacia cero.

10.7. Teoría de la decisión bayesiana

Previamente se postularon parámetros de tendencia central de la distribución posterior $\Pi(\theta|\mathrm{x})$ como estimadores puntuales del parámetro θ, típicamente se tiene la media, la mediana y la moda, las cuales, en algunas distribuciones como por ejemplo la normal o la t, son iguales. Probablemente, la media posterior $E[\theta|\mathrm{x}]$ es uno de los estimadores puntuales más usados. Surge entonces una pregunta, ¿existe algún criterio para justificar su uso? Una rama de la estadística conocida como teoría de la decisión, en particular, la teoría de la decisión bayesiana permite establecer la optimalidad de estos estimadores como minimizadores de algunas funciones que se discuten a continuación.

10.7.1. Elementos de la teoría de la decisión

El parámetro de interés es $\theta \in \Omega$, donde Ω es el espacio paramétrico, estamos interesados en encontrar una regla de decisión $\delta(\mathrm{x})$ que sirva como estimador puntual de este parámetro, esta regla es una función de los datos x, que toma valores en D, el espacio de decisión; además, $\mathrm{x} \in \mathfrak{X}$, donde $\mathfrak{X}$ es el soporte de la verosimilitud. Definimos una función de pérdida (o función de error) denotada como $\mathcal{L}(\theta; \delta(\mathrm{x}))$, que es cualquier función no negativa de valor real que satisface $\mathcal{L}(a; a) = 0$.

El riesgo frecuentista o simplemente el riesgo, corresponde a la esperanza de la función de pérdida tomada con respecto a la verosimilitud, esto es:

$$R(\theta, \delta) = \int_{\mathfrak{X}} \mathcal{L}(\theta, \delta(\mathrm{x})) f(\mathrm{x}|\theta) d\mathrm{x}$$

Esta es una esperanza condicional, $E[\mathcal{L}(\theta, \delta(\mathrm{x}))|\theta]$.

Ahora consideramos la esperanza del riesgo frecuentista tomada con respecto a la distribución a priori. Sea $\pi(\theta)$ la densidad a priori, entonces:

$$r(\pi, \delta) = \int_{\Omega} R(\theta, \delta)\pi(\theta) d\theta$$

Reemplazando por la definición de $R(\theta, \delta)$ se obtiene:

$$r(\pi, \delta) = \int_{\Omega} \int_{\mathfrak{X}} \mathcal{L}(\theta, \delta(\mathrm{x})) f(\mathrm{x}|\theta)\pi(\theta) d\mathrm{x} d\theta$$

Esto es:

$$r(\pi, \delta) = E[R(\theta, \delta)]$$

Donde, esta esperanza se calcula respecto a la distribución a priori, pero $R(\theta, \delta)$ es a su vez una esperanza, entonces:

$$r(\pi, \delta) = E[E[\mathcal{L}(\theta, \delta(x))|\theta]].$$

A $r(\pi, \delta)$ se le conoce como el riesgo bayesiano de δ con respecto a π [91], aunque algunos autores como Ghosh [89] también se refieren al mismo como riesgo pre-posterior por el hecho de que la esperanza del riesgo frecuentista se toma con respecto a la distribución a priori.

Recordemos que: $\pi(\theta|\mathrm{x}) = \frac{f(\mathrm{x}|\theta)\pi(\theta)}{m(\mathrm{x})}$, entonces $\pi(\theta|\mathrm{x})m(x) = f(\mathrm{x}|\theta)\pi(\theta)$; y de aquí:

$$r(\pi, \delta) = \int_{\Omega} \int_{\mathfrak{X}} \mathcal{L}(\theta, \delta(x))\, \pi(\theta|\mathrm{x}) m(\mathrm{x}) d\mathrm{x} d\theta$$

$$= \int_{\mathbf{x}} \left\{ \int_{\Omega} \mathcal{L}(\theta, \delta(\mathrm{x}))\pi(\theta|\mathrm{x})d\theta \right\} m(\mathrm{x})d\mathrm{x}.$$

La integral interna es la esperanza posterior de la función de pérdida, esto es $E[\mathcal{L}(\theta, \delta(\mathrm{x})|\mathrm{x}]$, la cual se conoce como el riesgo posterior. Esta forma de expresar el riesgo bayesiano resulta útil porque nos indica que la regla de decisión o estimador que minimiza el riesgo posterior, también minimiza el riesgo bayesiano.

10.7.2. Criterios generales para derivar reglas de decisión

Rara vez tenemos reglas $\delta(\mathrm{x})$ con un riesgo uniformemente menor, es decir, que minimicen el riesgo sobre todo el espacio paramétrico Ω. Consideremos la Figura FIGURA NRO. 10.1, allí se representa el comportamiento de dos reglas de decisión hipotéticas δ_1 y δ_2 para un parámetro unidimensional, una con riesgo constante a través del espacio paramétrico y otra cuyo riesgo depende del valor del parámetro. En este caso, en los extremos del espacio paramétrico, antes del punto a y después del punto b se prefiere δ_1, mientras que entre a y b se prefiere δ_2.

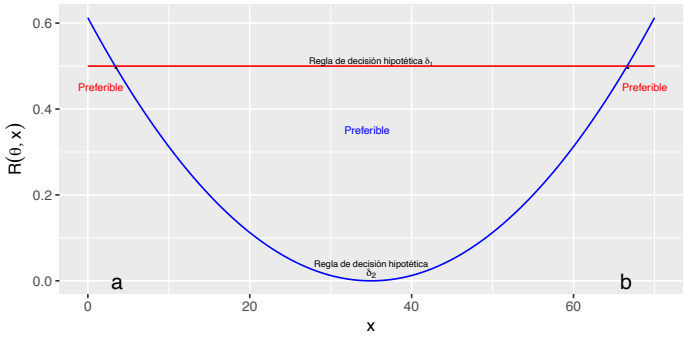

***Figura* 10.1:** Funciones de riesgo hipotéticas de dos reglas de decisión δ_1 y δ_2 para un parámetro unidimensional.
Fuente: elaboración propia generada en R-project [25].

En estos casos se buscan aproximaciones que permitan minimizar el riesgo de alguna forma general, dos principios que se emplean para derivar estimadores (reglas de decisión) con tal tipo de optimalidad son *Minimax* y bayes. El lector interesado en el principio *Minimax* puede consultar Lehmann et al [91]. Por su parte el principio bayes busca la regla de decisión o estimador que minimice el riesgo bayesiano, y en vista de lo discutido previamente, también el riesgo posterior.

10.7.3. Principio de Bayes

Una regla de decisión δ_0 se dice bayes con respecto a la a priori π, si:

$$r(\pi, \delta_0) = \underset{\delta \in D}{inf}\, r(\pi, \delta).$$

Donde inf representa el ínfimo de una función, que corresponde al más grande los límites inferiores de la misma. No se necesita entrar en detalles técnicos, simplemente se debe considerar que se adopta esta definición para dar mayor generalidad. Esta es simplemente la notación matemática para expresar que una regla bayes con respecto a π es aquella que minimiza el riesgo bayesiano (y por ende el riesgo posterior). Martínez et al [88] muestran casos exitosos del uso de la teoría de la decisión bayesiana para encontrar estimadores con propiedades atractivas en problemas de genética-estadística.

Recordemos el interés principal de esta sección, que es mostrar la motivación para emplear parámetros de tendencia central de la distribución posterior como estimadores puntuales. Nótese que la regla de decisión óptima depende de la función de pérdida que se elija y por su definición, sabemos que existen muchas opciones para ello. Se puede demostrar que el estimador bayes (aquel que satisface el principio de bayes) bajo la función de pérdida de error cuadrático medio, que en el caso unidimensional se escribe como $\mathcal{L}(\theta, \delta) = (\theta - \delta)^2$, es la media posterior; similarmente, bajo la pérdida de error absoluto $\mathcal{L}(\theta, \delta) = |\theta - \delta|$, el estimador bayes es una mediana posterior. Se habla de una mediana porque en el caso discreto puede haber más de una. Si Z es una variable aleatoria, entonces M se dice una mediana de la distribución de Z, si:

$$P(Z < M) \leq \frac{1}{2} \leq P(Z \leq M).$$

Finalmente, cuando se usa un tipo de pérdida conocida como 0-1, la moda posterior (si la distribución es unimodal) es el estimador puntual que la minimiza. La función de pérdida 0-1 puede verse como una forma límite de la función de pérdida de Minkowski. Estos resultados son válidos en el caso multidimensional, pero las funciones de pérdida no se escriben de la misma forma.

10.8. Estimador máximo a posteriori (MAP)

En la sección anterior se presentó una justificación para el uso de la media, la mediana y la moda posteriores como estimadores puntuales en inferencia bayesiana. Aquí se discute un tipo de estimador bayesiano puntual que aparece con frecuencia en la literatura y que corresponde, por definición, a la moda posterior. Si la distribución posterior es unimodal, entonces definimos el estimador MAP como:

$$\hat{\theta}_{MAP} = \underset{\theta \in \Omega}{argmax}(\pi(\theta|\mathrm{x}))$$

$$= \underset{\theta \in \Omega}{argmax} \left(\frac{f(\mathrm{x}|\theta)\pi(\theta)}{m(\mathrm{x})} \right)$$

$$= \underset{\theta \in \Omega}{argmax} \left(f(\mathrm{x}|\theta)\pi(\theta) \right)$$

$$= \underset{\theta \in \Omega}{argmax}\, f(\mathrm{x}, \theta).$$

Ahora se debe destacar un fenómeno interesante. Notemos que, si la distribución a priori es proporcional a una constante, como en el caso de una a priori impropia o una a priori uniforme cuando Ω es un subconjunto propio de los reales, el estimador $\hat{\theta}_{MAP}$ estaría maximizando la verosimilitud con respecto al parámetro, esto es, $\hat{\theta}_{MAP} = \underset{\theta \in \Omega}{argmax}\, f(\mathrm{x}|\theta)$, que no es otra cosa más que el estimador de máxima verosimilitud.

Existen casos en los que un estimador bayesiano coincide con un estimador obtenido bajo una aproximación frecuentista, incluso, con métodos determinísticos de estimación de parámetros como lo son las versiones penalizadas de los mínimos cuadrados. En el Capítulo 3 se presenta el BLUP y se menciona que este puede motivarse desde una perspectiva bayesiana, en tanto que, en el Capítulo 7 se menciona que la interpretación del Lasso como un estimador MAP dio lugar a su versión bayesiana. En Martínez et al [88] se muestra que bajo una a priori uniforme y una versión ponderada de la función de pérdida cuadrática, el estimador máximo-verosímil de las frecuencias génicas de un locus bialélico en una especie diploide es igual al estimador bayes.

Como la mayoría de los métodos, modelos y principios empelados en estadística y áreas afines (sino todos), el uso de estos parámetros o "medidas" de tendencia central como estimadores puntuales padece de ciertas falencias o limitantes en situaciones particulares; por ejemplo, consideremos una distribución con una FDP simétrica alrededor de la media (lo que hace que la media sea igual a la mediana) y bimodal como se ilustra en la FIGURA NRO. 10.2 En este caso se tienen dos modas y el estimador MAP no sería único, mientras que la media y la mediana se ubican en una región de baja densidad (en un "valle" de la FDP) y por consiguiente la probabilidad de que la variable tome valores en una vecindad pequeña del estimador es baja, es decir, son valores poco frecuentes.

En el caso de la evaluación genética animal, una alternativa a los estimadores puntuales de los valores genéticos para ordenar los animales candidatos a selección, podría ser el uso de la probabilidad posterior de que el valor genético exceda o sea inferior a un umbral dado, uno que se defina bajo principios zootécnicos. Así, los animales seleccionados serían aquellos con las mayores probabilidades de que su valor genético supere o sea menor al umbral. Para variables como ganancia de peso se buscaría que exceda el umbral, en otras como los días abiertos se buscaría que el valor genético sea menor al umbral. Nótese que se estaría trabajando con la FDA.

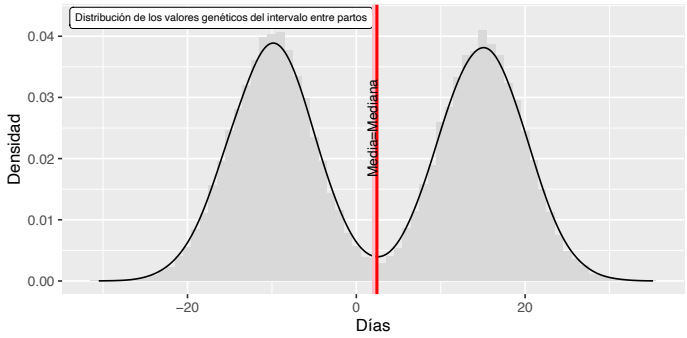

***Figura* 10.2:** Representación de una densidad hipotética que es simétrica alrededor de la media y bimodal. Med = Mediana.
Fuente: elaboración propia generada en R-project [25].

10.9. Modelos jerárquicos

Se inicia esta sección aclarando que el término "modelos jerárquicos" en el ámbito bayesiano no corresponde al que se le da en diseño experimental. Un modelo jerárquico bayesiano es aquel en el que la distribución a priori se especifica mediante una serie de distribuciones condicionales. Consideremos la siguiente partición del parámetro (multidimensional) del modelo $\theta := (\theta_1, \theta_2, \dots, \theta_p)$, donde θ_j, $j = 1, 2, \dots, p$, puede ser uni o multidimensional. Entonces, el modelo jerárquico es como sigue:

$$X|\theta_1, \theta_2, \dots, \theta_p \sim F(\mathrm{x}|\theta_1, \theta_2, \dots, \theta_p)$$

$$\theta_1|\theta_2, \theta_3, \dots, \theta_p \sim \Pi_1(\theta_1|\theta_2, \theta_3, \dots, \theta_p)$$

$$\theta_2|\theta_3, \theta_4, \dots, \theta_p \sim \Pi_2(\theta_2|\theta_3, \theta_4, \dots, \theta_p)$$

$$\vdots$$

$$\theta_p \sim \Pi_p(\theta_p)$$

Por lo tanto, la distribución a priori conjunta $\Pi(\theta) = \Pi(\theta_1, \theta_2, \dots, \theta_p)$ corresponde al producto de las distribuciones condicionales $\Pi_1, \Pi_2, \dots, \Pi_p$, así, si denotamos las FDP (caso absolutamente continuo) o FMP (caso discreto) correspondientes como $\pi_1, \pi_2, \dots, \pi_p$, tenemos:

$$\pi(\theta) = \pi(\theta_1, \theta_2, \dots, \theta_p)$$

$$= \pi(\theta_1|\theta_2, \theta_3, \dots, \theta_p)\pi(\theta_2|\theta_3, \theta_4, \dots, \theta_p) \dots \pi(\theta_p)$$

$$= \pi(\theta_p) \prod_{j=1}^{p-1} \pi(\theta_j|\theta_{j+1}, \dots, \theta_p)$$

Este resultado es consecuencia de extender la regla del producto a p variables. De aquí se sigue que:

$$\pi(\theta|\mathrm{x}) \propto \pi(\theta) f(\mathrm{x}|\theta)$$

$$= \pi(\theta_p) \prod_{j=1}^{p-1} \pi(\theta_j|\theta_{j+1}, \ldots, \theta_p) \times f(\mathrm{x}|\theta)$$

Esta sería la FDP/FMP posterior conjunta, pero podemos hablar de otras distribuciones de interés como: $\Pi(\theta_j|\mathrm{x})$: la distribución posterior marginal de $\theta_j, j = 1, 2, \ldots p$

$\Pi(\theta_j, \theta_k|\mathrm{x})$: la distribución posterior marginal conjunta de $\theta_j, \theta_k; 1 \leq j, k \leq p, j \neq k$.

10.10. Inferencia aproximada por métodos de tipo Monte Carlo Cadenas de Markov

Este grupo de métodos conocido como MCMC por sus siglas en inglés (*Markov Chain Monte Carlo*) es una familia de algoritmos que permiten obtener muestras aproximadas de la distribución posterior cuando esta no se puede encontrar de manera exacta. Muchos de los modelos empleados en análisis bayesianos, incluso algunos relativamente simples, exhiben la particularidad de que la distribución posterior no es fácilmente analizable desde el punto de vista matemático, a esto se le conoce como intratabilidad matemática, o equivalentemente, se dice que la posterior no es tratable matemáticamente. Así, parámetros de interés como la media o la varianza posterior no se pueden encontrar de manera analítica, es decir, no se pueden expresar mediante una fórmula como se hizo en la sección 10.6. Esto resulta porque la sumatoria o la integral que se debe calcular para obtener estos momentos no tiene una solución exacta; lo cual no significa que no se puedan obtener aproximaciones numéricas. También existen distribuciones para las que la constante normalizadora no se conoce, un ejemplo es la distribución G-Wishart o Wishart gráfica, que es una extensión de la Wishart utilizada en una clase de modelos estadísticos llamada modelos gráficos. En general su constante normalizadora no se conoce, solo en algunos casos particulares, razón por la cual, los momentos de la distribución tampoco se conocen de manera exacta.

En ramas de la matemática como el cálculo integral se cuenta con métodos numéricos como el del trapecio o el método de Simpson para encontrar valores numéricos de integrales definidas que no se pueden resolver de manera exacta, caso que también ocurre en ecuaciones diferenciales donde se usan métodos numéricos como el de Runge-Kutta para resolver problemas que no son tratables matemáticamente. Estos métodos son de tipo determinístico; sin embargo, se cuenta con otros que son de tipo estocástico porque se basan en la generación de variables aleatorias. Entre tales métodos, algunos para integrar, otros para derivar y optimizar funciones (encontrar puntos máximos o mínimos). Cuando el problema implica variables

aleatorias estándar, de las que se pueden obtener muestras directas, se emplean métodos de Monte Carlo, los cuales usan estas muestras y algunos resultados de la teoría de la probabilidad para tamaños de muestra grandes (asintóticos) para llegar a una respuesta aproximada que goza de algunas garantías teóricas de convergencia. En estos casos no hay necesidad de construir cadenas de Markov y por ello, el nombre de este grupo de métodos es simplemente Monte Carlo. Cuando la distribución posterior no es conocida, pero el parámetro θ se puede particionar en subconjuntos de parámetros (uni o multidimensionales) $\theta_1, \theta_2, \dots, \theta_k$, tales que, las distribuciones de cada θ_j, $j = 1, 2, \dots, k$, dados los datos x y los demás parámetros (denotados como $\theta_{[-j]}$), permiten que se obtengan muestras directas de la misma, existe un algoritmo de gran popularidad que se denomina muestreador de Gibbs. Este se presta particularmente bien para los modelos bayesianos jerárquicos descritos en la sección 10.9, especialmente los modelos lineales empleados en evaluación genómica que se estudiaron en al Capítulo 7. Las distribuciones θ_j dados los datos y los demás parámetros se notan como $F(\theta_j|\theta_{[-j]}, \mathrm{x})$ y se conocen como condicionales completas, pero también se escriben $F(\theta_j|Else)$ en inglés, o $F(\theta_j|lo\ demás)$.

Explícitamente:

$$\theta_{[-j]} = (\theta_1, \theta_2, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p) \forall j = 1, 2, \dots, p.$$

Antes de presentar el muestreador de Gibbs, introducimos algunos conceptos sobre cadenas de Markov.

Cadenas de Markov. Secuencias de variables aleatorias $\{X_t\}_{t\geq 0}$ que satisfacen:

$$X_t|X_0, X_1, \dots, X_{t-1} \overset{d}{=} X_t|X_{t-1}$$

A esta propiedad se le conoce como propiedad de Markov y nos muestra independencia condicional de X_t y $X_0, X_1, \dots, X_{t-2}$ dada X_{t-1}.

"Kernel" de transición o "Kernel" de Markov: una función de dos argumentos tal que:

$$X_t|X_0, X_1, \dots, X_{t-1} \sim K(X_t, X_{t-1})$$

Distribución estacionaria: sea $F(\bullet)$ una FDA, esta se dice estacionaria si satisface:

$$X_t \sim F(\bullet) \Longrightarrow X_{t+1} \sim F(\bullet) \forall t = 0, 1, 2, \dots$$

El "kernel" de transición se relaciona con esta distribución así:

$$\int_{\boldsymbol{x}} K(x, y) f(x) dx = f(y)$$

Siendo $f(\bullet)$ la densidad de $F(\bullet)$.

Espacio de estados: conjunto de todos los valores que la cadena puede tomar.

Cadena irreducible: una cadena de Markov se dice irreducible si $\forall x_0$ (el valor inicial de la cadena) esta tiene una probabilidad positiva de eventualmente visitar cualquier subconjunto del espacio de estados. La irreductibilidad es necesaria para la existencia de $F(\bullet)$, esta distribución no siempre existe, pero la mayoría de las cadenas usadas en la praxis gozan de esta propiedad. La condición $K(x, \bullet) > 0$ en todas partes es suficiente para que la cadena sea irreducible.

Ergodicidad: cuando la distribución estacionaria es también la distribución límite para cualquier x_0, esta propiedad se denomina ergodicidad.

Recurrencia: la cadena regresa a cualquier punto no negligible un infinito número de veces. La importancia de las cadenas recurrentes es que satisfacen la siguiente propiedad:

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \longrightarrow E_F[h_{(x)}]$$

Un resultado que corresponde a la ley de los grandes números para el caso de cadenas de Markov y que brinda la base teórica que justifica el uso de los métodos MCMC para aproximar parámetros de interés de la distribución posterior. Cuando la distribución estacionaria existe, entonces la cadena es recurrente.

El procedimiento que se sigue al implementar un método MCMC para aproximar una distribución se resume a continuación. Se tiene una FDP/FMP objetivo f, entonces se construye un "Kernel" de Markov K, tal que la distribución estacionaria tenga FDP o FMP f, luego, usando K, se genera una cadena de Markov cuya distribución límite tenga FDP/FMP f. La dificultad de este procedimiento radica en construir K, pero, como lo destacan Robert et al [92], es casi milagroso que existan métodos universales para esta tarea. Son universales porque son teóricamente válidos para cualquier densidad f o función de masa y esta propiedad hace que estos métodos hayan ganado tanta popularidad.

10.10.1. El muestreador de Gibbs

Este y los demás algoritmos tipo MCMC son usados en diversas áreas, no solamente en inferencia bayesiana. Por lo tanto, en esta sección discutimos la forma general del algoritmo, empleando notación acorde. Iniciamos con el caso bivariado (dos estados), queremos obtener muestras de una función cuya PDF o PMF es $f(x, y)$. Se define un valor inicial X_0 que esté en el soporte de X. En la iteración $t, t = 1, 2, \ldots, T$, obtenemos:

$$y_t \sim F(y|x_{t-1})$$

$$x_t \sim F(x|y_t).$$

Esta notación quiere decir que se muestrea el valor y_t de la distribución condicional $F(y|x_{t-1})$ y x_t de la distribución condicional $F(x|y_t)$. Así, en la iteración t del algoritmo se obtiene y_t, empleando el valor más reciente de x que corresponde al de la iteración inmediatamente anterior, esto es, x_{t-1}. Por su parte, al muestrear x_t, el valor más reciente de y es y_t y este es el valor que se emplea. Nótese que se está usando la distribución condicional completa, es decir, la distribución de cada variable dadas todas las demás, que en este caso bivariado es simplemente la distribución de Y dado $X = x_{t-1}$ y la de X dado $Y = y_t$.

En el caso general consideramos p variables aleatorias $X_1, X_2, \dots, X_p$ y el objetivo es obtener muestras de su distribución conjunta. Denotamos la j-ésima variable aleatoria en la t-ésima iteración como $X_j^{(t)}$ y como $x_j^{(t)}$ su respectiva realización, es decir, el valor muestreado. En la iteración t:

$$X_1^{(t)} \sim F(x_1|x_2^{(t-1)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})$$

$$X_2^{(t)} \sim F(x_2|x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})$$

$$X_3^{(t)} \sim F(x_3|x_1^{(t)}, x_2^{(t)}, x_4^{(t-1)} \dots, x_p^{(t-1)})$$

$$\vdots$$

$$X_p^{(t)} \sim F(x_p|x_1^{(t)}, x_2^{(t)}, \dots, x_{p-1}^{(t)})$$

En el límite (cuando t tiende a infinito) se obtienen muestras de la distribución conjunta $F(x_1, x_2, \dots, x_p)$ y las marginales $F(x_j)$, $j = 1, 2, \dots, p$. Esto indica que eventualmente se obtienen muestras de las distribuciones objetivo.

Al implementar el muestreador de Gibbs en un modelo bayesiano, $X_1, X_2, \dots, X_p$ corresponden a los parámetros del modelo $\theta_1, \theta_2, \dots, \theta_p$ y para obtener muestras de cada uno, se emplea la respectiva distribución condicional completa con los valores más recientes de los demás parámetros y los datos observados x. Explícitamente, en la iteración t:

$$\theta_1^{(t)} \sim F(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_p^{(t-1)}, \mathrm{x})$$

$$\theta_2^{(t)} \sim F(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_p^{(t-1)}, \mathrm{x})$$

$$\theta_3^{(t)} \sim F(\theta_3|\theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)} \dots, \theta_p^{(t-1)}, \mathrm{x})$$

$$\vdots$$

$$\theta_p^{(t)} \sim F(\theta_p|\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{p-1}^{(t)}, \mathrm{x})$$

El muestreador de Gibbs es un caso especial de un método más general llamado Metropolis-Hastings, el cual tiene varias versiones que no se discuten aquí y se emplea cuando no es posible obtener muestras directas de las condicionales completas. Existen casos en los que solo algunas condicionales completas no se pueden muestrear directamente, en ese caso se emplea un método "híbrido" llamado muestreador de Gibbs con paso Metrópolis.

Una vez se está satisfecho con el comportamiento de la cadena (tema que se discute brevemente más adelante), las muestras obtenidas se emplean para llevar a cabo las inferencias deseadas. La teoría detrás de las cadenas de Markov nos permite trabajar con las versiones muestrales de los parámetros de interés, de esta forma, si se quiere estimar la media posterior, su estimador MCMC no es más que la media aritmética de la muestra obtenida de la cadena de Markov; se procede igual para la varianza u otro momento de la distribución, así como con los cuantiles. Similarmente, si se quiere estimar la probabilidad de que un parámetro tome valores en un conjunto dado, se emplearía la frecuencia con que esto sucede en la muestra. Esta simplicidad es una propiedad muy atractiva de esta forma de inferencia aproximada.

Cuando se usan estos métodos se debe chequear convergencia de la cadena. La dificultad es que existen varias nociones de convergencia, un hecho que no es muy satisfactorio para algunos científicos. Discutir las nociones de convergencia y las técnicas para chequearlas en detalle va más allá del alcance de este texto, el lector interesado puede referirse a Robert et al [92] para una introducción amigable al tema. El paquete de R-project [25] llamado «CODA» [93] se ha diseñado explícitamente para chequear convergencia de cadenas de Markov, sin embargo, no podemos dejar de mencionar este importante resultado: cuando la posterior es impropia, no se tiene convergencia de los algoritmos MCMC.

Finalizamos este capítulo con dos comentarios relevantes, el primero tiene que ver con la manera en que comúnmente se implementan los métodos MCMC. En vista de que eventualmente el método simula variables aleatorias provenientes de la distribución objetivo, sucede que las primeras muestras no provienen de esta distribución y emplearlas podría "contaminar" de alguna manera la inferencia. Por esta razón se suelen descartar las primeras muestras obtenidas, a estas se les denomina "periodo de calentamiento" o simplemente "calentamiento". Por otro lado, debido a que las variables aleatorias muestreadas están correlacionadas, se suele ignorar un cierto número de muestras consecutivas y solo considerar las restantes; por ejemplo, guardar una muestra, ignorar las diez siguientes, volver a guardar la onceava y continuar así con todas las muestras generadas.

El segundo comentario tiene que ver con la diferenciación entre los métodos numéricos que permiten hacer una inferencia aproximada cuando la posterior es intratable y la aproximación bayesiana a la estadística. En modelos sencillos en los que se puede hacer inferencia exacta (Ejemplo 10.4 y sección 10.6), no hay necesidad de métodos numéricos; por otro lado, existen alternativas como la aproximación de Laplace (como lo indicó Ghosh [89]) y su variante llamada aproximación de Laplace integrada y anidada que es aplicable en un tipo particular de modelos [94]. El punto relevante es que los métodos de Monte Carlo no hacen parte de la estadística bayesiana, ni esta hace parte de los métodos de Monte Carlo. En ciencias aplicadas como las pecuarias, es común la creencia incorrecta de que estudiar estadística bayesiana corresponde a estudiar los métodos MCMC. También es común encontrar expresiones como "los parámetros se estimaron empleando un muestreador de Gibbs", lo cual carece de sentido puesto que aquí se debe mencionar el método estadístico usado, por ejemplo, cuando se emplean procedimientos de modelos lineales mixtos para

encontrar el mejor predictor lineal insesgado MPLI de los valores genéticos, se utilizan métodos numéricos para resolver las ecuaciones de modelos mixtos, como por ejemplo el método de Jacobi, pero el método empleado para predecir los valores genéticos fue el MPLI no Jacobi. Similarmente, bajo la aproximación frecuentista, cuando se estiman componentes de varianza mediante máxima verosimilitud, se acude a métodos numéricos para resolver los sistemas de ecuaciones resultantes, como por ejemplo Newton-Raphson; no sería adecuado decir que los componentes de varianza se estimaron con el método Newton-Raphson. Muchos métodos estadísticos implican resolver sistemas de ecuaciones, integración u optimización y es frecuente que se deba acudir a métodos numéricos para dicha tarea, pero el método numérico no es el método de inferencia.

11 CAPÍTULO ONCE

RENUMERACIÓN DE ANIMALES CON R-PROJECT

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

A continuación desarrollaremos el EJEMPLO 11 con una de las maneras de realizar la renumeración de animales, mediante la librería ≪ggroups≫ [95]. Adicionalmente, utilizaremos la librería ≪doBy≫ [40] para el ordenamiento de datos.

Librerías utilizadas:

```
library(doBy)
library(ggroups)
```

Genealogiá:

```
GENEALOGIA=data.frame(matrix(ncol=4,byrow=TRUE,c(
"B1","A1","A2",1,
"B3","A1","A2",1,
"B5","A1","A4",1,
"B7",NA,NA,1,
"C02","A1","A2",2,
"C04","A3","A2",2,
"C06",NA,NA,2,
"C08",NA,NA,2,
"H06","B1","C06",2,
"H12","B3","C08",2,
"A1",NA,NA,1,
"A2",NA,NA,2,
"A3",NA,NA,1,
"A4",NA,NA,2,
"H02","B1","C02",2,
"H04","B1","C04",2,
"H14","B3","C02",2)))
colnames(GENEALOGIA)=c("id","sire","dam","sex")
```

```
GENEALOGIA
```

```
##     id sire  dam sex
## 1   B1   A1   A2   1
## 2   B3   A1   A2   1
## 3   B5   A1   A4   1
## 4   B7 <NA> <NA>   1
## 5  C02   A1   A2   2
## 6  C04   A3   A2   2
## 7  C06 <NA> <NA>   2
## 8  C08 <NA> <NA>   2
## 9  H06   B1  C06   2
## 10 H12   B3  C08   2
## 11  A1 <NA> <NA>   1
## 12  A2 <NA> <NA>   2
## 13  A3 <NA> <NA>   1
## 14  A4 <NA> <NA>   2
## 15 H02   B1  C02   2
## 16 H04   B1  C04   2
## 17 H14   B3  C02   2
```

Relaciones entre padres y madres:

```
table(GENEALOGIA$sire,GENEALOGIA$dam)
```

```
##
##      A2 A4 C02 C04 C06 C08
##   A1  3  1   0   0   0   0
##   A3  1  0   0   0   0   0
##   B1  0  0   1   1   1   0
##   B3  0  0   1   0   0   1
```

Renumeración de animales con la librería «ggroups»:

Primero: utilizar solo las tres columnas y poner ceros en padres y madres desconocidos.

```
Pedigree=GENEALOGIA[,c("id","sire","dam")]
Pedigree$sire=ifelse(is.na(Pedigree$sire),0,Pedigree$sire)
Pedigree$dam=ifelse(is.na(Pedigree$dam),0,Pedigree$dam)
Pedigree
```

```
##     id sire dam
## 1   B1   A1  A2
## 2   B3   A1  A2
## 3   B5   A1  A4
## 4   B7    0   0
## 5  C02   A1  A2
## 6  C04   A3  A2
## 7  C06    0   0
## 8  C08    0   0
## 9  H06   B1 C06
## 10 H12   B3 C08
## 11  A1    0   0
## 12  A2    0   0
## 13  A3    0   0
## 14  A4    0   0
## 15 H02   B1 C02
## 16 H04   B1 C04
## 17 H14   B3 C02
```

```
NuevoPed=renum(Pedigree)$newped
```

```
## Found 3 generations
```

```
Codigos=renum(Pedigree)$xrf

## Found 3 generations

colnames(Codigos)=c("id","ID")
Renumerada=merge(NuevoPed,Codigos)
Renumerada

##     ID SIRE DAM  id
## 1    1    0   0  B7
## 2    2    0   0 C06
## 3    3    0   0 C08
## 4    4    0   0  A1
## 5    5    0   0  A2
## 6    6    0   0  A3
## 7    7    0   0  A4
## 8    8    4   5  B1
## 9    9    4   5  B3
## 10 10    4   5 C02
## 11 11    6   5 C04
## 12 12    4   7  B5
## 13 13    8  10 H02
## 14 14    9  10 H14
## 15 15    8  11 H04
## 16 16    8   2 H06
## 17 17    9   3 H12
```

Ahora pegamos las dos bases de datos:

```
GENEALOGIA=merge(GENEALOGIA,Renumerada)

GENEALOGIA=orderBy(~ID, data=GENEALOGIA)
GENEALOGIA

##      id sire  dam sex ID SIRE DAM
## 8    B7 <NA> <NA>   1  1    0   0
## 11 C06 <NA> <NA>   2  2    0   0
## 12 C08 <NA> <NA>   2  3    0   0
## 1    A1 <NA> <NA>   1  4    0   0
## 2    A2 <NA> <NA>   2  5    0   0
## 3    A3 <NA> <NA>   1  6    0   0
## 4    A4 <NA> <NA>   2  7    0   0
## 5    B1   A1   A2   1  8    4   5
```

```
## 6   B3   A1   A2   1  9    4   5
## 9  C02   A1   A2   2 10    4   5
## 10 C04   A3   A2   2 11    6   5
## 7   B5   A1   A4   1 12    4   7
## 13 H02   B1  C02   2 13    8  10
## 17 H14   B3  C02   2 14    9  10
## 14 H04   B1  C04   2 15    8  11
## 15 H06   B1  C06   2 16    8   2
## 16 H12   B3  C08   2 17    9   3
```

11.1. Errores frecuentes en los nombres de los individuos

Es recurrente que se tengan errores en los nombres de los individuos, por consiguiente, el analista debe tener mucho cuidado con la confirmación de la información. Esta tarea es la más compleja de las evaluaciones genéticas y requiere de tiempo, paciencia, cuidado y habilidad para detectar incoherencias. Veamos a continuación algunos ejemplos:

1) Un individuo con nombres diferentes (ej. *GABOR*7*HO*8477 y *Gabor*7*Ho*8477, *GABOR*, 7*HO*8477, *GAVOR*7*HO*8477, etc.)

2) Un nombre que aparezca como madre y como padre (ej. como padre *Aike* y como madre *Aike*)

3) Un individuo que tenga un nombre en la columna de individuo y nombre diferente como padre o como madre (ej. como individuo $Willow-Marsh-CCGabor-ET$ y como padre *GABOR*7*HO*8477).

A continuación traemos el EJEMPLO 11.1 con una lista de 9 nombres de toros que aparentemente son distintos, pero que en realidad hay errores de escritura.

```
Toros=data.frame(c(
"BALZAIR 507HO07947",
"BLADE 7HO7744",
"GABOR 7HO8477",
"COLDSPRINGS",
"CARUSO_7HO8866",
"GABOR__7HO8477",
"COLDSPRINGS 7HO7536",
"GABOR7HO8477",
"GABOR-7HO8477"))
Lista=data.frame(table(Toros))
colnames(Lista)=c("Sire","Veces")
Lista
```

```
##                   Sire Veces
## 1  BALZAIR 507HO07947     1
## 2        BLADE 7HO7744     1
## 3      CARUSO_7HO8866     1
## 4          COLDSPRINGS     1
## 5 COLDSPRINGS 7HO7536     1
## 6        GABOR 7HO8477     1
## 7       GABOR__7HO8477     1
## 8        GABOR-7HO8477     1
## 9         GABOR7HO8477     1


#convertir la columna como una cadena de caracteres
Lista$Sire=as.character(Lista$Sire)
```

Como pueden observar, hay nombres que identifican a un solo toro (caso de *COLDSPRINGS* y *GABOR*), por consiguiente, es necesario hacer comparaciones entre todos los nombres para detectar similitud, la cual puede ser por observación directa y también, con ayuda de procedimientos para detectar similitud.

A continuación, usaremos dos librerías: ≪gtools≫ [96] con el comando ≪combinations≫ para hacer dos columnas que permitan la comparación pareada de todos los nombres y ≪synthesisr≫ [97], con el comando ≪fuzzdist≫ para detectar similitud, la cual puede ser por diversas metodologías que comparan coincidencias de caracteres.

```
library(gtools)
Base<- as.matrix(combinations(n= nrow(Lista),
                              r = 2, v = Lista$Sire))
#Algunas combinaciones posibles de nombres
Base[c(1:2,30:36),]


##       [,1]                  [,2]
##  [1,] "BALZAIR 507HO07947"  "BLADE 7HO7744"
##  [2,] "BALZAIR 507HO07947"  "CARUSO_7HO8866"
##  [3,] "COLDSPRINGS 7HO7536" "GABOR7HO8477"
##  [4,] "GABOR 7HO8477"       "GABOR__7HO8477"
##  [5,] "GABOR 7HO8477"       "GABOR-7HO8477"
##  [6,] "GABOR 7HO8477"       "GABOR7HO8477"
##  [7,] "GABOR__7HO8477"      "GABOR-7HO8477"
##  [8,] "GABOR__7HO8477"      "GABOR7HO8477"
##  [9,] "GABOR-7HO8477"       "GABOR7HO8477"
```

Para la similtud, utilizaremos el método ≪fuzz_m_ratio≫ que genera una medida del número de letras que coinciden entre dos cadenas de caracteres (en nuestro caso

nombres de toros). Se calcula como uno menos dos veces el número de caracteres coincidentes, dividido por el número de caracteres en ambas cadenas:

```
library(synthesisr)


##
## Attaching package: 'synthesisr'

## The following object is masked from 'package:knitr':
##
##     write_bib


#Hacemos un vector que tenga la medida de comparación
#de las dos columnas de la matriz Base para cada fila
#(i desde 1 hasta la última fila)
Vector=matrix(ncol=1,nrow=nrow(Base))
for (i in 1:nrow(Base)){
  Vector[i,1]=(fuzzdist(Base[i,1],
                        Base[i,2],
                        method = "fuzz_m_ratio"))
}
Base1=data.frame(Base,Vector)
Base1$Vector=round(Base1$Vector,2)
#entre menor sea el valor, mayor similitud de
#cadenas de caracteres
library(doBy)
similitud=orderBy(~Vector,Base1)
head(similitud)


##                X1                   X2 Vector
## 32  GABOR 7HO8477       GABOR-7HO8477   0.08
## 22    COLDSPRINGS COLDSPRINGS 7HO7536   0.27
## 33  GABOR 7HO8477        GABOR7HO8477   0.52
## 36  GABOR-7HO8477        GABOR7HO8477   0.52
## 31  GABOR 7HO8477      GABOR__7HO8477   0.56
## 34 GABOR__7HO8477       GABOR-7HO8477   0.56
```

12 CAPÍTULO DOCE

Cálculo del tamaño efectivo de la población que se reproduce

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

Para la inplementación de programas de mejora genética o evaluación genética de una población, es conveniente hacer estudios iniciales, como es el caso, del coeficiente de consanguinidad (f), cálculo del tamaño efectivo de la población que se reproduce (Ne), el reconocimiento de las principales familias de la población y análisis de posibles apareamientos que se realizarán, para conocer el coeficiente de consanguinidad que se tendrá en la próxima generación (f_s). Esto permite establecer los lineamientos de los apareamientos para equilibrar el efecto de la presión de selección, con el aumento de la consanguinidad, la pérdida de variabilidad genética y la presencia de cuellos de botella.

En términos evolutivos, el *Ne* contempla los individuos que contribuirán genéticamente en la conformación de las generaciones siguientes. Una población puede tener una gran cantidad de hembras y machos, pero, si la mayoría son parientes y endogámicos, implica que se tiene un *Ne* reducido. Lo que podría ocasionar que la siguiente generación presente un bajo porcentaje de heterocigosis y un alto porcentaje de consanguinidad.

En genética de poblaciones, el *Ne* está dado por número de individuos aptos para reproducción, que efectivamente contribuyen para mantener la variabilidad genética de una población ideal (supuestos de Hardy-Weinberg) en la siguiente generación, la cual puede estar afectada por la consanguinidad o la deriva genética (como lo

indicaron Kimura y Crow [98]). El efecto de consanguinidad es una consecuencia de la autocigosidad (homocigosidad para un alelo que es identico por ascendencia) y la deriva genética (cambios de las frecuencia génicas por muestreo de una generación a otra), como lo indicaron Kimura y Crow [98].

El Ne se puede calcular de diversas formas, la más sencilla y generalmente usada está dada por la relación entre el número de indivíduos que se reproducen (N) y el coeficiente de endogamia promedio ($\hat{f}$), como se indica a continuación:

$$Ne = \frac{N}{1 + \hat{f}} \tag{12.1}$$

El Ne podría calcularse teniendo en cuenta el promedio de los coeficientes de endogamia de los hijos $\overline{C_{jk}}$ que tendría la población que se reproduce, así:

$$Ne^{*} = \frac{N}{1 + \overline{C_{jk}}} \tag{12.2}$$

Donde C_{jk} es el coeficiente de endogamia de un hijo de j y k.

Teniendo en cuenta el procedimiento anterior, podemos construir una matriz de coancestría con los posibles apareamientos de los animales actuales, dejando en las filas los individuos machos (padre j) y en las columnas las hembras (madre k), quedando la matriz C_{pxm} de tamaño igual al número de padres p y número de madres m. El número de elementos de esta matriz correspondería al número de posibles combinaciones de padres para gestar hijos ($Nd = pxm$) .

Coeficiente de endogamia promedio en los posibles apareamientos, sería:

$$\overline{C_{jk}} = \frac{\sum_{j=1}^{p} \sum_{k=1}^{m} c_{jk}}{pxm}$$

El tamaño efectivo de la población Ne^{**} será:

$$Ne^{**} = \frac{p + m}{1 + \overline{C_{pm}}} \tag{12.3}$$

Sin embargo, el tamaño efectivo de la población en la generación siguiente (Ne_d) dependerá del número de posibles nacimientos (Nac) en un periodo de partos y los coeficientes de endogamia de los individuos que nazcan:

El tamaño efectivo de la población, teniendo en cuenta los hijos que se tendrían en una estación de partos, se calcula así:

$$Ne_e = \frac{Nac}{1 + \overline{C_{pm}}} \tag{12.4}$$

Para garantizar que no exista aumento en los coeficientes de endogamia, se establece un límite máximo que podrán tener los hijos y se recalcula el tamaño efectivo de la población que se tendría en la siguiente generación.

Para el cálculo del tamaño efectivo desarrollaremos el EJEMPLO 12, usando una genealogía con presencia y ausencia de animales consanguíneos y de animales emparentados. Los individuos *R*1, *R*2, *R*3, *R*4, *R*5 y *R*6 (N=6, 3 machos y 3 hembras); son los que están activos para reproducción (Estado actual ≪blue≫), el restante de animales, son ancestros de la población actual (Estado actual ≪red≫), como se indica en la TABLA NRO. 12.1.

TABLA 12.1: Información genelógica de una población

Animal	Madre	Padre	Sexo	Estado actual
A1			1	red
A2			2	red
A3			1	red
A4			2	red
A5			1	red
A6			2	red
A7			1	red
A8			2	red
B1	A1	A2	1	red
B2	A3	A4	2	red
B3	A5	A6	1	red
B4	A7	A8	2	red
R1	B1	B2	1	blue
R2	B3	B4	2	blue
R3			1	blue
R4	R1	R2	2	blue
R5	B3	R2	1	blue
R6	R3	R2	2	blue
R8	R1	R6	2	blue

Nota: Información de 19 animales, de los cuales algunos están para reproducción (*blue*) y otros hacen parte de la genealogía (*red*).
Fuente: elaboración propia (2024).

Matriz de parentesco:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0.25 & 0 & 0 & 0.12 & 0 & 0 & 0.12 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0.25 & 0 & 0 & 0.12 & 0 & 0 & 0.12 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0.25 & 0 & 0 & 0.12 & 0 & 0 & 0.12 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0.25 & 0 & 0 & 0.12 & 0 & 0 & 0.12 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0.25 & 0 & 0.12 & 0.38 & 0.12 & 0.06 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.12 & 0.12 & 0.12 & 0.12 & 0.12 & \cdots & 0.5 & 0.5 & 0 & 1 & 0.38 & 0.25 & 0.38 \\ 0 & 0 & 0 & 0 & 0.38 & \cdots & 0 & 0.75 & 0 & 0.38 & 1.25 & 0.38 & 0.19 \\ 0 & 0 & 0 & 0 & 0.12 & \cdots & 0 & 0.5 & 0.5 & 0.25 & 0.38 & 1 & 0.5 \\ 0.12 & 0.12 & 0.12 & 0.12 & 0.06 & \cdots & 0.5 & 0.25 & 0.25 & 0.38 & 0.19 & 0.5 & 1 \end{bmatrix}$$

Animales de la generación con potencial para reproducirse (R1, R2, R3, R4, R5, R6 y R8).

Número de animales:

$$N = 7$$

Matriz de parentesco de los animales seleccionados:

$$A_s = \begin{bmatrix} 1 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0.5 & 0.75 & 0.5 & 0.25 \\ 0 & 0 & 1 & 0 & 0 & 0.5 & 0.25 \\ 0.5 & 0.5 & 0 & 1 & 0.38 & 0.25 & 0.38 \\ 0 & 0.75 & 0 & 0.38 & 1.25 & 0.38 & 0.19 \\ 0 & 0.5 & 0.5 & 0.25 & 0.38 & 1 & 0.5 \\ 0.5 & 0.25 & 0.25 & 0.38 & 0.19 & 0.5 & 1 \end{bmatrix}$$

Coeficientes de endogamia:

$$f = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.25 \\ 0 \\ 0 \end{bmatrix}$$

Promedio de consanguinidad

$$\widehat{f} = 0.04$$

Tamaño efectivo de la población:

$$Ne = \frac{7}{1 + 0.04} = 6.73$$

Matriz de coancestría:

$$C = \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0 & \cdots & 0.12 & 0 & 0 & 0.06 & 0 & 0 & 0.06 \\ 0 & 0.5 & 0 & 0 & 0 & \cdots & 0.12 & 0 & 0 & 0.06 & 0 & 0 & 0.06 \\ 0 & 0 & 0.5 & 0 & 0 & \cdots & 0.12 & 0 & 0 & 0.06 & 0 & 0 & 0.06 \\ 0 & 0 & 0 & 0.5 & 0 & \cdots & 0.12 & 0 & 0 & 0.06 & 0 & 0 & 0.06 \\ 0 & 0 & 0 & 0 & 0.5 & \cdots & 0 & 0.12 & 0 & 0.06 & 0.19 & 0.06 & 0.03 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.06 & 0.06 & 0.06 & 0.06 & 0.06 & \cdots & 0.25 & 0.25 & 0 & 0.5 & 0.19 & 0.12 & 0.19 \\ 0 & 0 & 0 & 0 & 0.19 & \cdots & 0 & 0.38 & 0 & 0.19 & 0.62 & 0.19 & 0.09 \\ 0 & 0 & 0 & 0 & 0.06 & \cdots & 0 & 0.25 & 0.25 & 0.12 & 0.19 & 0.5 & 0.25 \\ 0.06 & 0.06 & 0.06 & 0.06 & 0.03 & \cdots & 0.25 & 0.12 & 0.12 & 0.19 & 0.09 & 0.25 & 0.5 \end{bmatrix}$$

Machos: R1, R3 y R5.

Hembras: R2, R4, R6 y R8.

Matriz de coancestría de los posibles apareamientos de los animales actuales (en las filas los machos y las hembras en las columnas)

$$C = \begin{bmatrix} 0 & 0.25 & 0 & 0.25 \\ 0 & 0 & 0.25 & 0.12 \\ 0.38 & 0.19 & 0.19 & 0.09 \end{bmatrix}$$

Número de combinaciones o apareamientos posibles:

$$Nd = 12$$

Coeficiente de endogamia promedio en los posibles apareamientos, sería:

$$\overline{C_{jk}} = 0.14$$

El tamaño efectivo de la población en este escenario, teniendo en cuenta los 3 machos y las 4 hembras, sería:

$$Ne^{**} = \frac{3+4}{1+0.14} = 6.14$$

Si consideramos que en una estación de partos nace un hijo por hembra, entonces:

$$Nac = 4$$

Tamaño efectivo de la población teniendo en cuenta los hijos que se tendrían en una estación de partos:

$$Ne_e^{**} = \frac{4}{1 + 0.14} = 3.44$$

Listado de posibles apareamientos con bajo porcentaje de endogámia (por ejemplo, menor que 10 %):

$$C = \begin{bmatrix} 0 & NA & 0 & NA \\ 0 & 0 & NA & NA \\ NA & NA & NA & 0.09 \end{bmatrix}$$

Coeficiente de endogamia promedio en los posibles apareamientos:

$$\overline{C_{jk}} = 0.02$$

El número de hijos que pueden nacer en una estación de partos, es:

$$Nac = 4$$

El tamaño efectivo de la población teniendo en cuenta los hijos que se tendrían en una estación de partos:

$$Ne_d^{**} = \frac{4}{1 + 0.02} = 3.9$$

12.0.1. Ejercicios en R-project

Utilizaremos las librerías ≪doBy≫ [40] para organizar datos y ≪kinship2≫ de [28] para obtener la matriz de parentesco A y la matriz de coancestría C. Para el cálculo del tamaño efectivo utilizaremos procedimientos descritos por MacCluer et al [99], Cervantes et al [100] y la información suministrada en la librería ≪optiSel≫ [101]:

```
library(kinship2)
```

La población activa que es apta para reproducción son R1, R2, R3, R4, R5, R6 y R8 son siete individuos ($N = 7$).

```
Pedigri=data.frame(matrix(ncol=5,c(
"A1", NA, NA,1,"red",
"A2", NA, NA,2,"red",
"A3", NA, NA,1,"red",
"A4", NA, NA,2,"red",
"A5", NA, NA,1,"red",
"A6", NA, NA,2,"red",
"A7", NA, NA,1,"red",
"A8", NA, NA,2,"red",
"B1","A1","A2",1,"red",
"B2","A3","A4",2,"red",
"B3","A5","A6",1,"red",
"B4","A7","A8",2,"red",
"R1","B1","B2",1,"blue",
"R2","B3","B4",2,"blue",
"R3",NA,NA,1,"blue",
"R4","R1","R2",2,"blue",
"R5","B3","R2"  ,1,"blue",
"R6","R3","R2",2,"blue",
"R8","R1","R6",2,"blue"
), byrow=TRUE))
colnames(Pedigri)=c("id","sire","dam", "sex","Estado")
Pedigri$sex=as.numeric(Pedigri$sex)
```

Generamos la base de datos *Genea* con la información de genealogía y sexo, con el comando ≪pedigree≫:

```
Genea=pedigree(id = Pedigri$id, dadid = Pedigri$sire,
    momid = Pedigri$dam, sex=as.numeric(Pedigri$sex))
```

Generamos ahora un gráfico con el árbol genealógico (FIGURA NRO. 12.1), identificando de color rojo individuos que hacen parte de los ancestros y que no están activos. Puede apreciarse que hay una línea negra gruesa entre los individuos $B3$ y $R2$, indicando que tienen una relación de parentesco y además son padres de $R5$:

```
plot(Genea, cex=0.5,col=Pedigri$Estado,
   mar=c(bottom=0, left=1, top=1, right=1))
```

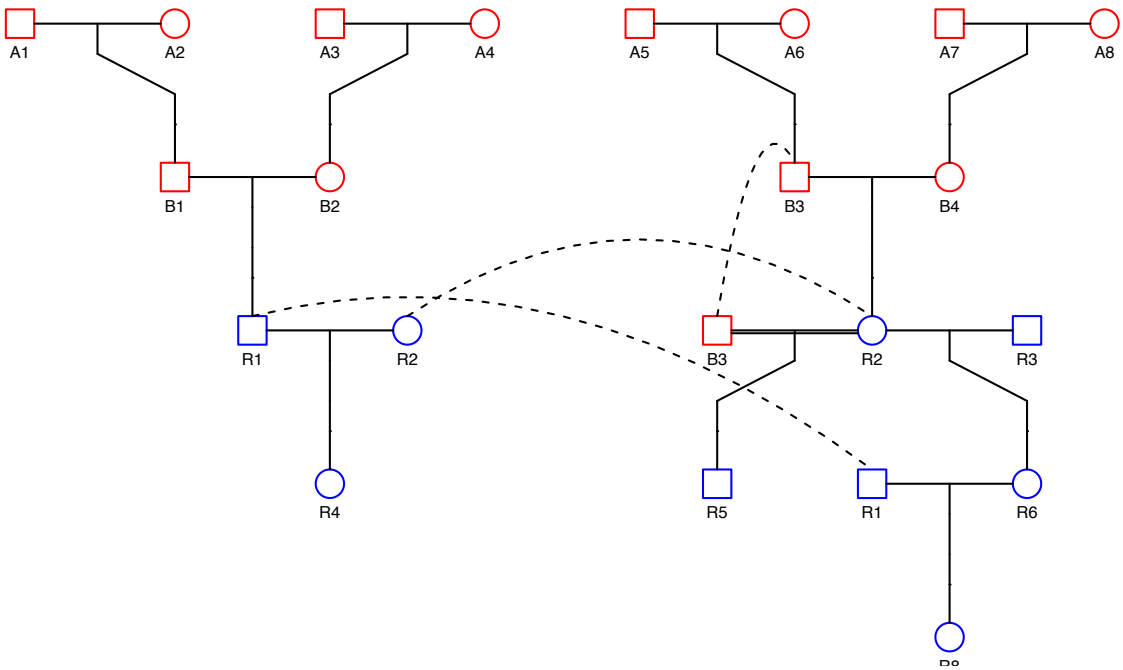

***Figura* 12.1:** Genealogía de los animales para el ejercicio de tamaño efectivo. Nota: genealogía con animales activos o aptos para reproducción (color azul) y animales que solamente hacen parte de la genealogía (color rojo). Fuente: elaboración propia generada en R-project [25].

Matriz de parentesco:

```
A=2*kinship(Pedigri$id,Pedigri$sire,Pedigri$dam);
A[6:13,2:13]
```

```
##       A2   A3   A4  A5  A6  A7  A8  B1  B2  B3  B4  R1
## A6 0.00 0.00 0.00 0.0 1.0 0.0 0.0 0.0 0.0 0.5 0.0 0.0
## A7 0.00 0.00 0.00 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.5 0.0
## A8 0.00 0.00 0.00 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.5 0.0
## B1 0.50 0.00 0.00 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.5
## B2 0.00 0.50 0.50 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.5
## B3 0.00 0.00 0.00 0.5 0.5 0.0 0.0 0.0 0.0 1.0 0.0 0.0
## B4 0.00 0.00 0.00 0.0 0.0 0.5 0.5 0.0 0.0 0.0 1.0 0.0
## R1 0.25 0.25 0.25 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 1.0
```

Para los cálculos del tamaño efectivo ecuación 12.1:

```
s=c("R1","R2","R3","R4","R5","R6","R8");s
```

```
## [1] "R1" "R2" "R3" "R4" "R5" "R6" "R8"
```

```
As=A[s,s];As
```

```
##      R1   R2   R3   R4   R5   R6   R8
## R1 1.0 0.00 0.00 0.50 0.00 0.00 0.50
## R2 0.0 1.00 0.00 0.50 0.75 0.50 0.25
## R3 0.0 0.00 1.00 0.00 0.00 0.50 0.25
## R4 0.5 0.50 0.00 1.00 0.38 0.25 0.38
## R5 0.0 0.75 0.00 0.38 1.25 0.38 0.19
## R6 0.0 0.50 0.50 0.25 0.38 1.00 0.50
## R8 0.5 0.25 0.25 0.38 0.19 0.50 1.00
```

```
f=as.matrix(diag(As)-1);f
```

```
##    [,1]
## R1 0.00
## R2 0.00
## R3 0.00
## R4 0.00
## R5 0.25
## R6 0.00
## R8 0.00
```

```
fprom=round(mean(f),2);fprom
```

```
## [1] 0.04
```

```
N=length(s);N
```

```
## [1] 7
```

```
Ne=round(N/(1+fprom),2);Ne
```

```
## [1] 6.7
```

Cálculo del tamaño efectivo con la ecuación 12.3:

```
C=kinship(Pedigri$id,Pedigri$sire,Pedigri$dam);C[1:7,1:7]
```

```
##     A1  A2  A3  A4  A5  A6  A7
## A1 0.5 0.0 0.0 0.0 0.0 0.0 0.0
## A2 0.0 0.5 0.0 0.0 0.0 0.0 0.0
## A3 0.0 0.0 0.5 0.0 0.0 0.0 0.0
## A4 0.0 0.0 0.0 0.5 0.0 0.0 0.0
## A5 0.0 0.0 0.0 0.0 0.5 0.0 0.0
## A6 0.0 0.0 0.0 0.0 0.0 0.5 0.0
## A7 0.0 0.0 0.0 0.0 0.0 0.0 0.5
```

```
machos <- Pedigri$sex==1 & (Pedigri$id %in% s)
machos  <- Pedigri$id[machos]
machos
```

```
## [1] "R1" "R3" "R5"
```

```
hembras <- Pedigri$sex==2 & (Pedigri$id %in% s)
hembras  <- Pedigri$id[hembras]
hembras
```

```
## [1] "R2" "R4" "R6" "R8"
```

```
C_jk=C[machos, hembras]
C_jk
```

```
##      R2   R4   R6    R8
## R1 0.00 0.25 0.00 0.250
## R3 0.00 0.00 0.25 0.125
## R5 0.38 0.19 0.19 0.094
```

```
Nd=length(machos)*length(hembras);Nd
```

```
## [1] 12
```

```
media_C_jk=round(sum(C_jk)/Nd,2);media_C_jk
```

```
## [1] 0.14
```

```
Ne_dosasteriscos=(length(machos)+length(hembras))/
  (1+media_C_jk)

Ne_dosasteriscos


## [1] 6.1


Nac=length(hembras); Nac


## [1] 4


Ned_dosasteriscosp=round(Nac*(1-media_C_jk),2)
Ned_dosasteriscosp


## [1] 3.4
```

```
#Apareamientos deseados
maximo=0.10
Deseados=ifelse(C_jk>=maximo,NA,C_jk)
Deseados


##    R2 R4 R6    R8
## R1  0 NA  0    NA
## R3  0  0 NA    NA
## R5 NA NA NA 0.094


Apareamientosdeseados=ifelse(C_jk>=maximo,NA,1)
Apareamientosdeseados


##    R2 R4 R6 R8
## R1  1 NA  1 NA
## R3  1  1 NA NA
## R5 NA NA NA  1


machose=as.matrix(apply(Apareamientosdeseados,1,
                        sum, na.rm=TRUE));machose


##    [,1]
## R1    2
```

```
## R3    2
## R5    1

machose=subset(machose,machose>0);machose

##    [,1]
## R1    2
## R3    2
## R5    1

hembrase=as.matrix(apply(Apareamientosdeseados,2,
                         sum, na.rm=TRUE))
hembrase=subset(hembras,hembrase>0);hembrase

## [1] "R2" "R4" "R6" "R8"

Deseados[rownames(machose),rownames(hembrase)]

##
## R1
## R3
## R5

Deseados

##    R2 R4 R6    R8
## R1  0 NA  0    NA
## R3  0  0 NA    NA
## R5 NA NA NA 0.094

coefdese=round(mean(Deseados,na.rm=TRUE),3);coefdese

## [1] 0.019

Nac1=length(hembrase); Nac1

## [1] 4

Nedese_dosasteriscos=round(Nac1/(1+coefdese),3)
Nedese_dosasteriscos

## [1] 3.9
```

13 CAPÍTULO TRECE

CONECTIVIDAD GENÉTICA ENTRE NIVELES DE EFECTO FIJO

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

Para una buena evaluación genética se requiere que exista conectividad genética entre los niveles de los efectos fijos, por ejemplo, es indispensable que un toro tenga hijas en varios rebaños para evitar confusión de la respuesta atribuida al efecto de la genética del toro y el efecto atribuido al ambiente de rebaño.

La medida más apropiada de conectividad es la varianza del error de predicción del promedio (VEP) de las diferencias en estimaciones de los valores de cría (EVC) entre animales en distintos niveles de factores no genéticos [102]. Cuando las comparaciones son entre niveles de un efecto fijo, la relación genética dentro de la unidad aumenta el PEV y la relación genética entre niveles disminuye el PEV. Por consiguiente, la matriz X^TZAZ^TX mide la suma de la relación genética dentro y entre niveles. Veamos el EJEMPLO 13, donde se tiene una genelalogía con individuos que tienen datos productivos en dos rebaños.

Para el desarrollo en R-project [25] utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos y ≪stringr≫ [30] para modificar los nombres de las columnas:

```
library(stringr)
library(kinship2)
library (MatrixModels)
```

Consideremos la siguiente base de información genealógica de 12 individuos, de los cuales seis son fundadores. Además, nueve animales tienen información productiva y están distribuidos en tres fincas (1, 2 y 3):

```
Base=data.frame(matrix(ncol=5,c(
"A1", NA, NA,1,NA,
"A2", NA, NA,2,"Finca 1",
"A3", NA, NA,1,NA,
"A4", NA, NA,2,"Finca 1",
"A5", NA, NA,1,NA,
"A6", NA, NA,2,"Finca 3",
"B1","A1","A2",1,"Finca 1",
"B2","A1","A2",2,"Finca 1",
"B3","A3","A4",1,"Finca 2",
"B4","A3","A4",2,"Finca 2",
"B5","A5","A6",1,"Finca 3",
"B6","A5","A6",2,"Finca 3"), byrow=TRUE))
colnames(Base)=c("id","sire","dam", "sex", "Finca")
Base$sex=as.numeric(Base$sex)

x=(pedigree(id = Base$id, dadid = Base$sire,
          momid = Base$dam, sex=Base$sex))
```

La matriz de parentesco sería:

```
A=2*kinship(Base$id,Base$sire,Base$dam)
A
```

```
##      A1  A2  A3  A4  A5  A6  B1  B2  B3  B4  B5  B6
## A1 1.0 0.0 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0
## A2 0.0 1.0 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0
## A3 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0
## A4 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0
## A5 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.5 0.5
## A6 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.5 0.5
## B1 0.5 0.5 0.0 0.0 0.0 0.0 1.0 0.5 0.0 0.0 0.0 0.0
## B2 0.5 0.5 0.0 0.0 0.0 0.0 0.5 1.0 0.0 0.0 0.0 0.0
## B3 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0 1.0 0.5 0.0 0.0
## B4 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0 0.5 1.0 0.0 0.0
## B5 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0 1.0 0.5
## B6 0.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 0.0 0.0 0.5 1.0
```

Con la siguiente programación obtendremos el árbol genealógico (FIGURA NRO. 13.1:

```
plot(x, paste(Base$id,"\n",Base$Finca), cex = 0.7,
   mar=c(bottom=0, left=1, top=1, right=1))
```

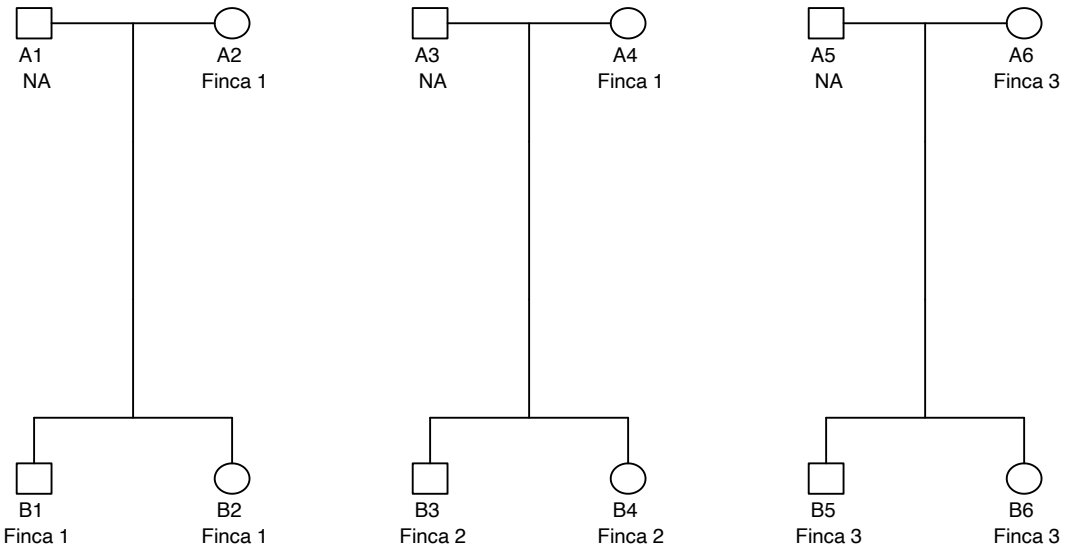

***Figura* 13.1:** Genealogía de los animales pertenecientes a tres fincas.
Fuente: elaboración propia generada en R-project [25].

Montaje de la matriz X:

```
Grupos=subset(Base, !is.na(Base$Finca))
X=as.matrix(model.Matrix(~ as.factor(Finca) -1,data=Grupos))
colnames(X)=word(colnames(X), 2, sep = fixed(')'))
rownames(X)=Grupos$id
X


##    Finca 1 Finca 2 Finca 3
```

```
## A2        1        0        0
## A4        1        0        0
## A6        0        0        1
## B1        1        0        0
## B2        1        0        0
## B3        0        1        0
## B4        0        1        0
## B5        0        0        1
## B6        0        0        1


Xp=t(X)
```

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Montaje de la matriz Z:

```
Pro=as.matrix(model.Matrix(~ as.factor(Grupos$id) -1))
colnames(Pro)=word(colnames(Pro), 2, sep = fixed(')'))
rownames(Pro)=Grupos$id
Z=matrix(nrow=nrow(Pro),ncol=ncol(A),0)
colnames(Z)=colnames(A)
rownames(Z)=colnames(Pro)
Z[colnames(Pro),colnames(Pro)]=Pro
Z


##    A1 A2 A3 A4 A5 A6 B1 B2 B3 B4 B5 B6
## A2  0  1  0  0  0  0  0  0  0  0  0  0
## A4  0  0  0  1  0  0  0  0  0  0  0  0
## A6  0  0  0  0  0  1  0  0  0  0  0  0
## B1  0  0  0  0  0  0  1  0  0  0  0  0
## B2  0  0  0  0  0  0  0  1  0  0  0  0
## B3  0  0  0  0  0  0  0  0  1  0  0  0
## B4  0  0  0  0  0  0  0  0  0  1  0  0
## B5  0  0  0  0  0  0  0  0  0  0  1  0
## B6  0  0  0  0  0  0  0  0  0  0  0  1
```

```
Zp=t(Z)
```

$$Z = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La deriva genética según el procedimiento de Kenney y Trus [102], se tendría:

```
DG=Xp%*%Z%*%A%*%Zp%*%X
DG
```

```
##          Finca 1 Finca 2 Finca 3
## Finca 1        7       1       0
## Finca 2        1       3       0
## Finca 3        0       0       6
```

$$X^T ZAZ^T X = \begin{bmatrix} 7 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

Los elementos diagonales son mayores que el número de animales debido a las relaciones dentro de los niveles de un efecto fijo, por consiguiente, creamos la matriz $\overline{A}$ con las relaciones promedio entre y dentro de los niveles de efecto fijo (incluyendo la relación de un animal consigo mismo). Los elementos de la diagonal de $\overline{A}$ se obtienen dividiendo los elementos de la diagonal de $X^T ZAZ^T X$ con el respectivo elemento de la diagonal de $X^T X$, y los elementos fuera de la diagonal de $\overline{A}$ se obtienen dividiendo el respectivo elemento fuera de la diagonal de $X^T ZAZ^T X$ con la multiplicación de los dos elementos de la diagonal de $X^T X$ que lo relacionan. Para un mejor entendimiento, veamos la siguiente matriz:

$$\overline{A} = \begin{bmatrix} \frac{7}{4^2} & \frac{1}{4x2} & \frac{0}{3x3} \\ \frac{1}{4x2} & \frac{3}{2^2} & \frac{0}{2x3} \\ \frac{0}{4x3} & \frac{0}{2x3} & \frac{6}{3^2} \end{bmatrix}$$

En R-project [25] crearemos una función donde se conjugan dos matrices: La matriz UNO corresponde a $X^T ZAZ^T X$ y la matriz DOS a $X^T X$. Cada elemento de la diagonal de la matriz UNO (i,i), se divide por el cuadrado del elemento correspondiente de la diagonal de la matriz DOS (i,i). Cada elemento fuera de la diagonal de la matriz UNO (i,j) se dividen por el producto de los elementos que corresponden a la matriz DOS (i,j) y (j,i):

```
PromRelac <- function(Uno,Dos){
  n=ncol(Uno)
  V <- matrix(nrow=nrow(Uno),ncol=ncol(Uno),0)
  for (i in 1:n){
    for(j in 1:n){
      if (i == j) {V[i,j]=Uno[i,i]/Dos[i,i]^2
      }else{V[i,j]=Uno[i,j]/(Dos[i,i]*Dos[j,j])}
    }
  }
  return(V)
}
XPX=Xp%*%X
Abarra=PromRelac(DG,XPX)
round(Abarra,2)


##      [,1] [,2] [,3]
## [1,] 0.44 0.12 0.00
## [2,] 0.12 0.75 0.00
## [3,] 0.00 0.00 0.67
```

$$\overline{A} = \begin{bmatrix} 0.44 & 0.12 & 0 \\ 0.12 & 0.75 & 0 \\ 0 & 0 & 0.67 \end{bmatrix}$$

Los elementos de $\overline{A}$ pueden interpretarse como los componentes genéticos de la varianza y la covarianza de la deriva entre los niveles del efecto fijo [103]. La varianza de la deriva genética entre dos niveles, estaría dada por la suma de los dos elementos diagonales, restándole el producto de los dos elementos fuera de la diagonal. Si el valor es bajo, indica alto grado de conectividad.

Crearemos una función para crear una matriz cuyos elementos fuera de la diagonal corresponden a la descripción anterior:

```
DG_UM <- function(TRES){
  n=ncol(TRES)
  V <- matrix(nrow=nrow(TRES),ncol=ncol(TRES),0)
  for (i in 1:n){
```

```
    for(j in 1:n){
      if (i == j) {V[i,j]=0
      }else{V[i,j]=TRES[i,i]+TRES[j,j]-2*TRES[i,j]}
    }
  }
  return(V)
}
```

Aplicamos la función:

```
Conectividad=DG_UM(Abarra)
diag(Conectividad) = NA
colnames(Conectividad)=rownames(Conectividad)= colnames(X)
triu(round(Conectividad,3))


## 3 x 3 Matrix of class "dtrMatrix"
##          Finca 1 Finca 2 Finca 3
## Finca 1       NA    0.94    1.10
## Finca 2        .      NA    1.42
## Finca 3        .       .      NA
```

Si tenemos en cuenta las varianzas genética y residual de qualquier característica analizada, podemos calcular la varianza de los niveles de un efecto fijo:

$$[X^TX - X^TZ(Z^TZ + A^{-1}\alpha)^{-1}Z^TX]^1\sigma_e^2$$

Si los niveles de un efecto fijo no estuvieran conectados, los elementos fuera de la diagonal de la matriz de varianza-covarianza serían cero. Los elementos fuera de la diagonal positivos son el resultado de relaciones genéticas entre animales en las diferentes niveles (conectividad genética):

```
vara=1
vare=1
alpha=vare/vara
VDGentreGrupos=solve((Xp%*%X) -(
  (Xp%*%Z)%*%(solve(Zp%*%Z + (
    solve(A)%x%alpha)))%*%Zp%*%X))%x%vare
round(VDGentreGrupos,2)


##      [,1] [,2] [,3]
## [1,] 0.67 0.17    0
## [2,] 0.17 1.17    0
## [3,] 0.00 0.00    1
```

```
Conectividad=DG_UM(VDGentreGrupos)

diag(Conectividad) = NA
colnames(Conectividad)=rownames(Conectividad)= colnames(X)
triu(round(Conectividad,3))


## 3 x 3 Matrix of class "dtrMatrix"
##          Finca 1 Finca 2 Finca 3
## Finca 1       NA     1.5     1.7
## Finca 2        .      NA     2.2
## Finca 3        .       .      NA
```

14 CAPÍTULO CATORCE

Matriz G^{-1} en modelo multirracial

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

Como lo describe Arnold et al [39], la inversa de la matriz de relaciones genéticas G^{-1} es el resultado de la combinaciones de A^{-1} y la inversa de las covarianzas G_0^{-1}. En una raza puede expresarse como el producto directo de estas dos matrices, así:

$$G^{-1} = A^{-1} \otimes G_0^{-1}$$

El el caso de la existencia de más de un grupo racial ($\phi = 1, 2.., k$) que presentan diferentes varianzas genéticas ($\sigma_1^2.., \sigma_k^2$), esta relación no es directa y G^{-1} estaría dada por un valor compuesto que depende linealmente de la composición del grupo racial del animal ($c_1\sigma_1^2, c_2\sigma_2^2,..,c_k\sigma_k^2$), donde c_k se refiere a la proporción de genes de la k-ésima raza y cuya sumatoria es igual que 1 ($c_1 + c_2 + .., + c_k = 1$). Esto evita expresar G^{-1} como un producto directo para datos de cruzamiento [104].

La construcción de G^{-1} ha sido ampliamente estudiada por Arnold et al [39], Elzo [104] y Lo et al [105] a partir de las reglas de Henderson [7]. Entretanto, existen diversos criterios para construirla, por ejemplo, Poulsen [106] comparó cuatro métodos que usan diferentes matrices de relaciones multirraciales en términos de capacidad para predecir valores genéticos precisos e imparciales con fenotipos de animales cruzados (en especial cuando se emplean modelos genómicos y multirraciales), los cuales se describen a continuación:

Matrices de relaciones del numerador (NRM): este método secciona la matriz de parentesco de la población en matrices de parentesco para cada raza pura, con su respectiva varianza genética aditiva, lo que permite dividir los valores genéticos de los animales mestizos en términos de la proporción de las razas puras. El vector de efectos genéticos aditivos tendría la siguiente estructura para dos razas ($R1$ y $R2$):

$$\begin{bmatrix} a_{R1} \\ a_{R2} \end{bmatrix} =\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{R1} A_{R1} & 0 \\ \sigma^2_{R2} 0 & A_{R2} \end{bmatrix}\right)$$

Para Poulsen et al [106], este método tiene las suposiciones inexactas para los efectos genéticos aditivos, si las variaciones genéticas aditivas parciales debidas a efectos específicos de la raza no son proporcionales a la composición de la raza.

El método de relaciones parciales (GT): este método propuesto por García-Cortés y Toro [107], divide la relación genética aditiva en varias matrices de relaciones parciales para las razas puras y para cada par de razas (matrices de parentesco parciales para términos de segregación). Continuando con el ejemplo de la estructura para dos razas ($R1$ y $R2$), se tendría:

$$\begin{bmatrix} a_{R1} \\ a_{R2} \\ a_{R1R2} \end{bmatrix} =\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{R1} A_{R1} & 0 & 0 \\ 0 & \sigma^2_{R2} A_{R2} & 0 \\ 0 & 0 & \sigma^2_{R1R2} A_{R1R2} \end{bmatrix}\right)$$

Método GT modificado (SM): este método es una aproximación del método GT, propuesto por Strandén y Mäntysaari [108] y reparte la varianza genética aditiva de la misma forma. Las varianzas genéticas para las razas y cruces son las mismas que el método GT, pero las matrices A_1, A_2.., A_k serían aproximadamente similares. Su diferencia estaría dada por la inclusión de matrices que tienen una diagonal conformada por la raiz cuadrada de las proporciones raciales de cada individuo (la llamaremos matriz Φ), por consiguiente, el vector de efectos genéticos para el caso de dos razas, estaría establecido de las siguiente forma:

$$\begin{bmatrix} a_{R1} \\ a_{R2} \\ a_{R1R2} \end{bmatrix} =\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{R1} \Phi_{R1} A_{R1} \Phi_{R1} & 0 & 0 \\ 0 & \sigma^2_{R2} \Phi_{R2} A_{R2} \Phi_{R2} & 0 \\ 0 & 0 & \sigma^2_{R1R2} \Phi_{R1R2} A_{R1R2} \Phi_{R1R2} \end{bmatrix}\right)$$

Método Metafundadores (MF): este método modela poblaciones como subpoblaciones derivadas de una población ancestral común. En la práctica, esto se hace identificando cada subpoblación a través de un metafundador, calculando una matriz de relación genética aditiva entre metafundadores (Γ), y luego incorporando esta información a A, compartida para todas las poblaciones [106], así:

$$\Gamma = \begin{bmatrix} \gamma_{R1} & \gamma_{R1R2} \\ \gamma_{R2R1} & \gamma_{R2} \end{bmatrix}$$

Donde γ_{R1} es la relación del metafundador de la subpoblación $R1$; γ_{R2} es la relación metafundador para la subpoblación $R2$ y γ_{R1R2} es la relación metafundadora entre las subpoblaciones $R1$ y $R2$. El vector de efectos genéticos seguiría una distribución dada por:

$$a =\sim \mathcal{N}\left(0, \sigma^2_{mf} A\Gamma\right)$$

Donde σ^2_{mf} es la varianza genética aditiva ancestral común en la población.

En el trabajo de Poulsen et al [106] en poblaciones en cruzamiento rotativo, todos los métodos fueron casi igualmente precisos para la predicción de valores genéticos en animales de raza pura; sin embargo, con información genómica, los métodos MF y GT fueron los más precisos, concluyendo que existe la necesidad de investigar cómo se comparan los modelos con estas matrices de relación, para predecir valores de reproducción precisos e imparciales con fenotipos de animales cruzados por rotación.

El método NRM es un enfoque común para los análisis de razas múltiples. El argumento principal para el método NRM es que comúnmente se implementa en software para evaluaciones genéticas. Sin embargo, se presentan resultados erróneos en los valores genéticos en animales mestizos. Los elementos diagonales para los animales mestizos no están escalados según las proporciones de su raza, y este error afecta tanto a los elementos diagonales como fuera de la diagonal para los descendientes de los animales mestizos. El método SM produce los mismos elementos diagonales que el método GT en ausencia de consanguinidad, punto fundamental, porque a pesar de ser poblaciones cruzadas, el número de fundadores es muy reducido y por la dinámica de los apareamientos, al paso de las generaciones, hay aumento de la consanguinidad.

Como observaron, existen diversas controversias sobre la construcción de G^{-1}, y además, si se ignoran las diferencias genéticas entre poblaciones, se tendrían estimaciones de parámetros sesgados, en particular para la varianza genética aditiva. Frente a esta dualidad, nos enfocaremos al trabajo de Muff et al [109], quienes aplicaron la descomposición de Cholesky a la matriz A y escalonaron los componentes de la matriz Φ que contiene las proporciones respectivas para cada raza; y advierten que a pesar de ser un método conveniente, es aproximado. Con base al método GT, García-Cortés y Toro [107] dividieron de forma aditiva el valor genético total de cada individuo en componentes específicos de la raza, que covarían de acuerdo con una

matriz de parentesco específica de la raza. La descomposición de A, está dada por las multiplicación de las matrices LDL^T, donde L es una matriz triangular inferior con 1 en la diagonal y D es una matriz diagonal.

Nosotros realizaremos modificaciones a lo propuesto por Muff et al [109] en la matriz Φ. Los anteriores autores indicaron que Φ es una matriz con las proporciones de cada raza que tiene cada individuo, en nuestro caso, Φ es una matriz de diseño que relaciona los individuos con los grupos genéticos.

Para el desarrollo de esta propuesta de aproximación, utilizaremos el EJEMPLO 14 con la genealogía de la TABLA NRO. 14.1 con individuos de dos razas ($R1$ y $R2$) y el cruzamiento entre ellas (RC). La genealogía tiene 9 individuos, de los cuales 5 no tienen ancestros y pertenecen a las razas puras (1, 2, 3 y 4) y uno es cruzado (6). Hay 4 animales con ascendencia conocida, dos de los cuales puros y dos son cruzados. Cabe destacar que el animal 9 es cruzado, pero también es endogámico porque es hijo del animal 1, el cual es su abuelo materno.

TABLA 14.1: Proporción racial y grupos genéticos de 9 animales puros y cruzados

id	padre	madre	Proporción		Grupo genético	Φ		
			c_{R1}	c_{R2}		ϕ_{R1}	ϕ_{R2}	ϕ_{RC}
1			1	0	R1	1	0	0
2			1	0	R1	1	0	0
3			0	1	R2	0	1	0
4			0	1	R2	0	1	0
5	1	2	1	0	R1	1	0	0
6	3	4	.5	.5	RC	0	0	1
7			0	1	R2	0	1	0
8	1	6	.75	.25	RC	0	0	1
9	1	8	.875	.125	RC	0	0	1

Nota: $c_{R1,R2}$ son las proporciones raciales de cada individuo y Φ relaciona los individuos con el grupo racial a que pertenece (matriz Φ).
Fuente: elaboración propia (2024).

La matriz de parentesco sería:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0.75 \\ 0 & 1 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0.25 & 0.38 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.25 & 0.5 & 0 & 1 & 0.75 \\ 0.75 & 0 & 0 & 0 & 0.38 & 0.25 & 0 & 0.75 & 1.25 \end{bmatrix}$$

La descomposición de Cholesky de la matriz de parentesco sería:

$$A_d = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 1 & 0 \\ 0.75 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.5 & 1 \end{bmatrix}$$

Por consiguiente, la matriz L y la matriz con la diagonal D, serían:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 1 & 0 \\ 0.75 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.5 & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

Si consideramos las varianzas de cada raza:

$$\sigma^2_{R1} = 10$$

$$\sigma^2_{R2} = 5$$

$$\sigma^2_{RC} = 7.5$$

Para cada grupo racial L y D, serían:

$$L_{R1} = L(I \otimes \Phi_{R1}) \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D_{R1} = \sigma^2_{R1} D(I \otimes \Phi_{R1}) = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$L_{R2} = L(I \otimes \Phi_{R2}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D_{R2} = D\sigma^2_{R2}(I \otimes \Phi_{R2}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$L_{RC} = L(I \otimes \Phi_{RC}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0.5 & 1 \end{bmatrix}$$

$$D_{RC} = \sigma^2_{RC} D(I \otimes \Phi_{RC}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.75 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.75 \end{bmatrix}$$

Las matrices G de cada grupo genético serían:

$$G_{R1} = L_{R1} D_{R1} L_{R1}^T = \begin{bmatrix} 10 & 0 & 0 & 0 & 5 & 0 & 0 & 5 & 7.5 \\ 0 & 10 & 0 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 5 & 0 & 0 & 10 & 0 & 0 & 2.5 & 3.75 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 2.5 & 0 & 0 & 2.5 & 3.75 \\ 7.5 & 0 & 0 & 0 & 3.75 & 0 & 0 & 3.75 & 5.62 \end{bmatrix}$$

$$G_{R2} = L_{R2} D_{R2} L_{R2}^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 2.5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 2.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.5 & 2.5 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$G_{RC} = L_{RC} D_{RC} L_{RC}^{T} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7.5 & 0 & 3.75 & 1.88 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.75 & 0 & 5.62 & 2.81 \\ 0 & 0 & 0 & 0 & 0 & 1.88 & 0 & 2.81 & 5.16 \end{bmatrix}$$

G sería la sumatoria de las matrices de relaciones de cada grupo racial, así:

$$G = G_{R1} + G_{R2} + G_{RC} = \begin{bmatrix} 10 & 0 & 0 & 0 & 5 & 0 & 0 & 5 & 7.5 \\ 0 & 10 & 0 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 2.5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 2.5 & 0 & 0 \\ 5 & 5 & 0 & 0 & 10 & 0 & 0 & 2.5 & 3.75 \\ 0 & 0 & 0 & 0 & 0 & 7.5 & 0 & 3.75 & 1.88 \\ 0 & 0 & 2.5 & 2.5 & 0 & 0 & 5 & 0 & 0 \\ 5 & 0 & 0 & 0 & 2.5 & 3.75 & 0 & 8.12 & 6.56 \\ 7.5 & 0 & 0 & 0 & 3.75 & 1.88 & 0 & 6.56 & 10.78 \end{bmatrix}$$

$$G^{-1} = \begin{bmatrix} 0.28 & 0.05 & 0 & 0 & -0.1 & 0.07 & 0 & -0.07 & -0.13 \\ 0.05 & 0.15 & 0 & 0 & -0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.1 & 0 & 0 & -0.2 & 0 & 0 \\ 0 & 0 & 0.1 & 0.3 & 0 & 0 & -0.2 & 0 & 0 \\ -0.1 & -0.1 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 \\ 0.07 & 0 & 0 & 0 & 0 & 0.2 & 0 & -0.13 & 0 \\ 0 & 0 & -0.2 & -0.2 & 0 & 0 & 0.4 & 0 & 0 \\ -0.07 & 0 & 0 & 0 & 0 & -0.13 & 0 & 0.33 & -0.13 \\ -0.13 & 0 & 0 & 0 & 0 & 0 & 0 & -0.13 & 0.27 \end{bmatrix}$$

14.0.1. Ejercicios en R-project

La descomposición de Cholesky la realizaremos con el comando *gchol*, de la librería ≪bdsmatrix≫ [110] y para extraer los componentes de la descomposición se utilizaron las funciones diag y as.matrix, también utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
Base=data.frame(matrix(ncol=4,byrow=TRUE, c(
1,NA,NA,"R1",
2,NA,NA,"R1",
3,NA,NA,"R2",
4,NA,NA,"R2",
5, 1, 2,"R1",
6,NA,NA,"RC",
7, 3, 4,"R2",
8, 1, 6,"RC",
9, 1, 8,"RC"
)))
colnames(Base)=c("id", "sire", "dam","GrupoGen")
rownames (Base)=seq(1,nrow(Base),1)
head(Base)
```

```
##   id sire  dam GrupoGen
## 1  1 <NA> <NA>       R1
## 2  2 <NA> <NA>       R1
## 3  3 <NA> <NA>       R2
## 4  4 <NA> <NA>       R2
## 5  5    1    2       R1
## 6  6 <NA> <NA>       RC
```

```
attach(Base)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     GrupoGen
```

```
library(kinship2)
A=2*kinship(id,sire,dam);colnames(A)=rownames(A)=Base$id;A
```

```
##      1   2   3   4    5    6   7    8    9
```

```
## 1 1.00 0.0 0.0 0.0 0.50 0.00 0.0 0.50 0.75
## 2 0.00 1.0 0.0 0.0 0.50 0.00 0.0 0.00 0.00
## 3 0.00 0.0 1.0 0.0 0.00 0.00 0.5 0.00 0.00
## 4 0.00 0.0 0.0 1.0 0.00 0.00 0.5 0.00 0.00
## 5 0.50 0.5 0.0 0.0 1.00 0.00 0.0 0.25 0.38
## 6 0.00 0.0 0.0 0.0 0.00 1.00 0.0 0.50 0.25
## 7 0.00 0.0 0.5 0.5 0.00 0.00 1.0 0.00 0.00
## 8 0.50 0.0 0.0 0.0 0.25 0.50 0.0 1.00 0.75
## 9 0.75 0.0 0.0 0.0 0.38 0.25 0.0 0.75 1.25
```

```
library (MatrixModels)
Grup=as.matrix(model.Matrix(~ as.factor(Base$GrupoGen) -1))
library(stringr)
colnames(Grup)=word(colnames(Grup), 2, sep = fixed(')'))
rownames(Grup)=Base$id
Grup
```

```
##   R1 R2 RC
## 1  1  0  0
## 2  1  0  0
## 3  0  1  0
## 4  0  1  0
## 5  1  0  0
## 6  0  0  1
## 7  0  1  0
## 8  0  0  1
## 9  0  0  1
```

```
Phi=matrix(nrow=nrow(Base),ncol=ncol(Grup),0)
colnames(Phi)=colnames(Grup)
rownames(Phi)=Base$id
Phi[rownames(Grup),colnames(Grup)]=
  Grup[rownames(Grup),colnames(Grup)]
Phi
```

```
##   R1 R2 RC
## 1  1  0  0
## 2  1  0  0
## 3  0  1  0
## 4  0  1  0
## 5  1  0  0
## 6  0  0  1
## 7  0  1  0
```

```
## 8  0  0  1
## 9  0  0  1
```

```
#considerando la misma varianza para
#todos los grupos genéticos
var=10;var
```

```
## [1] 10
```

```
G=A%x%var;G
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,] 10.0    0    0    0  5.0  0.0    0  5.0  7.5
##  [2,]  0.0   10    0    0  5.0  0.0    0  0.0  0.0
##  [3,]  0.0    0   10    0  0.0  0.0    5  0.0  0.0
##  [4,]  0.0    0    0   10  0.0  0.0    5  0.0  0.0
##  [5,]  5.0    5    0    0 10.0  0.0    0  2.5  3.8
##  [6,]  0.0    0    0    0  0.0 10.0    0  5.0  2.5
##  [7,]  0.0    0    5    5  0.0  0.0   10  0.0  0.0
##  [8,]  5.0    0    0    0  2.5  5.0    0 10.0  7.5
##  [9,]  7.5    0    0    0  3.8  2.5    0  7.5 12.5
```

```
library(MASS)
Gin=ginv(G);round(Gin,2)
```

```
##       [,1]  [,2]  [,3]  [,4] [,5]  [,6] [,7]  [,8] [,9]
##  [1,]  0.25  0.05  0.00  0.00 -0.1  0.05  0.0 -0.05 -0.1
##  [2,]  0.05  0.15  0.00  0.00 -0.1  0.00  0.0  0.00  0.0
##  [3,]  0.00  0.00  0.15  0.05  0.0  0.00 -0.1  0.00  0.0
##  [4,]  0.00  0.00  0.05  0.15  0.0  0.00 -0.1  0.00  0.0
##  [5,] -0.10 -0.10  0.00  0.00  0.2  0.00  0.0  0.00  0.0
##  [6,]  0.05  0.00  0.00  0.00  0.0  0.15  0.0 -0.10  0.0
##  [7,]  0.00  0.00 -0.10 -0.10  0.0  0.00  0.2  0.00  0.0
##  [8,] -0.05  0.00  0.00  0.00  0.0 -0.10  0.0  0.25 -0.1
##  [9,] -0.10  0.00  0.00  0.00  0.0  0.00  0.0 -0.10  0.2
```

```
library(bdsmatrix)
```

```
##
## Attaching package: 'bdsmatrix'
```

```
## The following object is masked from 'package:base':
##
##     backsolve

#Descomposición de Cholesky para A, denominada Ad
A_d=gchol(A);A_d
```

```
##      1   2   3   4   5    6   7   8   9
## 1 1.00 0.0 0.0 0.0 0.0 0.00 0.0 0.0 0.0
## 2 0.00 1.0 0.0 0.0 0.0 0.00 0.0 0.0 0.0
## 3 0.00 0.0 1.0 0.0 0.0 0.00 0.0 0.0 0.0
## 4 0.00 0.0 0.0 1.0 0.0 0.00 0.0 0.0 0.0
## 5 0.50 0.5 0.0 0.0 0.5 0.00 0.0 0.0 0.0
## 6 0.00 0.0 0.0 0.0 0.0 1.00 0.0 0.0 0.0
## 7 0.00 0.0 0.5 0.5 0.0 0.00 0.5 0.0 0.0
## 8 0.50 0.0 0.0 0.0 0.0 0.50 0.0 0.5 0.0
## 9 0.75 0.0 0.0 0.0 0.0 0.25 0.0 0.5 0.5
```

```
D=matrix(ncol=ncol(A), nrow=ncol(A),0);diag(D)=diag(A_d);D
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,]    1    0    0    0  0.0    0  0.0  0.0  0.0
##  [2,]    0    1    0    0  0.0    0  0.0  0.0  0.0
##  [3,]    0    0    1    0  0.0    0  0.0  0.0  0.0
##  [4,]    0    0    0    1  0.0    0  0.0  0.0  0.0
##  [5,]    0    0    0    0  0.5    0  0.0  0.0  0.0
##  [6,]    0    0    0    0  0.0    1  0.0  0.0  0.0
##  [7,]    0    0    0    0  0.0    0  0.5  0.0  0.0
##  [8,]    0    0    0    0  0.0    0  0.0  0.5  0.0
##  [9,]    0    0    0    0  0.0    0  0.0  0.0  0.5
```

```
L=as.matrix(A_d); L
```

```
##      1   2   3   4 5    6 7   8 9
## 1 1.00 0.0 0.0 0.0 0 0.00 0 0.0 0
## 2 0.00 1.0 0.0 0.0 0 0.00 0 0.0 0
## 3 0.00 0.0 1.0 0.0 0 0.00 0 0.0 0
## 4 0.00 0.0 0.0 1.0 0 0.00 0 0.0 0
## 5 0.50 0.5 0.0 0.0 1 0.00 0 0.0 0
## 6 0.00 0.0 0.0 0.0 0 1.00 0 0.0 0
## 7 0.00 0.0 0.5 0.5 0 0.00 1 0.0 0
## 8 0.50 0.0 0.0 0.0 0 0.50 0 1.0 0
## 9 0.75 0.0 0.0 0.0 0 0.25 0 0.5 1
```

```
comprobaciónCHOL=L%*%D%*%t(L);A-comprobaciónCHOL#son iguales
```

```
##    1 2 3 4 5 6 7 8 9
## 1  0 0 0 0 0 0 0 0 0
## 2  0 0 0 0 0 0 0 0 0
## 3  0 0 0 0 0 0 0 0 0
## 4  0 0 0 0 0 0 0 0 0
## 5  0 0 0 0 0 0 0 0 0
## 6  0 0 0 0 0 0 0 0 0
## 7  0 0 0 0 0 0 0 0 0
## 8  0 0 0 0 0 0 0 0 0
## 9  0 0 0 0 0 0 0 0 0
```

```
#incluyendo varianzas diferentes a los grupos raciales
var_R1=matrix(10)
var_R2=matrix(5)
var_RC=matrix(7.5)
```

```
L1=L%*%diag(Phi[,1]);colnames(L1)=Base$id;L1
```

```
##       1   2 3 4 5 6 7 8 9
## 1  1.00 0.0 0 0 0 0 0 0 0
## 2  0.00 1.0 0 0 0 0 0 0 0
## 3  0.00 0.0 0 0 0 0 0 0 0
## 4  0.00 0.0 0 0 0 0 0 0 0
## 5  0.50 0.5 0 0 1 0 0 0 0
## 6  0.00 0.0 0 0 0 0 0 0 0
## 7  0.00 0.0 0 0 0 0 0 0 0
## 8  0.50 0.0 0 0 0 0 0 0 0
## 9  0.75 0.0 0 0 0 0 0 0 0
```

```
L2=L%*%diag(Phi[,2]);colnames(L2)=Base$id;L2
```

```
##    1 2   3   4 5 6 7 8 9
## 1  0 0 0.0 0.0 0 0 0 0 0
## 2  0 0 0.0 0.0 0 0 0 0 0
## 3  0 0 1.0 0.0 0 0 0 0 0
## 4  0 0 0.0 1.0 0 0 0 0 0
## 5  0 0 0.0 0.0 0 0 0 0 0
## 6  0 0 0.0 0.0 0 0 0 0 0
## 7  0 0 0.5 0.5 0 0 1 0 0
## 8  0 0 0.0 0.0 0 0 0 0 0
## 9  0 0 0.0 0.0 0 0 0 0 0
```

```
L3=L%*%diag(Phi[,3]);colnames(L3)=Base$id;L3
```

```
##   1 2 3 4 5    6 7   8 9
## 1 0 0 0 0 0 0.00 0 0.0 0
## 2 0 0 0 0 0 0.00 0 0.0 0
## 3 0 0 0 0 0 0.00 0 0.0 0
## 4 0 0 0 0 0 0.00 0 0.0 0
## 5 0 0 0 0 0 0.00 0 0.0 0
## 6 0 0 0 0 0 1.00 0 0.0 0
## 7 0 0 0 0 0 0.00 0 0.0 0
## 8 0 0 0 0 0 0.50 0 1.0 0
## 9 0 0 0 0 0 0.25 0 0.5 1
```

```
D1=D%*%diag(Phi[,1])%x%var_R1
rownames(D1)=colnames(D1)=Base$id;D1
```

```
##    1  2 3 4 5 6 7 8 9
## 1 10  0 0 0 0 0 0 0 0
## 2  0 10 0 0 0 0 0 0 0
## 3  0  0 0 0 0 0 0 0 0
## 4  0  0 0 0 0 0 0 0 0
## 5  0  0 0 0 5 0 0 0 0
## 6  0  0 0 0 0 0 0 0 0
## 7  0  0 0 0 0 0 0 0 0
## 8  0  0 0 0 0 0 0 0 0
## 9  0  0 0 0 0 0 0 0 0
```

```
D2=D%*%diag(Phi[,2])%x%var_R2
rownames(D2)=colnames(D2)=Base$id;D2
```

```
##   1 2 3 4 5 6   7 8 9
## 1 0 0 0 0 0 0 0.0 0 0
## 2 0 0 0 0 0 0 0.0 0 0
## 3 0 0 5 0 0 0 0.0 0 0
## 4 0 0 0 5 0 0 0.0 0 0
## 5 0 0 0 0 0 0 0.0 0 0
## 6 0 0 0 0 0 0 0.0 0 0
## 7 0 0 0 0 0 0 2.5 0 0
## 8 0 0 0 0 0 0 0.0 0 0
## 9 0 0 0 0 0 0 0.0 0 0
```

```
D3=D%*%diag(Phi[,3])%x%var_RC
rownames(D3)=colnames(D3)=Base$id;D3
```

```
##   1 2 3 4 5   6 7   8   9
## 1 0 0 0 0 0 0.0 0 0.0 0.0
## 2 0 0 0 0 0 0.0 0 0.0 0.0
## 3 0 0 0 0 0 0.0 0 0.0 0.0
## 4 0 0 0 0 0 0.0 0 0.0 0.0
## 5 0 0 0 0 0 0.0 0 0.0 0.0
## 6 0 0 0 0 0 7.5 0 0.0 0.0
## 7 0 0 0 0 0 0.0 0 0.0 0.0
## 8 0 0 0 0 0 0.0 0 3.8 0.0
## 9 0 0 0 0 0 0.0 0 0.0 3.8
```

```
G_R1=L1%*%(D1)%*%t(L1);G_R1
```

```
##      1  2 3 4    5 6 7   8   9
## 1 10.0  0 0 0  5.0 0 0 5.0 7.5
## 2  0.0 10 0 0  5.0 0 0 0.0 0.0
## 3  0.0  0 0 0  0.0 0 0 0.0 0.0
## 4  0.0  0 0 0  0.0 0 0 0.0 0.0
## 5  5.0  5 0 0 10.0 0 0 2.5 3.8
## 6  0.0  0 0 0  0.0 0 0 0.0 0.0
## 7  0.0  0 0 0  0.0 0 0 0.0 0.0
## 8  5.0  0 0 0  2.5 0 0 2.5 3.8
## 9  7.5  0 0 0  3.8 0 0 3.8 5.6
```

```
G_R2=L2%*%(D2)%*%t(L2);G_R2
```

```
##   1 2   3   4 5 6   7 8 9
## 1 0 0 0.0 0.0 0 0 0.0 0 0
## 2 0 0 0.0 0.0 0 0 0.0 0 0
## 3 0 0 5.0 0.0 0 0 2.5 0 0
## 4 0 0 0.0 5.0 0 0 2.5 0 0
## 5 0 0 0.0 0.0 0 0 0.0 0 0
## 6 0 0 0.0 0.0 0 0 0.0 0 0
## 7 0 0 2.5 2.5 0 0 5.0 0 0
## 8 0 0 0.0 0.0 0 0 0.0 0 0
## 9 0 0 0.0 0.0 0 0 0.0 0 0
```

```
G_RC=L3%*%(D3)%*%t(L3);G_RC
```

```
##   1 2 3 4 5   6 7   8   9
## 1 0 0 0 0 0 0.0 0 0.0 0.0
## 2 0 0 0 0 0 0.0 0 0.0 0.0
## 3 0 0 0 0 0 0.0 0 0.0 0.0
## 4 0 0 0 0 0 0.0 0 0.0 0.0
## 5 0 0 0 0 0 0.0 0 0.0 0.0
## 6 0 0 0 0 0 7.5 0 3.8 1.9
## 7 0 0 0 0 0 0.0 0 0.0 0.0
## 8 0 0 0 0 0 3.8 0 5.6 2.8
## 9 0 0 0 0 0 1.9 0 2.8 5.2
```

```
G_d=G_R1+G_R2+G_RC;G_d
```

```
##      1  2   3   4    5   6   7   8    9
## 1 10.0  0 0.0 0.0  5.0 0.0 0.0 5.0  7.5
## 2  0.0 10 0.0 0.0  5.0 0.0 0.0 0.0  0.0
## 3  0.0  0 5.0 0.0  0.0 0.0 2.5 0.0  0.0
## 4  0.0  0 0.0 5.0  0.0 0.0 2.5 0.0  0.0
## 5  5.0  5 0.0 0.0 10.0 0.0 0.0 2.5  3.8
## 6  0.0  0 0.0 0.0  0.0 7.5 0.0 3.8  1.9
## 7  0.0  0 2.5 2.5  0.0 0.0 5.0 0.0  0.0
## 8  5.0  0 0.0 0.0  2.5 3.8 0.0 8.1  6.6
## 9  7.5  0 0.0 0.0  3.8 1.9 0.0 6.6 10.8
```

```
Gin=ginv(G_d);round(Gin,2)
```

```
##        [,1]  [,2] [,3] [,4] [,5]  [,6] [,7]  [,8]  [,9]
##  [1,]  0.28  0.05  0.0  0.0 -0.1  0.07  0.0 -0.07 -0.13
##  [2,]  0.05  0.15  0.0  0.0 -0.1  0.00  0.0  0.00  0.00
##  [3,]  0.00  0.00  0.3  0.1  0.0  0.00 -0.2  0.00  0.00
##  [4,]  0.00  0.00  0.1  0.3  0.0  0.00 -0.2  0.00  0.00
##  [5,] -0.10 -0.10  0.0  0.0  0.2  0.00  0.0  0.00  0.00
##  [6,]  0.07  0.00  0.0  0.0  0.0  0.20  0.0 -0.13  0.00
##  [7,]  0.00  0.00 -0.2 -0.2  0.0  0.00  0.4  0.00  0.00
##  [8,] -0.07  0.00  0.0  0.0  0.0 -0.13  0.0  0.33 -0.13
##  [9,] -0.13  0.00  0.0  0.0  0.0  0.00  0.0 -0.13  0.27
```

15 CAPÍTULO QUINCE

EVALUACIONES GENÓMICAS

Carlos Alberto Martínez Niño
Universidad Nacional de Colombia, sede Bogotá

Mario Fernando Cerón-Muñoz
Universidad de Antioquia

15.1. No identificabilidad de los efectos de los marcadores moleculares

Esta sección muestra el procedimiento para establecer la no identificabilidad de algunos efectos individuales de los marcadores moleculares en un modelo de regresión, a través del genoma basado en el procedimiento descrito por Gianola [54], desde una perspectiva bayesiana.

En el caso bayesiano, no existe una única noción de identificabilidad como en el caso frecuentista, la siguiente es la que se emplea aquí [111, 112]. Consideremos un modelo bayesiano con parámetro de localización θ particionado como $\theta = (\theta_1, \theta_2)$ y función de verosimilitud $f(y|\theta_1, \theta_2)$. El parámetro θ_2 se dice no identificable si su distribución condicional posterior es tal que:

$$f(\theta_2|\theta_1, y) = f(\theta_2|\theta_1)$$

Nótese que en este caso los datos y, no brindan información sobre θ_2 más allá de la que provee θ_1, hay una independencia condicional entre θ_2 y los datos dado θ_1. Bajo el modelo de regresión a través del genoma presentado en el capítulo 7, cuando

$p > n$, el rango de la matriz Z es a lo más n y en consecuencia no es de rango columna completo. Típicamente esta matriz tendrá rango n, así que se continúa con este supuesto. Entonces, podemos permutar las columnas de Z de la siguiente forma:

$$Z = [Z_1 Z_2]$$

Z_1 es de dimensión nxn de rango completo y Z_2 de dimensión nx$(p-n)$. Llevamos a cabo la misma permutación en el vector de efectos aditivos de los marcadores, esto es:

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Ahora definimos nuevos parámetros $\theta_1 = Zu = Z_1 u_1 + Z_2 u_2$ y $\theta_2 = u_2$, esto puede escribirse matricialmente así:

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \\ 0 & I_{(n-p)} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Estos parámetros corresponden al vector de valores genéticos aditivos (θ_1) y al vector de efectos aditivos de $p-n$ marcadores θ_2 correspondientes a la matriz Z_2. El objetivo es mostrar que $\theta_2 = u_2$ no es identificable en el sentido bayesiano definido arriba porque dado θ_1 (los valores genéticos aditivos de los n individuos evaluados), los datos no proveen más información sobre los efectos de los marcadores que están en u_2. Debido a que $\theta_1 = Zu$, el modelo en términos de los nuevos parámetros sería $y = \theta_1 + e$. A continuación, se formaliza este resultado. Como Z_1 es invertible, la transformación inversa existe y tiene la forma:

$$u_2 = \theta_2$$

$$u_1 = Z_1^{-1}(\theta_1 - Z_2\theta_2)$$

Utilizando la transformación inversa, podemos reparametrizar el modelo así:

$$y = Zu + e$$

$$= [Z_1 Z_2] \begin{bmatrix} Z_1^{-1}(\theta_1 - Z_2\theta_2) \\ \theta_2 \end{bmatrix} + e$$

$$= Z_1 Z_1^{-1}(\theta_1 - Z_2\theta_2) + Z_2\theta_2 + e$$

$$= \theta_1 - Z_2\theta_2 + Z_2\theta_2 + e$$

$$= \theta_1 + e$$

Esto indica que los datos no contienen información sobre θ_2 más allá de la que está en θ_1 y así, tenemos independencia condicional de y y θ_2 dado θ_1; es decir, $f(y|\theta_1, \theta_2) = f(y|\theta_1)$. Aplicando este resultado (tercera línea del siguiente procedimiento) se tiene:

$$f(\theta_2|\theta_1, y) = \frac{f(\theta_1, \theta_2, y)}{f(\theta_1, y)}$$

$$= \frac{f(y|\theta_1, \theta_2)f(\theta_2|\theta_1)f(\theta_1)}{f(y|\theta_1)f(\theta_1)}$$

$$= \frac{f(y|\theta_1)f(\theta_2|\theta_1)}{f(y|\theta_1)}$$

$$= f(\theta_2|\theta_1)$$

De acuerdo con la definición que se presentó antes, este resultado muestra que θ_2 no es identificable en el sentido bayesiano, es decir, los efectos de los $p - n$ marcadores contenidos en u_2 no son identificables. Entonces, lo que los datos nos pueden informar sobre u_2 se hace solo a través de $\theta_1 = Z_1u_1 + Z_2u_2$, los valores genéticos aditivos de los individuos evaluados. La siguiente expresión nos muestra que el aprendizaje se logra mediante la distribución posterior de θ_1 y la a priori condicional $f(\theta_2|\theta_1)$, para ello escribimos:

$$f(\theta_2|y) = \int f(\theta_1, \theta_2|y)d\theta_1$$

$$= \int f(\theta_2|\theta_1, y)f(\theta_1|y)d\theta_1$$

$$= \int f(\theta_2|\theta_1)f(\theta_1|y)d\theta_1$$

$$= E_{\theta_1|y}\left[f(\theta_2|\theta_1)\right]$$

Es decir, la esperanza de la densidad $f(\theta_2|\theta_1)$ calculada con respecto a la distribución posterior de θ_1.

15.2. Ejercicios de evaluaciones genómicas

15.2.1. Montaje matricial de un modelo *ss – GBLUP*

A continuación desarrollaremos el EJEMPLO 15.2.1 con información de la variable intervalo entre el primer y segundo parto en bovinos (días).

Montaje de la genealogía en R-project [25]:

```
B=data.frame(matrix(ncol=6,c(
 1,NA,NA, 1,0,NA,
 2,NA,NA, 2,403,1,
 3,NA,NA, 1,0,NA,
 4,NA,NA, 2,432,1,
 5,NA,NA, 1,0,NA,
 6,NA,NA, 2,421,2,
 7, 1, 2, 1,0,NA,
 8, 3, 4, 2,398,2,
 9, 7, 4, 1,0,NA,
10, 3, 2, 2,345,1,
11, 1, 8, 1,0,NA,
12, 3, 2, 2,370,2,
14, 3,12, 2,412,2,
15, 7, 4, 1,0,NA,
16, 5, 6, 2,365,1,
17, 7,10, 1,0,NA,
21, 5, 4, 1,0,NA
), byrow=TRUE))
colnames(B)=c("id","sire","dam","sex", "IEP","GC")
```

```
B


##    id sire dam sex IEP GC
## 1   1   NA  NA   1   0 NA
## 2   2   NA  NA   2 403  1
## 3   3   NA  NA   1   0 NA
## 4   4   NA  NA   2 432  1
## 5   5   NA  NA   1   0 NA
## 6   6   NA  NA   2 421  2
## 7   7    1   2   1   0 NA
## 8   8    3   4   2 398  2
## 9   9    7   4   1   0 NA
## 10 10    3   2   2 345  1
## 11 11    1   8   1   0 NA
```

```
## 12 12    3   2   2 370  2
## 13 14    3  12   2 412  2
## 14 15    7   4   1   0 NA
## 15 16    5   6   2 365  1
## 16 17    7  10   1   0 NA
## 17 21    5   4   1   0 NA
```

Número de animales por padre y madre:

```
table(B$sire, B$dam)


##
##     2 4 6 8 10 12
##   1 1 0 0 1  0  0
##   3 2 1 0 0  0  1
##   5 0 1 1 0  0  0
##   7 0 2 0 0  1  0
```

Utilizaremos las librerías ≪kinship2≫ [28] para generar la matriz de parentesco, ≪MatrixModels≫ [29] para la construcción de las matrices a partir de las bases de datos, ≪stringr≫ [30] para modificar los nombres de las columnas y la librería ≪MASS≫ [18] para calcular la inversa:

```
library(kinship2)
Genea1=pedigree(id=B$id,dadid=B$sire,momid=B$dam,
                sex=B$sex)
```

Montaje de la matriz de parentesco:

```
A=2*kinship(B$id,B$sire,B$dam);
colnames(A)=B$id
rownames(A)=B$id
A[1:8,1:8]


##     1   2   3   4 5 6   7   8
## 1 1.0 0.0 0.0 0.0 0 0 0.5 0.0
## 2 0.0 1.0 0.0 0.0 0 0 0.5 0.0
## 3 0.0 0.0 1.0 0.0 0 0 0.0 0.5
## 4 0.0 0.0 0.0 1.0 0 0 0.0 0.5
## 5 0.0 0.0 0.0 0.0 1 0 0.0 0.0
## 6 0.0 0.0 0.0 0.0 0 1 0.0 0.0
## 7 0.5 0.5 0.0 0.0 0 0 1.0 0.0
## 8 0.0 0.0 0.5 0.5 0 0 0.0 1.0
```

```
bitSize(Genea1)


## $bitSize
## [1] 16
##
## $nFounder
## [1] 6
##
## $nNonFounder
## [1] 11
```

El árbol genealógico está representado en la FIGURA NRO. 15.1:

```
plot(Genea1,mar=c(bottom=1,left=1,top=5,right=1),cex=0.8)
```

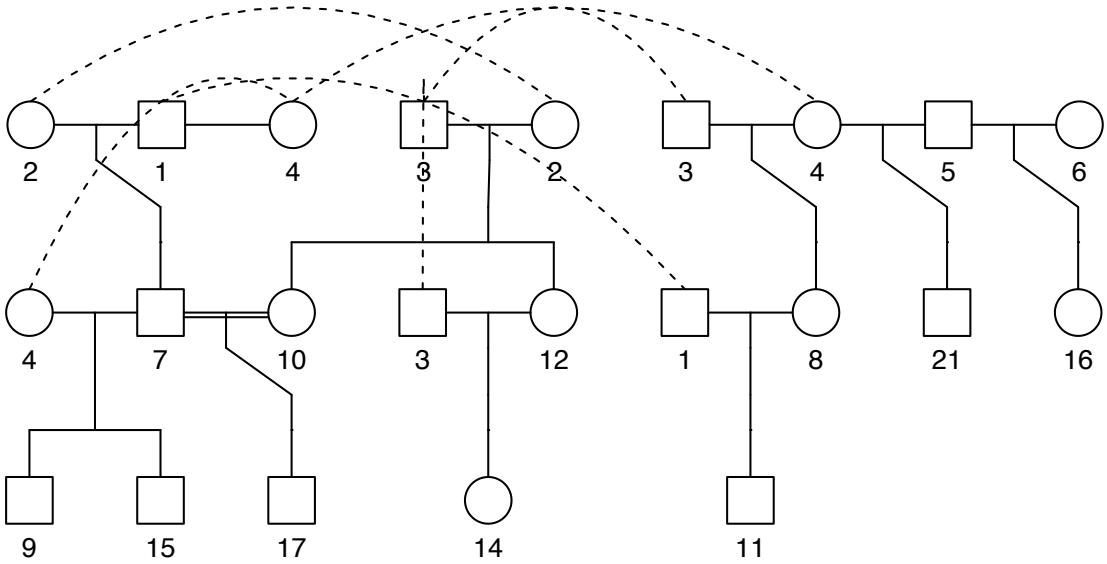

***Figura* 15.1:** Genealogía de los animales del ejercicio de modelo genómico
Fuente: elaboración propia generada en R-project [25].

Montaje de la matriz de parentesco:

```
A=2*kinship(B$id,B$sire,B$dam);
colnames(A)=B$id
rownames(A)=B$id
A[1:8,1:8]
##     1   2   3   4 5 6   7   8
## 1 1.0 0.0 0.0 0.0 0 0 0.5 0.0
## 2 0.0 1.0 0.0 0.0 0 0 0.5 0.0
## 3 0.0 0.0 1.0 0.0 0 0 0.0 0.5
## 4 0.0 0.0 0.0 1.0 0 0 0.0 0.5
## 5 0.0 0.0 0.0 0.0 1 0 0.0 0.0
## 6 0.0 0.0 0.0 0.0 0 1 0.0 0.0
## 7 0.5 0.5 0.0 0.0 0 0 1.0 0.0
## 8 0.0 0.0 0.5 0.5 0 0 0.0 1.0


bitSize(Genea1)


## $bitSize
## [1] 16
##
## $nFounder
## [1] 6
##
## $nNonFounder
## [1] 11
```

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0.5 & 0 & 0 & 0.25 & 0 & 0.25 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0.5 & 0.25 & 0.25 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0.25 & 0.5 & 0.75 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0.25 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.25 & 0.25 & 0 & 0.5 & 0 & \cdots & 0.25 & 0.12 & 0.06 & 1 & 0 & 0.31 & 0.25 \\ 0 & 0 & 0 & 0 & 0.5 & \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0.25 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & \cdots & 0.19 & 0.38 & 0.31 & 0.31 & 0 & 1.12 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & \cdots & 0.12 & 0 & 0 & 0.25 & 0.25 & 0 & 1 \end{bmatrix}$$

Información de cinco marcadores SNP:

```
Genotipo=data.frame(matrix(ncol=6,byrow=TRUE,c(
 1,      0,       1,       2,       1,       1,
 2,      0,       0,       0,       2,       2,
 3,      1,       1,       1,       1,       1,
 4,      0,       0,       0,       0,       0,
 5,      1,       2,       1,       2,       1,
#6,2,2,1,1,1, no lo tendremos en cuenta para el ejemplo
 7,      0,       0,       1,       2,       1,
 8,      1,       0,       0,       1,       0,
 9,      0,       0,       1,       1,       0,
10,      0,       1,       1,       2,       1,
11,      0,       0,       1,       2,       0,
12,      1,       1,       0,       2,       1,
14,      0,       1,       0,       2,       2,
15,      0,       0,       0,       1,       1,
16,      2,       2,       0,       2,       0,
17,      0,       0,       1,       2,       2,
21,      0,       1,       1,       1,  0)))
colnames(Genotipo)=c("Id","SNP1","SNP2","SNP3","SNP4","SNP5")
head(Genotipo)


##    Id SNP1 SNP2 SNP3 SNP4 SNP5
## 1   1    0    1    2    1    1
## 2   2    0    0    0    2    2
## 3   3    1    1    1    1    1
## 4   4    0    0    0    0    0
## 5   5    1    2    1    2    1
## 6   7    0    0    1    2    1


N_genotipados=nrow(Genotipo)
N_genotipados


## [1] 16
```

Como se indicó en la sección 7.8.2, construiremos la matriz A_{22} para generar posteriormente la matriz H^{-1}, así:

```
A22=matrix(ncol=N_genotipados,nrow=N_genotipados,0)
colnames(A22)=Genotipo$Id
rownames(A22)=Genotipo$Id
A22=A[rownames(A22),colnames(A22)]
```

$$A_{22} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0.5 & 0 & 0 & 0.25 & 0 & 0.25 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0.5 & 0.25 & 0.25 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0.25 & 0.5 & 0.75 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0.25 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.25 & 0.25 & 0 & 0.5 & 0 & \cdots & 0.25 & 0.12 & 0.06 & 1 & 0 & 0.31 & 0.25 \\ 0 & 0 & 0 & 0 & 0.5 & \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0.25 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & \cdots & 0.19 & 0.38 & 0.31 & 0.31 & 0 & 1.12 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & \cdots & 0.12 & 0 & 0 & 0.25 & 0.25 & 0 & 1 \end{bmatrix}$$

Ahora construiremos la matriz de diseño M, que relaciona los individuos (i) y los genotipos de los SNP (j), teniendo en cuenta lo descrito por VanRaden [48] y el ejercicio de Putz [113], donde:

$$m_{ij} = \begin{cases} 0 \text{ si es homocigótico 11} \\ 1 \text{ si es heterocigótico 21 o 12} \\ 2 \text{ si es homocigótico 22} \end{cases}$$

```
M=as.matrix(ncol=5,Genotipo[,-1], byrow=TRUE)
rownames(M)=Genotipo$Id
M
```

```
##    SNP1 SNP2 SNP3 SNP4 SNP5
## 1     0    1    2    1    1
## 2     0    0    0    2    2
## 3     1    1    1    1    1
## 4     0    0    0    0    0
## 5     1    2    1    2    1
## 7     0    0    1    2    1
## 8     1    0    0    1    0
## 9     0    0    1    1    0
## 10    0    1    1    2    1
## 11    0    0    1    2    0
## 12    1    1    0    2    1
## 14    0    1    0    2    2
## 15    0    0    0    1    1
## 16    2    2    0    2    0
## 17    0    0    1    2    2
## 21    0    1    1    1    0
```

Ahora calculamos las frecuencias de p_j:

```
fp_j=as.matrix(apply(M,2,mean)/2)
fp_j
```

```
##       [,1]
## SNP1 0.19
## SNP2 0.31
## SNP3 0.31
## SNP4 0.75
## SNP5 0.41
```

$$F_{p_j} = \begin{bmatrix} 0.19 \\ 0.31 \\ 0.31 \\ 0.75 \\ 0.41 \end{bmatrix}$$

Ahora calculamos la versión estandarizada de F_{p_j} denotada como p_j:

$$p_j = 2 * (F_{p_j})$$

```
p_j=2*(fp_j);p_j
```

```
##       [,1]
## SNP1 0.38
## SNP2 0.62
## SNP3 0.62
## SNP4 1.50
## SNP5 0.81
```

El vector p_j sería:

$$p_j = \begin{bmatrix} 0.38 \\ 0.62 \\ 0.62 \\ 1.5 \\ 0.81 \end{bmatrix}$$

La matriz P contiene los vectores p_j de todos los individuos:

```
P=matrix(ncol=5,rep(p_j,nrow(M)),byrow=TRUE)
colnames(P)=colnames(M)
rownames(P)=Genotipo$Id
P
```

```
##    SNP1 SNP2 SNP3 SNP4 SNP5
## 1  0.38 0.62 0.62  1.5 0.81
## 2  0.38 0.62 0.62  1.5 0.81
## 3  0.38 0.62 0.62  1.5 0.81
## 4  0.38 0.62 0.62  1.5 0.81
## 5  0.38 0.62 0.62  1.5 0.81
## 7  0.38 0.62 0.62  1.5 0.81
## 8  0.38 0.62 0.62  1.5 0.81
## 9  0.38 0.62 0.62  1.5 0.81
## 10 0.38 0.62 0.62  1.5 0.81
## 11 0.38 0.62 0.62  1.5 0.81
## 12 0.38 0.62 0.62  1.5 0.81
## 14 0.38 0.62 0.62  1.5 0.81
## 15 0.38 0.62 0.62  1.5 0.81
## 16 0.38 0.62 0.62  1.5 0.81
## 17 0.38 0.62 0.62  1.5 0.81
## 21 0.38 0.62 0.62  1.5 0.81
```

Ahora calculamos Q que es una matriz que contiene los efectos alélicos centrados:

$$Q = M - P$$

```
Q=M-P
Q
```

```
##     SNP1  SNP2  SNP3 SNP4  SNP5
## 1  -0.38  0.38  1.38 -0.5  0.19
## 2  -0.38 -0.62 -0.62  0.5  1.19
## 3   0.62  0.38  0.38 -0.5  0.19
## 4  -0.38 -0.62 -0.62 -1.5 -0.81
## 5   0.62  1.38  0.38  0.5  0.19
## 7  -0.38 -0.62  0.38  0.5  0.19
## 8   0.62 -0.62 -0.62 -0.5 -0.81
## 9  -0.38 -0.62  0.38 -0.5 -0.81
## 10 -0.38  0.38  0.38  0.5  0.19
## 11 -0.38 -0.62  0.38  0.5 -0.81
```

```
## 12  0.62  0.38 -0.62  0.5  0.19
## 14 -0.38  0.38 -0.62  0.5  1.19
## 15 -0.38 -0.62 -0.62 -0.5  0.19
## 16  1.62  1.38 -0.62  0.5 -0.81
## 17 -0.38 -0.62  0.38  0.5  1.19
## 21 -0.38  0.38  0.38 -0.5 -0.81
```

$$Q = \begin{bmatrix} -0.38 & 0.38 & 1.38 & -0.5 & 0.19 \\ -0.38 & -0.62 & -0.62 & 0.5 & 1.19 \\ 0.62 & 0.38 & 0.38 & -0.5 & 0.19 \\ -0.38 & -0.62 & -0.62 & -1.5 & -0.81 \\ 0.62 & 1.38 & 0.38 & 0.5 & 0.19 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -0.38 & -0.62 & -0.62 & -0.5 & 0.19 \\ 1.62 & 1.38 & -0.62 & 0.5 & -0.81 \\ -0.38 & -0.62 & 0.38 & 0.5 & 1.19 \\ -0.38 & 0.38 & 0.38 & -0.5 & -0.81 \end{bmatrix}$$

Ahora calculamos G, que es una matriz de relaciones genómicas, dada por:

$$G = \frac{QQ^T}{k}$$

Donde $k = (2_j \sum p_j(1 - pj))$:

```
fq_j=1-fp_j
k=2*(t(fq_j)%*%fp_j)
k
```

```
##      [,1]
## [1,]    2
```

```
G= (Q%*%t(Q))%x%(1/k)
colnames(G)=rownames(Q)
rownames(G)=rownames(Q)
round(G[1:6,1:6],2)
```

```
##       1     2     3     4     5     7
## 1  1.22 -0.49  0.35 -0.18  0.29  0.10
## 2 -0.49  1.28 -0.36 -0.39 -0.42  0.38
## 3  0.35 -0.36  0.47 -0.05  0.41 -0.27
## 4 -0.18 -0.39 -0.05  1.90 -1.10 -0.30
## 5  0.29 -0.42  0.41 -1.10  1.34 -0.33
## 7  0.10  0.38 -0.27 -0.30 -0.33  0.47
```

Ahora calculamos:

$$[G^{-1} - A_{22}^{-1}]$$

```
library(MASS)
GinvA22inv=ginv(G)-ginv(A22)
colnames(GinvA22inv)=rownames(Q)
rownames(GinvA22inv)=rownames(Q)
round(GinvA22inv[1:6,1:6],2)
```

```
##       1     2     3     4     5     7
## 1 -1.79 -0.55  0.22 -0.01  0.04  0.98
## 2 -0.55 -2.44 -1.00  0.00 -0.03  1.01
## 3  0.22 -1.00 -2.53 -0.57 -0.01  0.02
## 4 -0.01  0.00 -0.57 -2.79 -0.50 -1.13
## 5  0.04 -0.03 -0.01 -0.50 -1.77 -0.04
## 7  0.98  1.01  0.02 -1.13 -0.04 -3.39
```

Teniendo como base el BLUP genómico de un solo paso, reemplazando la matriz A^{-1} por la matriz H^{-1} (matriz de parentesco modificado, incorporando información genómica), se tendría:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

```
n=nrow(B)
Ad=matrix(nrow=n,ncol=n,0)
rownames(Ad)=B$id
colnames(Ad)=B$id
Ad[rownames(GinvA22inv),colnames(GinvA22inv)]=GinvA22inv
round(Ad[1:6,1:6],2)
```

```
##       1     2     3     4     5 6
## 1 -1.79 -0.55  0.22 -0.01  0.04 0
## 2 -0.55 -2.44 -1.00  0.00 -0.03 0
## 3  0.22 -1.00 -2.53 -0.57 -0.01 0
## 4 -0.01  0.00 -0.57 -2.79 -0.50 0
## 5  0.04 -0.03 -0.01 -0.50 -1.77 0
## 6  0.00  0.00  0.00  0.00  0.00 0
```

```
Hinv=ginv(A)+Ad
round(Hinv[1:6,1:6],2)
```

```
##        1     2     3     4     5   6
## 1   0.21 -0.05  0.22 -0.01  0.04 0.0
## 2  -0.05  0.06  0.00  0.00 -0.03 0.0
## 3   0.22  0.00  0.47 -0.07 -0.01 0.0
## 4  -0.01  0.00 -0.07  0.21  0.00 0.0
## 5   0.04 -0.03 -0.01  0.00  0.23 0.5
## 6   0.00  0.00  0.00  0.00  0.50 1.5
```

```
rownames(B)=B$id
```

Ahora desarrollamos el sistema de ecuaciones:

Matriz Z:

```
library(MatrixModels)
Z=as.matrix(model.Matrix(~ as.factor(B$id) -1))
colnames(Z)=B$id
rownames(Z)=B$id
Z[1:4,1:4]
```

```
##   1 2 3 4
## 1 1 0 0 0
## 2 0 1 0 0
## 3 0 0 1 0
## 4 0 0 0 1
```

```
diag(Z)=ifelse(B$IEP==0,0,1)
```

Matriz X:

```
Fijos=subset(B,!is.na(B$GC), select=c("GC"))
Fijos
```

```
##    GC
## 2   1
## 4   1
## 6   2
## 8   2
## 10  1
## 12  2
## 14  2
## 16  1
```

```
Fijo=as.matrix(model.Matrix(~ as.factor(Fijos$GC) -1))
rownames(Fijo)=rownames(Fijos)
Fijo
```

```
##    as.factor(Fijos$GC)1 as.factor(Fijos$GC)2
## 2                     1                     0
## 4                     1                     0
## 6                     0                     1
## 8                     0                     1
## 10                    1                     0
## 12                    0                     1
## 14                    0                     1
## 16                    1                     0
```

```
X=matrix(nrow=n,ncol=ncol(Fijo),0)
rownames(X)=B$id
X[rownames(Fijo),]=Fijo
library(stringr)
colnames(X)=word(colnames(Fijo), 2, sep = fixed(')'))
X
```

```
##    1 2
## 1  0 0
## 2  1 0
## 3  0 0
## 4  1 0
## 5  0 0
## 6  0 1
## 7  0 0
## 8  0 1
## 9  0 0
## 10 1 0
## 11 0 0
## 12 0 1
## 14 0 1
## 15 0 0
## 16 1 0
## 17 0 0
## 21 0 0
```

Operaciones matriciales:

```
XpX=t(X)%*%X
XpZ=t(X)%*%Z
```

```
ZpX=t(Z)%*%X;ZpX
```

```
##    1 2
## 1  0 0
## 2  1 0
## 3  0 0
## 4  1 0
## 5  0 0
## 6  0 1
## 7  0 0
## 8  0 1
## 9  0 0
## 10 1 0
## 11 0 0
## 12 0 1
## 14 0 1
## 15 0 0
## 16 1 0
## 17 0 0
## 21 0 0
```

```
ZpZ=t(Z)%*%Z;ZpZ
```

```
##    1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 21
## 1  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 2  0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 3  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 4  0 0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0
## 5  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 6  0 0 0 0 0 1 0 0 0  0  0  0  0  0  0  0  0
## 7  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 8  0 0 0 0 0 0 0 1 0  0  0  0  0  0  0  0  0
## 9  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 10 0 0 0 0 0 0 0 0 0  1  0  0  0  0  0  0  0
## 11 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 12 0 0 0 0 0 0 0 0 0  0  0  1  0  0  0  0  0
## 14 0 0 0 0 0 0 0 0 0  0  0  0  1  0  0  0  0
## 15 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
```

```
## 16 0 0 0 0 0 0 0 0 0  0  0  0  0  0  1  0  0
## 17 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
## 21 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0
```

```
Xpy=t(X)%*%B$IEP;Xpy
```

```
##   [,1]
## 1 1545
## 2 1601
```

```
Zpy=t(Z)%*%B$IEP;Zpy
```

```
##    [,1]
## 1     0
## 2   403
## 3     0
## 4   432
## 5     0
## 6   421
## 7     0
## 8   398
## 9     0
## 10  345
## 11    0
## 12  370
## 14  412
## 15    0
## 16  365
## 17    0
## 21    0
```

Consideremos a manera de ejemplo que la varianza aditiva es de 90 d^2 y la residual de 812 d^2:

```
vara=90
vare=812
```

```
Hinvmasalfa=Hinv+(vare/vara)
ZpZHinvmasalfa=ZpZ+Hinvmasalfa
```

Montaje del sistema:

```
Izq=rbind(
  cbind(XpX,XpZ),
  cbind(ZpX,ZpZHinvmasalfa))
```

```
Der=rbind(Xpy,Zpy)
```

Solución del sitema:

```
Sol=round(ginv(Izq)%*%Der,20)
rownames(Sol)=c(rownames(XpX),paste0("a_",rownames(ZpZ)))
Sol
```

```
##         [,1]
## 1      340.6
## 2      400.7
## a_1     29.4
## a_2     62.4
## a_3    -11.1
## a_4     91.4
## a_5    -12.1
## a_6     20.3
## a_7      7.1
## a_8     -2.7
## a_9    -33.4
## a_10     4.4
## a_11    22.6
## a_12   -30.7
## a_14    11.3
## a_15 -134.5
## a_16    24.4
## a_17    -8.3
## a_21   -40.8
```

15.2.2. Ejemplo del alfabeto bayesiano con datos simulados

Para el desarrollo del EJEMPLO 15.2.2 descargaremos la base de datos denominada *Pop1_mrk_001.txt* con los genotipos de lo 430 animales y 4752 SNPs y la base *Pop1_data_001.txt* con información fenotípica, respectivamente (Cliclar los archivos txt anteriores)

Una imagen de la base de datos está en la FIGURA NRO. 15.2, donde se tienen dos columnas separadas por espacio. En la primera columna aparece el nombre del invididuo y en la segunda la secuencia de los SNPs, con los números: 2 (homocigótico de referencia), 3 y 4 (heterocigótico) y 0 (el otro homocigótico). Estos valores se modificarán cuando se construya la matriz W, quedando: el 2 como 1 (homocigótico de referencia), 3 y 4 como 0 y el 0 como -1 (otro homocigótico).

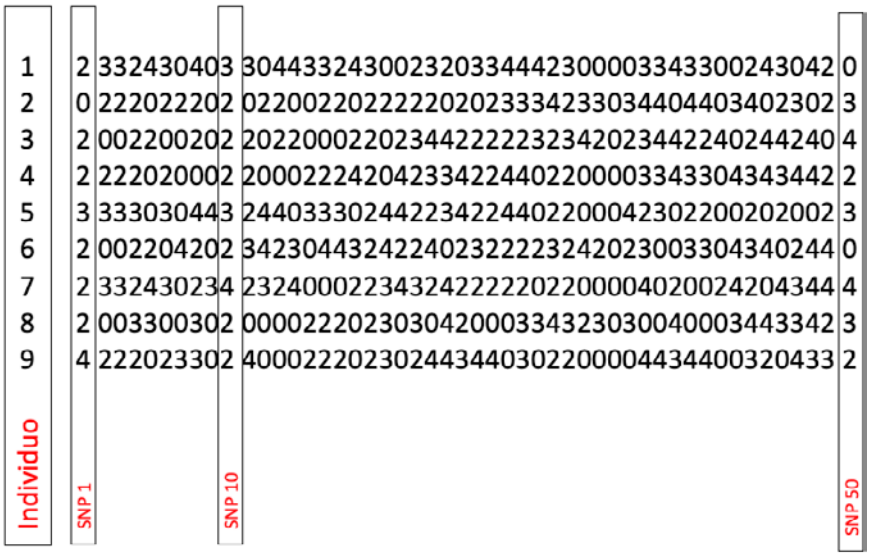

***Figura* 15.2:** Estructura de la base de datos con información de los primeros nueve individuos y los primeros 50 SNP.
Fuente: elaboración propia (2024).

Para leer la base de datos utilizaremos el comando read.delim el cual generará las dos columnas (identificación del animal y los genotipos). También utilizamos la librería ≪stringr≫ [30] para separar los caracteres que tiene una columna.

```
rm(list=ls())
#cambiar la ruta según la organización de su computador
SNP = read.delim(
"~/Desktop/Desktop/EsTuDiAnTeS/LM/01-Apoy/Pop1_mrk_001.txt",
                  colClasses = c("character"),header = FALSE)
library(stringr)
#usar el primer conjunto de caracteres (antes del espacio)
SNP$Ind=word(SNP$V1,1)
#usar el conjunto de caracteres después del espacio
```

```
SNP$Geno=word(SNP$V1,-1)
SNP$V1=NULL
```

Antes de trabajar la base de datos con todos los SNPs disponibles, utilizaremos los primeros 50 SNP con el objetivo de reducir el tiempo de ejecución de las rutinas. Cuando ya entiendas todo el procedimiento, puedes utilizar los 4752 SNP para tener un resultado más completo:

```
nSNP=nchar(SNP[1,2]);nSNP


## [1] 4752


#inicialmente trabajaremos con los primeros 50 SNP
#borrar la siguiente linea para todos los SNP
nSNP=50;nSNP


## [1] 50
```

```
individuos=length(SNP[,1]);individuos


## [1] 430


for(j in 1:nSNP)
{
  columna = c()
  for(i in 1:individuos)
  {
    n = strsplit(SNP[, 2][i], '')
    columna = c(columna, unlist(n)[j])
  }
  SNP = cbind(SNP, as.numeric(columna))
}
colnames(SNP)=c("id","geno", paste0("SNP",seq(1,nSNP,1)))
```

La base de datos tiene 50 SNPs y 430 individuos:

```
Z=as.matrix(SNP[,3:ncol(SNP)])
colnames(Z)=paste0("SNP",seq(1,nSNP,1))
rownames(Z)=SNP$Ind
Z[1:3,1:3]
```

```
##       SNP1 SNP2 SNP3
## [1,]    2    3    3
## [2,]    0    2    2
## [3,]    2    0    0
```

```
W=matrix(0,nrow=nrow(Z),ncol=ncol(Z))
for(i in 1:nrow(Z)){
      for(j in 1:ncol(Z)){
                if(Z[i,j]==0 || Z[i,j]==2){
                                        W[i,j]=Z[i,j]-1
                }
                if(Z[i,j]==3 || Z[i,j]==4){
                                        W[i,j]=0
                }
      }
   }
dim(W)


## [1] 430  50
```

```
W[1:3,1:3]


##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]   -1    1    1
## [3,]    1   -1   -1
```

Información genealógica y productiva:

```
base=read.table(
"~/Desktop/Desktop/EsTuDiAnTeS/LM/01-Apoy/Pop1_data_001.txt"
,header=T)
n=nrow(base);n


## [1] 430


Entrenam=base[-which(base$G==3), ]
nrow(Entrenam)


## [1] 330
```

```
W.Entren=W[-which(base$G==3), ]
nrow(W.Entren)


## [1] 330
```

```
Prueba=base[which(base$G==3), ]
nrow(Prueba)


## [1] 100
```

```
W.Prueb=W[which(base$G==3), ]
nrow(W.Prueb)


## [1] 100
```

Análisis con BGLR bayes B:

```
library(BGLR)
R2=0.5 #heredabilidad
#Para reducir el tiempo de cómputo y
#solo a manera de ejemplo
#inicialmente utilizaremos 200 iteraciones y 50 de arranque,
#después puede utilizar otros valores
Iteracciones=200
Arranque=50
ETA1=list(list(X=W.Entren, model='BayesB',R2=R2))
PostmeanMarkBB1=as.matrix((BGLR(y=Entrenam$Phen,
                        ETA=ETA1,nIter=Iteracciones,
                        burnIn=Arranque))$ETA[[1]]$b)
```

Generación de las medias posteriores de los efectos alélicos de los primeros marcadores:

```
PrimerosMarcadores=6
round(PostmeanMarkBB1[1:PrimerosMarcadores],3)


## [1] -0.41 -0.03 -0.38 -0.18 -0.15 -0.43
```

Con la siguiente programación obtendremos el histograma de las medias posteriores de los efectos alélicos de todos los marcadores (FIGURA NRO. 15.3):

```
hist(PostmeanMarkBB1,border = "BLUE",
   mar=c(bottom=0, left=1, top=1, right=1))
```

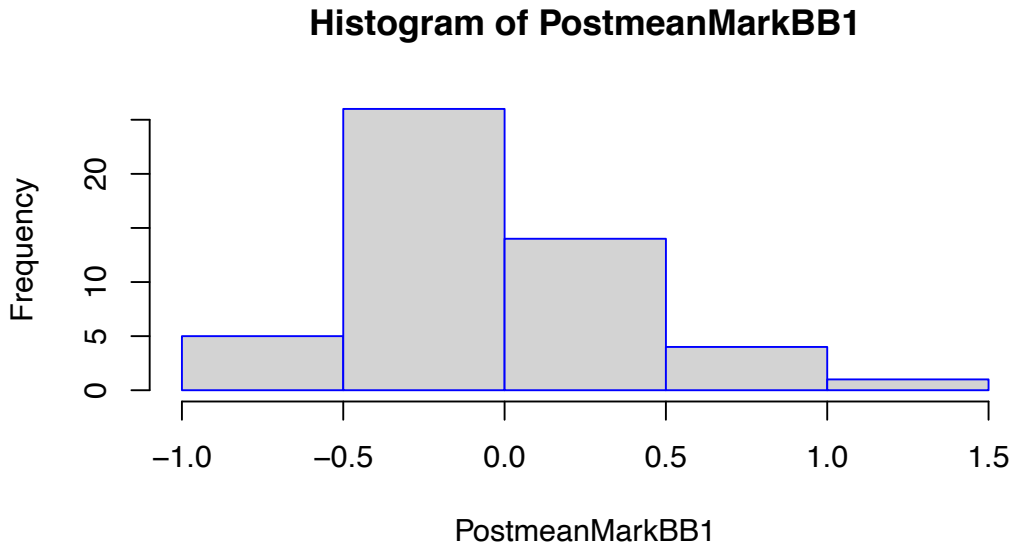

***Figura* 15.3:** Histograma de frecuencias de las medias posteriores de los efectos alélicos de los marcadores
Fuente: elaboración propia generada en R-project [25].

Valores genéticos predichos de los conjuntos de entrenamiento y prueba:

```
#Conjunto de entrenamiento
PredBVTrainBB1=W.Entren%*%PostmeanMarkBB1
round(PredBVTrainBB1[c(1:4)],2)


## [1]  2.73  0.54 -2.07 -1.64
```

```
# Conjunto de prueba
PredBVTestBB1=W.Prueb%*%PostmeanMarkBB1
PredBVTestBB1[c(1:3)]


## [1] -0.18 -2.20  0.51
```

```
# correlación predictiva
Predabil.BB1=cor(Prueba$Phen,PredBVTestBB1,
                 method="pearson")
round(Predabil.BB1,3)


##        [,1]
## [1,] 0.023
```

```
# Confiabilidad de los valores genéticos conjunto de prueba
corrBVBB1=cor(Prueba$QTL,PredBVTestBB1,method="pearson")
round(corrBVBB1,3)


##       [,1]
## [1,] 0.12
```

```
# Confiabilidad de los valores genéticos conjunto de
#entrenamiento
corrBVTrainBB1=cor(Entrenam$QTL,PredBVTrainBB1,
                   method="pearson")
round(corrBVTrainBB1,3)


##       [,1]
## [1,] 0.21
```

Ahora realizamos los mismos cómputos con otros miembros del alfabeto bayesiano (bayes A):

```
ETA1=list(list(X=W.Entren, model='BayesA',R2=R2))
PostmeanMarkBA1=as.matrix((BGLR(y=Entrenam$Phen,
                      ETA=ETA1, nIter=Iteracciones,
                     burnIn=Arranque))$ETA[[1]]$b)
PredBVTrainBA1=W.Entren%*%PostmeanMarkBA1
PredBVTestBA1=W.Prueb%*%PostmeanMarkBA1
Predabil.BA1=cor(Prueba$Phen,PredBVTestBA1,
                 method="pearson")
corrBVBA1=cor(Prueba$QTL,PredBVTestBA1,method="pearson")
round(corrBVBA1,3)


##       [,1]
## [1,] 0.13
```

```
corrBVTrainBA1=cor(Entrenam$QTL,PredBVTrainBA1,
                   method="pearson")
round(corrBVTrainBA1,3)


##       [,1]
## [1,] 0.32
```

Bayes C:

```
ETA1=list(list(X=W.Entren, model='BayesC',R2=R2))
PostmeanMarkBC1=as.matrix((BGLR(y=Entrenam$Phen,
                        ETA=ETA1,nIter=Iteracciones,
                        burnIn=Arranque))$ETA[[1]]$b)
PredBVTrainBC1=W.Entren%*%PostmeanMarkBC1
PredBVTestBC1=W.Prueb%*%PostmeanMarkBC1
Predabil.BC1=cor(Prueba$Phen,PredBVTestBC1,
                 method="pearson")
```

```
round(Predabil.BC1,3)


##       [,1]
## [1,] 0.039


corrBVBC1=cor(Prueba$QTL,PredBVTestBC1,
              method="pearson")
round(corrBVBC1,3)


##       [,1]
## [1,] 0.041
```

```
corrBVTrainBC1=cor(Entrenam$QTL,PredBVTrainBC1,
                   method="pearson")
round(corrBVTrainBC1,3)


##       [,1]
## [1,] 0.18
```

Regresión ridge bayesiana, similar a bayes A pero se asigna la misma varianza a priori condicional a los efectos de los marcadores:

```
ETA1=list(list(X=W.Entren, model='BRR',R2=R2))
PostmeanMarkBRR1=as.matrix((BGLR(y=Entrenam$Phen,
                                 ETA=ETA1,
                            nIter=Iteracciones,
                            burnIn=Arranque))$ETA[[1]]$b)
PredBVTrainBRR1=W.Entren%*%PostmeanMarkBRR1
PredBVTestBRR1=W.Prueb%*%PostmeanMarkBRR1
Predabil.BRR1=cor(Prueba$Phen,PredBVTestBRR1,
```

```
                    method="pearson")
corrBVBRR1=cor(Prueba$QTL,PredBVTestBRR1,
                  method="pearson")
corrBVBRR1
```

```
##       [,1]
## [1,] 0.14
```

```
corrBVTrainBRR1=cor(Entrenam$QTL,PredBVTrainBRR1,
                      method="pearson")
round(corrBVTrainBRR1,3)
```

```
##       [,1]
## [1,] 0.32
```

Con la finalidad de comparar resultados, vamos a visualizar los histogramas de las medias posteriores de los efectos de los marcadores bajo los 4 modelos en un solo gráfico (FIGURA NRO. 15.4):

```
par(mfrow=c(2,2),mar=c(bottom=1, left=1, top=1, right=1))
hist(PostmeanMarkBB1,border = "BLUE")
hist(PostmeanMarkBA1,border = "BLUE")
hist(PostmeanMarkBC1,border = "BLUE")
hist(PostmeanMarkBRR1,border = "BLUE")
```

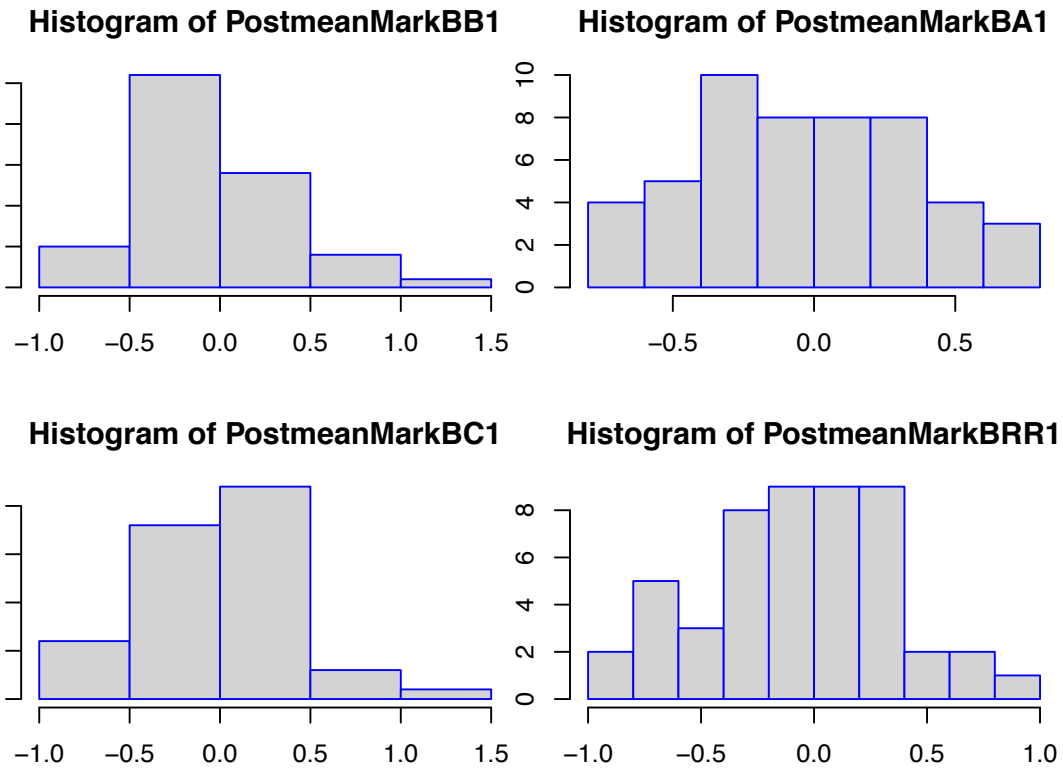

***Figura* 15.4:** Histograma de las medias posteriores de los efectos de los marcadores. Fuente: elaboración propia generada en R-project [25].

Bayes B presenta una mayor frecuencia relativa de marcadores con efectos cercanos a 0. Ahora visualizamos los histogramas de las medias posteriores de los valores genéticos aditivos bajo los 4 modelos en un solo gráfico (FIGURA NRO. 15.5):

```
par(mfrow=c(2,2),mar=c(bottom=1, left=1, top=1, right=1))
hist(PredBVTestBB1,border = "red")
hist(PredBVTestBA1,border = "red")
hist(PredBVTestBC1,border = "red")
hist(PredBVTestBRR1,border = "red")
```

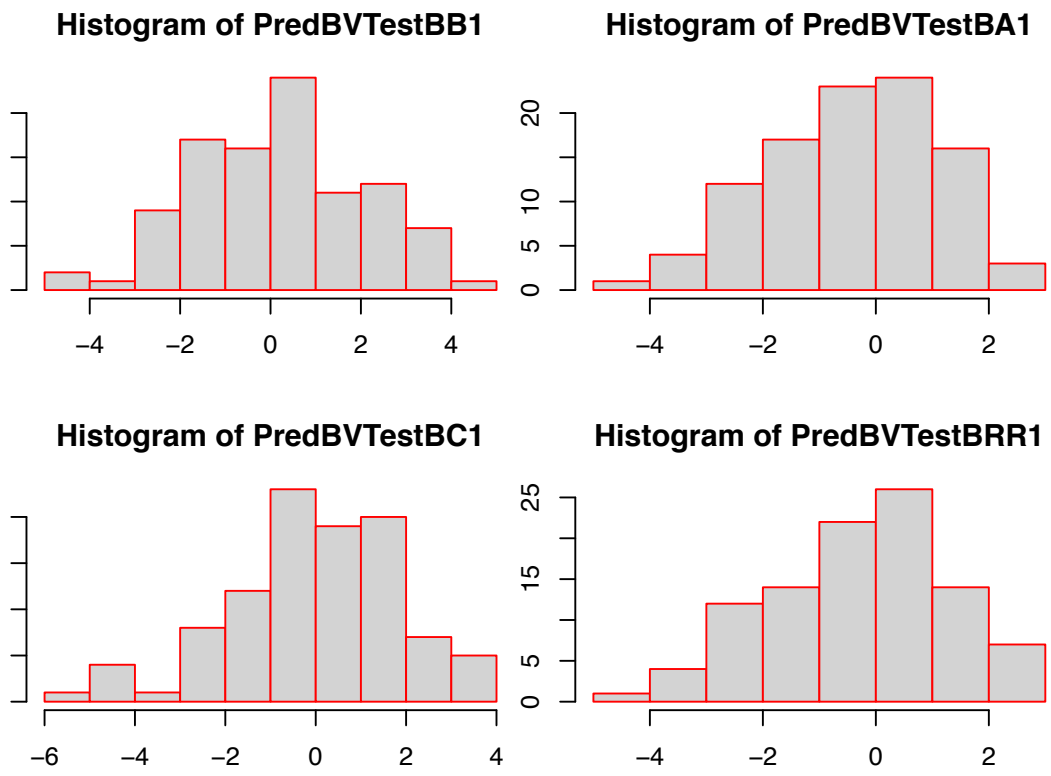

***Figura* 15.5:** Histograma de las medias posteriores de los efectos de los valores genéticos aditivos.
Fuente: elaboración propia generada en R-project [25].

Finalmente, creamos tablas con estadísticas descriptivas de las medias posteriores de 1) efectos alélicos de los marcadores y 2) valores genéticos aditivos de los individuos bajo los 4 modelos y una tabla con el resumen del desempeño predictivo:

```
Desc_Efectos_Marc=data.frame(matrix(NA,ncol=7,nrow=4))
colnames(Desc_Efectos_Marc)=c("Modelo","Min","1er cuart",
                              "Mediana","Media","3er cuart",
                              "Max")
Desc_Efectos_Marc[,1]=c("Bayes B","Bayes A","Bayes C",
                        "BayesRR")
Desc_Efectos_Marc[1,2:7]=round(summary(as.numeric
                                       (PostmeanMarkBB1)),2)
Desc_Efectos_Marc[2,2:7]=round(summary(as.numeric
                                       (PostmeanMarkBA1)),2)
Desc_Efectos_Marc[3,2:7]=round(summary(as.numeric
                                       (PostmeanMarkBC1)),2)
Desc_Efectos_Marc[4,2:7]=round(summary(as.numeric
```

```
                                          (PostmeanMarkBRR1)),2)
Desc_Efectos_Marc


##    Modelo   Min 1er cuart Mediana Media 3er cuart  Max
## 1 Bayes B -0.92     -0.38   -0.19 -0.09      0.13 1.19
## 2 Bayes A -0.70     -0.28   -0.10 -0.03      0.30 0.80
## 3 Bayes C -0.96     -0.21    0.01  0.00      0.24 1.26
## 4 BayesRR -0.98     -0.30   -0.06 -0.06      0.23 0.95
```

Como se apreciaba en los histogramas, bayes B y bayes A presentaron un rango un poco más amplio en las estimaciones de efectos de sustitución alélica:

```
Desc_VGA=data.frame(matrix(NA,ncol=7,nrow=4))
colnames(Desc_VGA)=c("Modelo","Min","1er cuart","Mediana",
                                "Media", "3er cuart","Max")
Desc_VGA[,1]=c("Bayes B","Bayes A","Bayes C","BayesRR")
Desc_VGA[1,2:7]=round(summary(as.numeric(PredBVTestBB1)),2)
Desc_VGA[2,2:7]=round(summary(as.numeric(PredBVTestBA1)),2)
Desc_VGA[3,2:7]=round(summary(as.numeric(PredBVTestBC1)),2)
Desc_VGA[4,2:7]=round(summary(as.numeric(PredBVTestBRR1)),2)
Desc_VGA


##    Modelo  Min 1er cuart Mediana Media 3er cuart Max
## 1 Bayes B -4.8      -1.2    0.14  0.18      1.49 4.9
## 2 Bayes A -4.7      -1.7   -0.32 -0.42      0.51 2.5
## 3 Bayes C -5.0      -1.1    0.06 -0.05      1.34 3.6
## 4 BayesRR -4.8      -1.5   -0.13 -0.26      0.85 3.0


Desemp_Pred=data.frame(matrix(NA,ncol=3,nrow=4))
colnames(Desemp_Pred)=c("Modelo","CorrelPred","Confiabi")
Desemp_Pred[,1]=c("Bayes B","Bayes A","Bayes C","BayesRR")
Desemp_Pred[1,2:3]=round(c(Predabil.BB1,corrBVBB1),2)
Desemp_Pred[2,2:3]=round(c(Predabil.BA1,corrBVBA1),2)
Desemp_Pred[3,2:3]=round(c(Predabil.BC1,corrBVBC1),2)
Desemp_Pred[4,2:3]=round(c(Predabil.BRR1,corrBVBRR1),2)
Desemp_Pred


##    Modelo CorrelPred Confiabi
## 1 Bayes B       0.02     0.12
## 2 Bayes A       0.06     0.13
## 3 Bayes C       0.04     0.04
## 4 BayesRR       0.10     0.14
```

15.2.3. Ejercicio en ratones

La librería ≪BGLR≫ [114] proporciona un conjunto de bases de datos con información de 1814 ratones (mice). Para el desarrollo del EJEMPLO 15.2.3 utilizaremos la variable respuesta obesidad y (variable Obesity.BMI de la base mice.pheno), la matriz de diseño de los genotipos (marcadores) con los 1814 ratones y 10346 SNPs (base de datos mice.X), la matriz de relaciones genéticas aditivas de los 1814 ratones (base de datos mice.A), la covariable tamaño de camada (variable Litter de la base mice.pheno), información del sexo de los individuos (variable GENDER de la base de datos mice.pheno) e información de procedencia de la jaula (variable cage, de la base mice.pheno):

```
rm(list=ls())
library(BGLR)
```

```
data(mice)
#Variables y matrices para iniciar el análisis
#Variable obesidad
Animal=mice.pheno$SUBJECT.NAME
Obesidad=mice.pheno$Obesity.BMI
media=mean(Obesidad);media
```

```
## [1] -0.46
```

```
desvio=sd(Obesidad);desvio
```

```
## [1] 0.06
```

```
#estandarización de la variable obesidad
y=(Obesidad-media)/desvio
round(min(y),2)
```

```
## [1] -2.8
```

```
round(max(y),2)
```

```
## [1] 3.1
```

```
round(mean(y),2)
```

```
## [1] 0
```

```
round(sd(y),2)
```

```
## [1] 1
```

```
X=mice.X;dim(X)
```

```
## [1]  1814 10346
```

```
#inicialmente trabajaremos con los primeros 50 SNP
#borrar la línea siguiente, para correr todos los SNP
X=X[,1:50]
A=mice.A;dim(A)
```

```
## [1] 1814 1814
```

```
TC=mice.pheno$Litter
table(TC)
```

```
## TC
##   1   2   3   4   5   6   7   8
## 673 515 339 176  69  24  14   4
```

```
Sexo=mice.pheno$GENDER
table(Sexo)
```

```
## Sexo
##   F   M
## 880 934
```

```
Jaula=mice.pheno$cage
```

```
#los primeros y últimos  animales de la base
#como referencia
animalesref=c(1:2,1813:1814)
Refer=data.frame(Animal[animalesref],TC[animalesref],
```

```
            Sexo[animalesref],Jaula[animalesref],
            round(Obesidad[animalesref],2),
            round(y[animalesref],2))
colnames(Refer)=
  c("Animal","TC","Sexo","Jaula", "Obesidad", "y")
Refer
```

```
##       Animal TC Sexo Jaula Obesidad     y
## 1 A048005080  2    F   19F    -0.52 -1.06
## 2 A048006063  4    M   13C    -0.40  0.94
## 3 A084291787  3    M   83A    -0.40  0.99
## 4 A084292044  2    M   73C    -0.45  0.19
```

```
X[animalesref,1:3]#primeros SNPs
```

```
##            rs3683945_G rs3707673_G rs6269442_G
## A048005080           1           1           1
## A048006063           1           1           2
## A084291787           1           1           1
## A084292044           0           2           0
```

```
A[animalesref,animalesref]
```

```
##            A048005080 A048006063 A084291787 A084292044
## A048005080          1          0          0          0
## A048006063          0          1          0          0
## A084291787          0          0          1          0
## A084292044          0          0          0          1
```

Ahora, dividimos la información para conformar conjunto de datos para el entrenamiento del modelo y para el análisis de prueba (200 animales). Al declarar los fenotipos del conjunto de prueba como perdidos (NA), el programa no los tendrá en cuenta en el entrenamiento del modelo, así, quedan definidos simultáneamente los dos conjuntos (entrenamiento y prueba):

```
tst=sample(1:length(y),size=200)
yNA=y
yNA[tst]=NA
```

Se crea matriz diseño con los efectos fijos de sexo y tamaño de camada y efecto aleatorio de compañeros de jaula:

```
XF=model.matrix(~as.factor(Sexo)+ as.factor(TC))[,-1]
dim(XF)


## [1] 1814    8


XR=model.matrix(~as.factor(Jaula)-1)
dim(XR)


## [1] 1814  552
```

Ajustamos un modelo con efectos fijos de sexo y tamaño de camada y efectos aleatorios de jaula, efectos de sustitución de los marcadores y efectos poligénicos usuales de cada animal cuya matriz de covarianzas es el producto de la varianza genética aditiva y la matriz de relaciones genéticas aditivas construída a partir del pedigrí.

En este modelo se busca que los efectos poligénicos capturen la variabilidad genética aditiva que no es explicada por los efectos de de los marcadores. FIXED especifica la distribución a priori de efectos "fijos", los efectos de jaula se modelan con una regresión Ridge (BRR), los efectos poligénicos de la manera usual que en este caso se especifica mediante una forma simplificada de un espacio de Hilbert de núcleo reproducible o RKHS por sus siglas en inglés (Reproducing Kernel Hilbert Space). Finalmente, la a priori asignada a los efectos de los marcadores moleculares es aquella que corresponde a bayes A:

```
ETA=list( list(X=XF,  model='FIXED'),
          list(X=XR, model='BRR'), #se podría usar otro
          list(K=A,  model='RKHS'),
          list(X=X,  model='BayesA')
)
```

Creamos el objeto fm que contiene los resultados de ajustar el modelo descrito a la base de datos de entrenamiento:

```
#inicialmente utilizaremos 200 iteraciones y 50 de arranque,
#después puede aumentarlas a 5000 con 2500 de arranque
#Puede utilizar otros valores
Iteracciones=200
Arranque=50
fm=BGLR(y=yNA,ETA=ETA,nIter=Iteracciones,
        burnIn=Arranque,saveAt='full_')
```

Generamos el resumen del modelo ajustado. A los efectos fijos se les asigna una a priori plana y los demás componentes son tal y como se describieron previamente:

```
summary(fm)
```

```
## --------------------> Summary of data & model <--------------------
##
##  Number of phenotypes= 1614
##  Min (TRN)=  -2.8
##  Max (TRN)=  3.1
##  Variance of phenotypes (TRN)= 1
##  Residual variance= 0.52
##  N-TRN= 1614  N-TST= 200
##  Correlation TRN= 0.8
##
##  -- Linear Predictor --
##
##  Intercept included by default
##  Coefficientes in ETA[ 1 ] (  ) were asigned a flat prior
##  Coefficientes in ETA[ 2 ] (  ) modeled as in  BRR
##  Coefficientes in ETA[ 3 ] (  ) were assumed to be normally distributed with zero mean and
##  covariance (or its eigendecoposition) provided by user
##  Coefficientes in ETA[ 4 ] (  ) modeled as in  BayesA
##
## ------------------------------------------------------------------
```

El siguiente parámetro indica cada cuantas iteraciones se guardan los valores de los parámetros:

```
fm$thin
```

```
## [1] 5
```

Número de iteraciones consideradas como calentamiento:

```
fm$burnIn
```

```
## [1] 50
```

Número total de iteraciones:

```
fm$nIter
```

```
## [1] 200
```

Histograma de las medias posteriores de los efectos de jaula (FIGURA NRO. 15.6) y estadísticas descriptivas:

```
hist(fm$ETA[[2]]$b,border="red", main=NULL,
   mar=c(bottom=0, left=1, top=1, right=1))
```

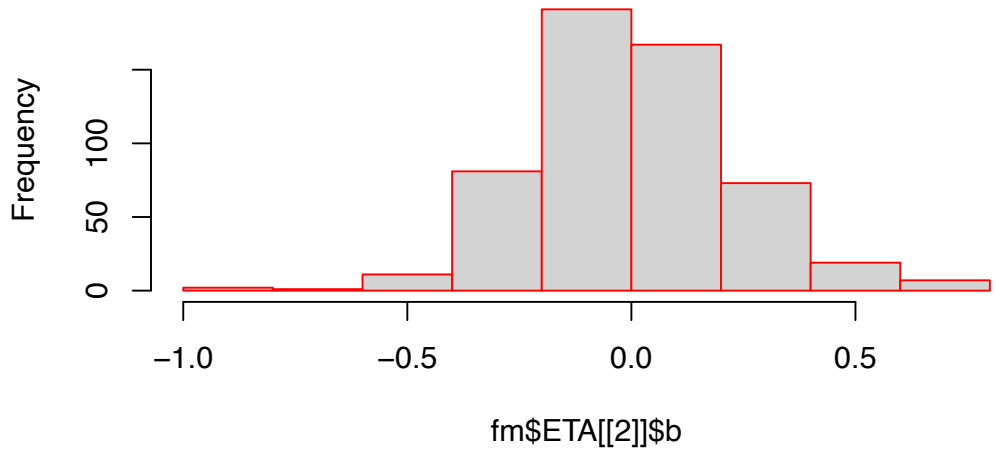

***Figura* 15.6:** Histograma de las medias posteriores de los efectos de jaula.
Fuente: elaboración propia generada en R-project [25].

Resumen del análisis:

```
round(summary(fm$ETA[[2]]$b),2)


##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.98   -0.14   -0.02    0.00    0.15    0.78
```

Histograma de las medias posteriores de los efectos poligénicos (FIGURA NRO. 15.7) y estadísticas descriptivas:

```
hist(fm$ETA[[3]]$u,border="blue",
   main=NULL,mar=c(bottom=0, left=1, top=1, right=1))
```

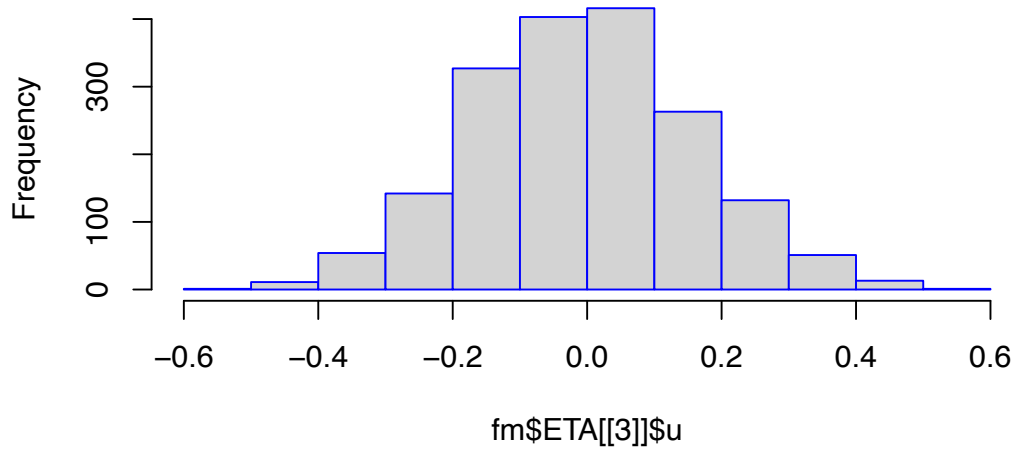

***Figura* 15.7:** Histograma de las medias posteriores de los efectos poligénicos. Fuente: elaboración propia generada en R-project [25].

Resumen:

```
round(summary(fm$ETA[[3]]$u),2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.50   -0.12   -0.01   -0.01    0.10    0.57
```

Histograma de las medias posteriores de los efectos de los marcadores (FIGURA NRO. 15.8) y estadísticas descriptivas :

```
hist(fm$ETA[[4]]$b,border="blue",
     main=NULL,mar=c(bottom=0, left=1, top=1, right=1))
```

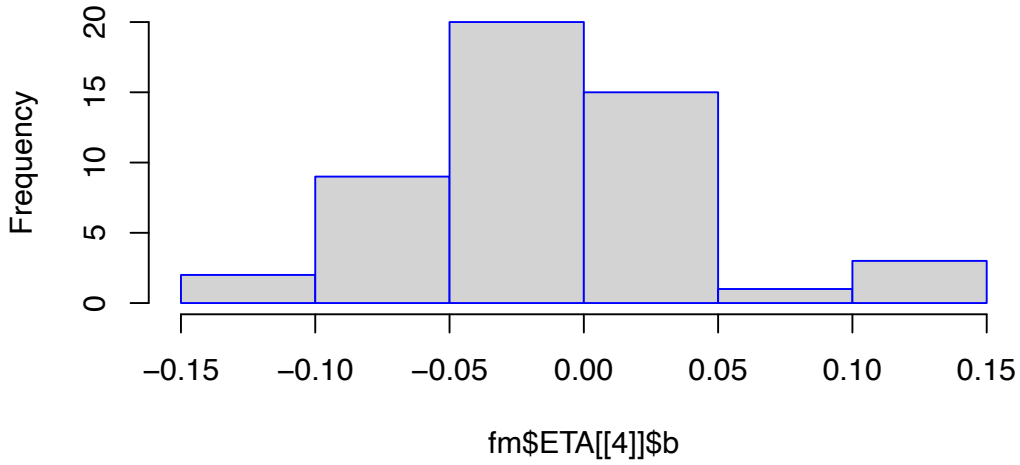

***Figura* 15.8:** Histograma de las medias posteriores de los efectos de los marcadores. Fuente: elaboración propia generada en R-project [25].

Estadísticas descriptivas de las medias posteriores:

```
round(summary(fm$ETA[[4]]$b),2)


##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.12   -0.04   -0.01   -0.01    0.02    0.13
```

Correlación entre el fenotipo observado y el predicho en el conjunto de prueba:

```
round(cor(y[tst],fm$yHat[tst]),2)


## [1] 0.61
```

Gráficos con los efectos cuadráticos de los marcadores (FIGURA NRO. 15.9):

```
plot( (fm$ETA[[4]]$b^2),xlab='SNP',
      ylab='Cuadrado del efecto',cex=.5,col=2,type='l',
      mar=c(bottom=0, left=1, top=1, right=1))
```

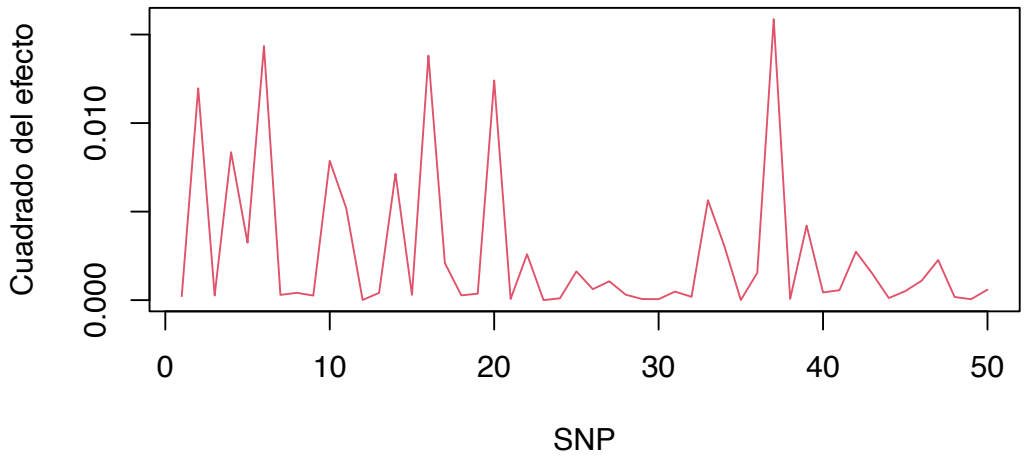

***Figura* 15.9:** Efectos cuadráticos de los marcadores.
Fuente: elaboración propia generada en R-project [25].

16 CAPÍTULO DIECISÉIS

ANEXO

16.1. Gnosa

Gnosa es un símbolo que representa la liberación de nuestras mentes para concentrarnos y potencializar nuestras habilidades en la escritura, la lectura y la oratoria. Nos permite concentrarnos en lo verdaderamente esencial de nuestros estudios y de nuestros proyectos de investigación.

Se dibuja repitiendo tres veces el infinito, el infinito vertical, el triángulo y el círculo. En nuestra programación en R-project, nos enfocamos en el número pi (movimiento de la naturaleza), su seno y su coseno. Cada figura tiene 73 divisiones (el 73 es un número primo).

Generación de las secuencias:

```
Divisiones=73
Veces=3
t=seq(-pi,pi,pi/(Divisiones/2));head(round(t,1))


## [1] -3.1 -3.1 -3.0 -2.9 -2.8 -2.7


t=rep(t,Veces)
```

Base de datos para graficar el infinito:

```
x = cos(t);head(round(x),1)


## [1] -1


y = sin(-2*t) / 2; head (round(y),1)


## [1] 0
```

Base para el triángulo triángulo:

```
dpi=pi/5
x1 =rep(c(seq(-dpi,dpi,dpi/(Divisiones/2)),
      rev(seq(0.0,dpi,dpi/Divisiones)),
      -1*(seq(0.0,dpi,dpi/Divisiones))),3)
y1=rep(c(rep(-dpi,Divisiones+1),
     seq(-dpi,dpi,dpi/(Divisiones/2)),
     rev(seq(-dpi,dpi,dpi/(Divisiones/2)))),3)
```

Base para el circulo:

```
r=.25;
x2=-r*cos(t);
y2=r*sin(t);
```

Base de datos con la secuencia completa (infinito, infinito vertical, triángulo y círculo):

```
X=c(x,NA,y,NA,x1,NA,x2)
Y=c(y,NA,x,NA,y1,NA,y2)
d = data.frame(x = X, y = Y)
```

Gráfico del símbolo Gnosa (FIGURA NRO. 16.1):

```
plot( X,Y,c("yellow","green","blue","red","pink"),
    mar=c(bottom=0, left=1, top=0, right=1))
```

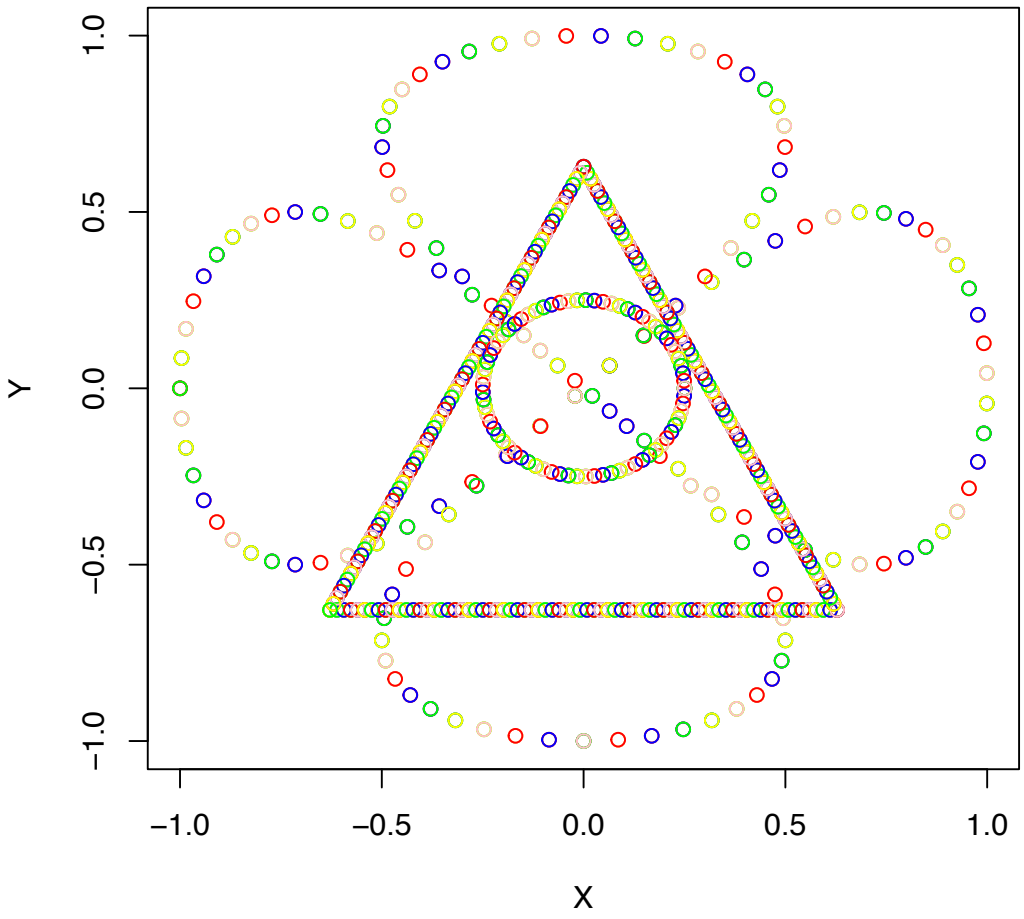

***Figura* 16.1:** Gnosa.
Fuente: elaboración propia generada en R-project [25].

REFERENCIAS

[1] Montaldo H, Barria N. Mejoramiento genético de animales. Ciencia al día [Internet]. 1998. [Consultado en marzo de 2024]; 2: 19. Dispobible en: http://www.ciencia.cl/CienciaAlDia/volumen1/numero2/articulos/articulo3.htm.

[2] San Primitivo Tirados F. La mejora genética animal en la segunda mitad del siglo XX. Arch. Zootec. 2001; 50: 517-546.

[3] Henderson C R. Applications of linear models in animal breeding.University of Guelph; 1984.

[4] Gianola D, Rosa G J. One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biosci. 2015; 3(1): 19-56.

[5] Henderson C R. Sire evaluation and gentic trends. J. Anim. Sci. [Internet]. 1973. [Consultado en marzo de 2024]; 10-41. Disponible en: https://doi.org/10.1093/ansci/1973.Symposium.10.

[6] Fernando R L, Gianola D. Optimal properties of the conditional mean as a selection criterion. Theor. Appl. Genet. 1986; 72(6): 822-825.

[7] Henderson C. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976; 32: 69-83.

[8] Martínez C A, Manrique Perdomo C, Elzo M A, Jiménez Rodriguez A. Additive genetic group and heterosis effects on growth and corporal composition of crossbred cattle in southern Cesar (Colombia. Rev. Colomb. de Cienc. Pecu. 2012; 25: 377-390.

[9] Searle S R. Linear models. Wiley Classics Library Edition; 1971.

[10] Littell R C, Milliken G A, Stroup W W. SAS for mixed models. SAS Institute; 2006.

[11] Elzo M A, Famula T R. Multibreed sire evaluation procedures within a country. J. Anim. Sci. [Internet]. 1985. [Consultado en marzo de 2024]; 60: 942-952. Disponible en: https://doi.org/10.2527/jas1985.604942x.

[12] Christensen R. Plane answers to complex questions: the theory of linear models. 4 ta ed. Springer; 2011.

[13] Kass R E, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. J. Am. Stat. Assoc. [Internet]. 1995. [Consultado en marzo de 2024];90: 928-934. Disponible en: http://www.jstor.org/stable/2291327.

[14] Henderson C R. Selection index and expected genetic advance. En: Statistical genetics and plant breeding. Editores: Hanson W D, Robinson H F. 982:141-163; 1963.

[15] Patterson H D, Thompson R. Recovery of Inter-Block Information When Block Size Are Unequal. Biometrika [Internet]. 1971. [Consultado en marzo de 2024]; 58: 545-554. Disponible en: http://dx.doi.org/10.1093/biomet/58.3.545.

[16] Kackar R N, Harville D A. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. J. Am. Stat. Assoc. [Internet]. 1984. [Consultado en marzo de 2024]; 79: 853-862. Disponible en: https://doi.org/10.1080/01621459.1984.10477102.

[17] Jiang J. On unbiasedness of the empirical BLUE and BLUP.Stat. Probab. Lett. 1999; 41: 19-24.

[18] Venables W, Ripley B. MASS: modern applied statistics with S. 4 ta ed. Springer [Internet]; 2002. [Consultado en marzo de 2024]. Disponible en: https://www.stats.ox.ac.uk/pub/MASS4/.

[19] Henderson C. Estimation of genetic parameters. Ann. Math. Stat. 1950; 21: 309.

[20] Kennedy B W. Animal model BLUP: Erasmus intensive graduate course, Trinity College, Dublin, 20-26 August 1989. Trinity College [Internet]; 1989. [Consultado en marzo de 2024]. Disponible en: https://books.google.com.co/books?id=ctX2oAEACAAJ.

[21] Mrode R A. Linear models for the prediction of animal breeding values. 2 da ed. CABI [Internet]; 2006. [Consultado en marzo de 2024]. Disponible en: 10.1079/9780851990002.0000.

[22] VanRaden P M, Wiggans G R. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 1991; 74(8):2737-2746.

[23] Kennedy B W, Schaeffer L R, Sorensen D A. Genetic properties of animal models. J. Dairy Sci. 1988; 17-26.

[24] Elzo M. Animal breeding notes: the mixed linear model; 2014. [Consultado en marzo de 2024]. Disponible en: https://animal.ifas.ufl.edu/elzo/animal_breeding_notes/uinbreed/011_abn94_2014.pdf.

[25] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing [Internet]; 2024. [Consultado en marzo de 2024]. Disponible en: https://www.R-project.org/.

[26] Solarte C E, Imuez A M, Pérez T. Modelo animal multicaracter para la estimación de parámetros genéticos del Cavia porcellus en Colombia. Revista Cubana de Ciencia Agrícola. 2002; 36: 19-24.

[27] Solarte C E, Soto F, Pérez T. Modelo animal multicarácter para la selección de reproductores Cavia porcellus en Colombia. Rev. Cub. de Cien. Agríc. [Internet]. 2002. [Consultado en marzo de 2024]; 36 (1): 25-29. Disponible en: https://www.redalyc.org/articulo.oa?id=193018091005.

[28] Sinnwell J, Therneau T. kinship2: pedigree functions. Comprehensive R Archive Network [Internet]; 2020. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=kinship2.

[29] Bates D, Maechler M. Matrix Models: modelling with sparse and dense matrices. Comprehensive R Archive Network [Internet]; 2021. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=MatrixModels.

[30] Wickham H. stringr: simple, consistent wrappers for common string operations. Comprehensive R Archive Network [Internet]; 2022. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=stringr.

[31] Solarte C, Zambrano G L. Characterization and genetic evaluation of Holstein cattle in Nariño, Colombia. Rev. Colomb. de Cienc. Pecu.2012; 25:539-547.

[32] Quaas R L, Anderson R D, Gilmour A R. BLUP School Handbook: Use of Mixed Models for Prediction and for Estimation of (co)variance Components. Animal Genetics and Breeding Unit, University of New England [Internet]; 1984. [Consultado en marzo de 2024]. Disponible en: https://books.google.com.co/books?id=klDVswEACAAJ.

[33] Willham R L. The role of maternal effects in animal breeding. J. Anim. Sci. 1972; 35:1288-1293.

[34] Quintanilla R, Piedrafita J. Efectos maternos en el peso al destete del ganado vacuno de carne: revisión. Itea. 2000; 96: 7-39.

[35] Cerón-Muñoz M F, Hurtado-Lugo N, Vergara O D. Evaluación genética de animales domésticos. En: Modelación Aplicada A Las Ciencias Animales: II Evaluaciones Genéticas. Editores: Elzo M A, Vergara O D. 2: 11-45; 2012.

[36] Henderson C. Theoretical basis and computational methods for a number of different animal models. J. Dairy Sci. 1988; 71: 1-16.

[37] Shull G. What is heterosis?. Genetics [Internet]. 1948. [Consultado en marzo de 2024]; 33: 439-446. Disponible en: https://doi.org/10.1093/genetics/33.5.439.

[38] Falconer D S, Mackay T F C. Introduction to quantitative genetics. 4 ta ed. Loagman group Ltd; 1996.

[39] Arnold J W, Bertrand J K, Benyshek L. Animal model for genetic evaluation of multibreed data. J. Anim. Sci. 1992; 70: 3322-3332.

[40] Højsgaard S, Halekoh U. doBy: groupwise statistics, LSmeans, linear estimates, utilities. Comprehensive R Archive Network [Internet]; 2023. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=doBy.

[41] Soller M, Beckmann J S. Restriction fragment length polymorphisms and genetic improvement. En: Proceedings of the 2nd World Congress on Genetics Applied to Livestock Production. Madrid. 6:396—404; 1982.

[42] Bullock K, Pollak E. Beef symposium: the evolution of beef cattle genetic evaluation. J. Anim. Sci. [Internet]. 2009. [Consultado en marzo de 2024]; 87: E1-E2. Dispobible en https://doi.org/10.2527/jas.2008-1738.

[43] Soller M. The use of loci associated with quantitative effects in dairy cattle improvement. Anim. Sci. [Internet]; 1978. [Consultado en marzo de 2024];27 (2): 133-139. Disponible en: 10.1017/S0003356100035960.

[44] Fernando R L, Grossman M. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. [Internet]. 1989. [Consultado en marzo de 2024]; 21: 467-477. Disponible en: https://doi.org/10.1186/1297-9686-21-4-467.

[45] Hill W G. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. Genetics. 2014; 196: 1-16.

[46] Misztal I. Challenges of application of marker assisted selection-a review. Anim. Sci. Pap. Rep. 2006; 24: 5-10.

[47] Meuwissen T H, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157: 1819-1829.

[48] VanRaden P M. Efficient methods to compute genomic predictions. J. Dairy Sci. 2008; 91 (11):4414-4423.

[49] Aguilar I, Misztal I; Johnson D L, Legarra A; Tsuruta S, Lawlor T J. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. [Internet]. 2010. [Consultado en marzo de 2024]; 93(2): 743-752. Disponible en: https://www.sciencedirect.com/science/article/pii/S0022030210715174.

[50] Gianola D, De los Campos G, Hill W, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics [Internet]. 2009.[Consultado en marzo de 2024]; 183: 347-363. Disponible en: https://doi.org/10.1534/genetics.109.103952.

[51] Habier D, Fernando R L, Dekkers J C M. The impact of genetic relationship information on genome-assisted breeding values. Genetics [Internet]. 2007.[Consultado en marzo de 2024]; 177: 2389-2397. Disponible en: https://pubmed.ncbi.nlm.nih.gov/18073436/.

[52] Goddard Mike. Genomic selection: prediction of accuracy and maximization of long term response. Genetics [Internet].2008. [Consultado en marzo de 2024]; 136: 245-257. Disponible en: 10.1007/s10709-008-9308-0.

[53] Zhong S, Dekkers J, Fernando R, Jannink J. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics. 2009; 182:355-364.

[54] Gianola D. Priors in whole-genome regression: the bayesian alphabet returns. Genetics [Internet]. 2013. [Consultado en marzo de 2024]; 194: 573-596. Disponible en: 10.1534/genetics.113.151753.

[55] Leon-Novelo L, Casella G. Prior influence in linear regression when the number of covariates increases to infinity. Stat. Probab. Lett. [Internet]. 2012 [Consultado en marzo 2024]; 438-445. Disponible en: 10.1016/j.spl.2011.10.018.

[56] Legarra A, Misztal I. Technical note: computing strategies in genome-wide selection. J. Dairy Sci. 2008; 91. 360-366.

[57] Wiggans G R, VanRaden P M, Cooper T A. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. [Internet]. 2011. [Consultado en marzo de 2024]; 94(6): 3202-3211. Disponible en: https://www.sciencedirect.com/science/article/pii/S0022030211003079.

[58] VanRaden P M, Van Tassell C P, Wiggans G R, Sonstegard T S, Schnabel R D, Taylor J F, Schenkel F S. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 2009; 92:16-24.

[59] Hayes B J, Bowman P J, Chamberlain A J, Goddard M E. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 2009; 92: 433-443.

[60] Legarra A, Aguilar I, Misztal, I. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 2009; 92: 4656-4663.

[61] Casella G, Berger R. Statistical Inference. 2 a ed. Thomson Learning; 2002.

[62] Colleau Jean-Jacques. An indirect approach to the extensive calculation of relationship coefficients. Genet. Sel. Evol. [Internet]. 2002. [Consultado en marzo de 2024]; 34: 409-421. Disponible en: https://doi.org/10.1186/1297-9686-34-4-409.

[63] Christensen O, Lund M S. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol [Internet]. 2010. [Consultado en marzo de 2024]; 42. Disponible en: https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-42-2.

[64] Lehmann E L, Casella G. Theory of point estimation. Springer Science & Business Media; 1998.

[65] Habier D, Fernando R L, Kizilkaya K; Garrick D J.Extension of the bayesian alphabet for genomic selection. BMC Bioinf. 2011; 12:186.

[66] Tibshirani R. Regression Shrinkage, selection via the Lasso. J. R. Stat. Soc. [Internet]. 996. [Consultado en marzo de 2024]; 58: 267-288. Disponible en: http://www.jstor.org/stable/2346178.

[67] Bishop C. Pattern recognition and machine learning. Springer; 2006.

[68] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction.2 a ed. Springer; 2009.

[69] Park T, Casella G. The bayesian lasso. J. Am. Stat. Assoc. [Internet]. 2008. [Consultado en marzo de 2024]; 103:681-686. Disponible en: https://doi.org/10.1198/016214508000000337.

[70] Erbe M, Hayes B J, Matukumalli L K, Goswami S, Bowman P J, Reich C M, Mason B A, Goddard M E. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 2012; 95(7): 4114-4129.

[71] Martínez C A, Khare K, Rahman S, Elzo M A .Gaussian covariance graph models accounting for correlated marker effects in genome-wide prediction. J. Anim. Breed. Genet. 2017; 134: 412-421.

[72] Martínez C A, Khare K, Rahman S, Elzo M A. Inferring the partial correlation structure of allelic effects and incorporating it in genome-wide prediction; 2017. [Consultado en marzo de 2024]. Disponible en: https://arxiv.org/abs/1705.02026.

[73] Martínez C A, Khare K, Rahman S, Elzo M A. Modeling correlated marker effects in genome-wide prediction via Gaussian concentration graph models .J. Theor. Biol. 2018; 437:67-78.

[74] Lauritzen S L. Graphical models. Clarendon Press [Internet]; 1996. [Consultado en marzo de 2024]. Disponible en: https://books.google.com.co/books?id=mGQWkx4guhAC.

[75] Wang C L, Ding X D, Wang J Y, Liu J F, Fu W X, Zhang Z, Yin Z J, Zhang Q. Bayesian methods for estimating GEBVs of threshold traits. Heredity [Internet]. 2013. [Consultado en marzo de 2024]; 110:213-219. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668646/.

[76] Lynch M, Walsh E. Genetics and analysis of quantitative traits. Sinauer Associates, Inc.; 1998.

[77] Harville D A. Matrix Algebra From a Statistician's Perspective. Springer [Internet]; 1997. [Consultado en marzo de 2024]. Disponible en: https://link.springer.com/book/10.1007/b98818.

[78] Searle S R. Matrix algebra for the biological sciences (including applications in statistics). Wiley [Internet]; 1966. [Consultado en marzo de 2024]. Disponible en: https://www.amazon.com/Biological-Sciences-Including-Applications-Statistics/dp/B0071IRNU4.

[79] Stewart J. Calculus: early transcendentals. 7 ma ed. Cengage Learning; 2012.

[80] Spivak M. Cálculus. 4 ta ed. Publish or Perish [Internet]; 2012. [Consultado en marzo de 2024]. Disponible en: https://www.academia.edu/39004372/Spivak_Michael_Calculus_2012_Revert%C3%A9_.

[81] Martínez C A, Khare K, Banerjee A, Elzo M A. Joint genome-wide prediction in several populations accounting for randomness of genotypes: A hierarchical Bayes approach. I: multivariate gaussian priors for marker effects and derivation of the joint probability mass function of genotypes. J. Theor. Biol. 2017; 417:8-19.

[82] Martínez C A, Khare K, Banerjee A, Elzo M A. Joint genome-wide prediction in several populations accounting for randomness of genotypes: A hierarchical Bayes approach. II: Multivariate spike and slab priors for marker effects and derivation of approximate Bayes and fractional Bayes factors for the complete family of models. J. Theor. Biol. 2017; 417:131-141.

[83] Martínez C A, Khare, K, Elzo M A. BIBI: Bayesian inference of breed composition. J. Anim. Breed. 2018; 135: 54-61.

[84] Cox R T. Probability, Frequency, and Reasonable Expectation. Amer. J. of Phys. [Internet]. 1946. [Consultado en marzo de 2024]; 14 (2):1–13. Disponible en: 10.2307/2272983.

[85] Cox R T. The algebra of probable inference. The Jhons Hopkins Press; 1961.

[86] Savage L J. The foundations of statistics. Courier Corporation; 1972.

[87] Wright S. The distribution of gene frequencies in populations. Proceedings of the National Academy of Sciences. 1937; 23(6): 307-320.

[88] Martínez C A, Khare K, Elzo M A. On the Bayesness, minimaxity and admissibility of point estimators of allelic frequencies. J. of Theo. Biol. 2015; 383: 106-115.

[89] Ghosh M. Objective Priors: An Introduction for Frequentists. Stat. Sci. [Internet]. 2011. [Consultado en marzo de 2024]; 26(2). Disponible en: https://doi.org/10.1214%2F10-sts338.

[90] Diaconis P, Ylvisaker D. Conjugate priors for exponential families. The Annals of statistics. 1979; 269-281.

[91] Lehmann E, Casella G. Theory of point estimation. Springer Science & Business Media; 2006.

[92] Robert C P, Casella G. Introducing monte carlo methods with r. Springer; 2010.

[93] Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R news. 2006; 6(1): 7-11.

[94] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. of the Royal Stat Society Series B: Statistical Methodology. 2009; 71(2):319-392.

[95] Nilforooshan M. ggroups: pedigree and genetic groups. Comprehensive R Archive Network [Internet]; 2022. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=ggroups.

[96] Bolker B, Warnes G, Lumley T. gtools: Various R programming tools. Comprehensive R Archive Network [Internet]; 2023. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=gtools.

[97] Westgate M, Grames E. synthesisr: Import, assemble, and deduplicate bibliographic datasets. Comprehensive R Archive Network [Internet]; 2020. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=synthesisr.

[98] Kimura M, Crow J. The Measurement of Effective Population Number. Evolution. 1963; 17(3): 279-288.

[99] MacCluer J W, Boyce A J, Dyke B, Weitkamp L R, Pfenning D W, Parsons Cindy J. Inbreeding and pedigree structure in Standardbred horses. J. of Hered. [Internet]. 1983. [Consultado en marzo de 2024]; 74 (6): 394-399. Disponible en: https://doi.org/10.1093/oxfordjournals.jhered.a109824.

[100] Cervantes I, Goyache F, Molina A, Valera M, Gutiérrez J P. Estimation of effective population size from the rate of coancestry in pedigreed populations. J. Anim. Breed Genet. [Internet]. 2011. [Consultado en marzo de 2024]; 128(1):56-63. Disponible en: https://pubmed.ncbi.nlm.nih.gov/21214645/.

[101] Wellmann R. optiSel: optimum contribution selection and population genetics. Comprehensive R Archive Network [Internet]; 2021. [Consultado en agosto de 2023]. Disponible en: https://CRAN.R-project.org/package=optiSel.

[102] Kennedy B W, Trus D. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 1993; 71(9): 2341-52.

[103] Sorensen D A, Kennedy B W. The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. Theor. Appl. Genet. 1983; 66:217-220.

[104] Elzo M A. Recursive procedures to compute the inverse of the multiple trait additive genetic covariance matrix in inbreed and noninbred multibred populations. J. Anim. Sci. 1989; 68: 1215-1228.

[105] Lo L, Fernando R, Grossman M. Covarianece between relatives in multibreed populations: additive model. Theor. Appl. Genet. 1993; 87: 423-430.

[106] Poulsen B, Ostersen T, Nielsen B, Christensen O. Predictive performances of animal models using different multibreed relationships matrices in systems with rotational crossbreeding. Genet. Sel. Evol. [Internet]. 2006. [Consultado en marzo de 2024]; 54(25):1-17. Disponible en: doi.org/10.1186/s12711-022-00714-w.

[107] García-Cortés L A, Toro M A. Multibreed analysis by splitting the breeding values.Genet. Sel. Evol. 2006; 38: 601-615.

[108] Strandén I, Mäntysaari E. Use of random regression model as an alternative for multibreed relationship matrix. J. Anim. Breed. Genet. 2013; 130:4-9.

[109] Muff S, Niskanen A, Saatoglu D, Keller L, Jensen H. Animal models with group-specific additive genetic variances: extending genetic group models. Genetic Selec. Evol. [Internet]. 2019. [Consultado en marzo de 2024]; 51. Disponible en: 10.1186/s12711-019-0449-7.

[110] Therneau T. bdsmatrix: routines for block diagonal symmetric matrices. Comprehensive R Archive Network [Internet]; 2024. [Consultado en marzo de 2024]. Disponible en: https://CRAN.R-project.org/package=bdsmatrix.

[111] Dawid A P. Conditional independence in statistical theory (with discussion). J. R. Stat. Soc. [Internet]. 1979. [Consultado en marzo de 2024]; 41: 1-31. Disponible en: https://www.jstor.org/stable/2984718.

[112] Gelfand A E, Sahu S K. Identifiability, improper priors, and gibbs sampling for generalized linear models. J. Am. Stat. Assoc. [Internet]. 1999. [Consultado en marzo de 2024]; 94: 247-253. Disponible en: https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473840.

[113] Putz A. Introduction to GBLUP and single-step GBLUP. Rpubs [Internet]. 2018. [Consultado en marzo de 2024]. Disponible en: https://rpubs.com/amputz/GBLUP_and_ssGBLUP.

[114] Pérez P, De los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014; 198: 483-495.

Autores

Carlos Eugenio Solarte Portilla (Guaitarilla, Nariño, Colombia, 1960)

Es Doctor en Ciencias Veterinarias de la Universidad Agraria de La Habana (Cuba), Zootecnista de la Universidad de Nariño (Colombia), y Profesor Asociado de la Universidad de Nariño, ha sido reconocido por la Presidencia SUE Caribe y SUE de Colombia en el 2022 "Por su invaluable trabajo y compromiso en la consolidación de la Educación Superior en Colombia y su sentido de pertenencia en la formación de centenares de estudiantes para ser profesionales de excelencia". Además, el Sistema Universitario Estatal lo ha reconocido y exaltado por su denodada, esforzada gestión y liderazgo desarrollado como Exvicepresidente del SUE y Exrector de la Universidad de Nariño, consolidando hechos significativos que contribuyen al fortalecimiento de la educación superior en Colombia, también en el año 2022. En el ámbito local, la Asamblea Departamental de Nariño lo ha reconocido y exaltado en el 2021, resaltando su vida y obra por los logros alcanzados en el desarrollo académico e investigativo. Asimismo, el Consejo Superior de la Universidad de Nariño reconoció su gestión como rector entre el 2014 y el 2021. También ha recibido reconocimientos de diversas facultades, programas, sedes regionales y dependencias universitarias por su gestión como rector en 2017 y 2021. Anteriormente, fue reconocido como Zootecnista Distinguido por la Asociación de Zootecnistas Egresados de la Universidad de Nariño en junio de 2007 y junio de 2008. Además, recibió reconocimiento por su trabajo en el Programa de Mejoramiento Genético Animal de Nariño, Universidad de Nariño, en febrero de 2010.

Ha publicado el libro *Bioestadística: Aplicaciones en Producción y Salud Animal* e(2003), y ha contribuido con un capítulo en el libro *Milk Protein* de la editorial Intech Open, editado por Walter Hurley. Su trabajo también ha sido indexado en el Book Citation Index (BKCI) en Web of Science Core Collection™ (2012) y ha publicado numerosos artículos en revistas especializadas nacionales e internacionales en mejoramiento animal.

Pertenece al Grupo de Investigación Producción y Salud Animal-Cuyes de la Universidad de Nariño.

csolarte@udenar.edu.co

Carlos Alberto Martínez Niño (Paipa, Boyacá, Colombia, 1985)

PhD en genética-estadística de la University of Florida, Gainesville, FL, y se graduó con honores como Zootecnista de la Universidad Nacional de Colombia, Sede Bogotá. Además, ha ejercido como Profesor Asistente en la Facultad de Medicina Veterinaria y de Zootecnia, Departamento de Producción Animal, Universidad Nacional de Colombia, Sede Bogotá. También ha desempeñado roles como Investigador PhD Asociado en la Corporación Colombiana de Investigación Agropecuaria – AGROSAVIA, Sede Central, hasta 2023, y como Profesor Asistente con dedicación cátedra en la misma facultad hasta 2023. Se destaca su publicación de más de 25 artículos en revistas especializadas nacionales e internacionales de su especialidad. Entre los reconocimientos recibidos se incluyen el Premio Grinter de la Universidad de la Florida en 2012, la Beca Francisco José de Caldas-Fulbright para estudios de doctorado en los Estados Unidos en el mismo año, y la participación en el Programa Jóvenes Investigadores e Innovadores de MINCIENCIAS entre abril de 2011 y abril de 2012. Asimismo, ha sido galardonado con diversos reconocimientos por su desempeño académico, incluyendo menciones de honor, becas y matrículas de honor durante su periodo de estudios universitarios en la Universidad Nacional de Colombia.

Sus publicaciones incluyen *Evaluación genética animal en poblaciones multirraciales* (2013), y es miembro del grupo de investigación en Recursos genéticos y biotecnología animal (A1).

camartinezn@unal.edu.co

Mario Fernando Cerón Muñoz (Pasto, Nariño, Colombia, 1972)

Doctor en Producción Animal de la Universidad Estadual Paulista (Brasil) y Zootecnista graduado de la Universidad de Nariño (Colombia). Actualmente, ejerce como Profesor Titular en la Universidad de Antioquia. Ha sido reconocido con varios premios y distinciones, entre los que se destacan los títulos de Doctor Honoris Causa en Andragogía y Educación Inclusiva, otorgados por la Organización Internacional para

la Inclusión y Calidad de la Educación en Chile en 2023 y en México en 2022 respectivamente. Además, ha sido premiado en la Universidad de Antioquia por su labor en extensión en 2023, excelencia docente en 2017 e investigación en 2010, este último acompañado de la Medalla Francisco José de Caldas.

Entre sus publicaciones se encuentran: *Manejo de información zootécnica en hatos lecheros. Evaluación genética de ganado Holstein de Antioquia* (2014) *Modelación aplicada a las ciencias animales: Diseño experimental, con implementación del programa R- Project* (2013); Modelación aplicada a las ciencias animales: Generalidades de R- Project (2013); *Modelos Aplicados a las Ciencias Animales: Evaluaciones genéticas* (2012), *Juzgamiento, clasificación y selección de ganado bubalino* (2011); *Modelos Aplicados a las Ciencias Animales: Genética cuantitativa* (2008). Forma parte del Grupo de Investigación Agrociencias Biodiversidad y Desarrollo GAMMA de la Universidad de Antioquia.

mario.ceron@udea.edu.co

La Editorial de la Universidad Tecnológica de Pereira tiene como política la divulgación del saber científico, técnico y humanístico para fomentar la cultura escrita a través de libros y revistas científicas especializadas.

Las colecciones de este proyecto son: Trabajos de Investigación, Ensayos, Textos Académicos y Tesis Laureadas.

Este libro pertenece a la Colección de Trabajos de Investigación

Este libro se enfoca en la aplicación de modelos lineales en evaluación genética animal. Se inicia con un recuento del desarrollo de la evaluación genética, con énfasis en la explicación de los métodos: el mejor predictor, el mejor predictor lineal y el mejor predictor lineal insesgado. Posteriormente, se presentan varios casos del modelo animal que obedecen a diferentes tipos de acción génica y tipos de datos. Además, incluye capítulos relacionados con los fundamentos de algebra matricial, probabilidad, introducción a la estadística Bayesiana e identicabilidad. También tiene capítulos relaciona-dos con conceptos y procedimientos básicos para el estudio previo de bases de datos. El libro se soporta con ejemplos desarrollados en R-project.

This book focuses on the application of linear models in animal genetic evaluation. It begins with an account of the development of genetic evaluation, with emphasis on the explanation of the methods: the best predictor, the best linear predictor, and the best unbiased linear predictor. Subsequently, several cases of the animal model that obey dierent types of gene action and types of data are presented. In addition, it includes chapters related to the fundamentals of matrix algebra, probability, introduction to Bayesian statistics, and identiability. It also has chapters related to basic concepts and procedures for the previous study of databases. The book is supported with examples developed in R-project.

Editorial UTP

Colección Trabajos de Investigación

eISBN: 978-958-722-930-1