Iniciación a la valoración genética animal. Metodología adaptada al EEES

Juan Pablo Gutiérrez García

Extensión

UCM EDITORIAL COMPLUTENSE

VALORACIÓN GENÉTICA ANIMAL.
JUAN PABLO GUTIÉRREZ

UCM EDITORIAL COMPLUTENSE

VALORACIÓN GENÉTICA ANIMAL.

Juan Pablo Gutiérrez

UCM EDITORIAL COMPLUTENSE

Todos los libros publicados por Editorial Complutense a partir de enero de 2007 han superado el proceso de evaluación experta.

Donoso Cortés, 63 – 4. planta (28015) Madrid
Tels.: 91 394 64 60/1 Fax: 91 394 64 58
e-mail: ecsa@rect.ucm.es
www.editorialcomplutense.com

Primera edición: Febrero 2010

ISBN: 978-84-9938-021-6
Depósito legal:

PRÓLOGO

Hasta ahora, año tras año, mis alumnos de Mejora Genética Animal de la Facultad de Veterinaria de la Universidad Complutense de Madrid me han preguntado por la bibliografía adecuada para estudiar la asignatura, y año tras año les he referido a mis propias fuentes advirtiendo al mismo tiempo que no existía un texto que tratase todo el contenido de la materia con la estructura que se presenta, ya que venía a ser un compendio de los capítulos de iniciación a muchos de ellos. Durante años he tratado de suplir esta carencia con pequeñas aportaciones, tablas, resúmenes, presentaciones, trasparencias, problemas y todo tipo de material que de forma concreta ayudaba a resolver parcialmente esta carencia. Comento esto en este prólogo ya que, si a alguien le resulta de ayuda este texto, es a mis alumnos a los que corresponde el principal agradecimiento. Su tesón en pedir material bibliográfico y sus siempre educadas reclamaciones, normalmente canalizadas a través de las reuniones de seguimiento docente, han provocado el mantenimiento de mi inquietud en este sentido. Un agradecimiento especial merece Andrea González Peña, alumna del curso 2007/2008, que escuchó mis quejas sobre la necesidad de

poseer un texto base en soporte informático tomado directamente de mis clases, para poder realizar modificaciones que permitiesen conseguir unos buenos apuntes. No queda aquí nada de lo que ella preparó inicialmente pero sin aquellos apuntes originales, este texto nunca habría visto la luz.

No pretende ser éste un texto para profesionales de la mejora genética. Será raro que ellos encuentren aquí algo que no supieran antes. Al contrario, se trata de un texto sin grandes pretensiones. Aunque se ha hecho un esfuerzo en evitar inconsistencias, no ha sido éste el primer objetivo de su desarrollo. Se ha intentado llegar con detalle a las manipulaciones que sufren los datos durante las valoraciones genéticas relacionándolas con las propiedades, las bondades o las utilidades de los métodos. Aunque se han incluido algunas demostraciones enmarcadas a lo largo del texto, en mi opinión, sólo los profesionales de la mejora genética, y no los estudiantes, están obligados a conocer estas demostraciones, pero han sido incluidas en el texto para satisfacer la curiosidad de los alumnos que acaban preguntando por ellas cuando han asimilado el resto del contenido.

El desarrollo del libro está pensado partiendo del absoluto desconocimiento de las bases necesarias por parte del lector, pero termina desembocando en el patrón general de los modelos más avanzados que se utilizan actualmente en la práctica. En algunos casos recomiendo que el lector reproduzca los ejemplos con ayuda de software que facilite las operaciones entre matrices. Esta forma de presentar el contenido permite así acceder a esta materia únicamente con el trabajo personal del lector adaptándose así a las enseñanzas encuadradas en el *Espacio Europeo de Educación Superior* (EEES). Todos los conocimientos necesarios procedentes de otras disciplinas se han ido incorporando en la medida que han ido siendo necesarias. Un pequeño recordatorio sobre estadística introduce el libro para poner al lector en situación de admitir conceptos relacionados con la misma. El álgebra de matrices, imprescindible en este contexto, también ha ido incorporándose a lo largo del texto cuando ha ido siendo necesario. Esta manera de ir introduciendo los conceptos forma parte de mi propio estilo, como también forma parte el reducidísimo grupo de citas

bibliográficas presentadas en este libro. Prácticamente todo el contenido ha pasado de mi cabeza a estos apuntes, y en muchos casos no estoy seguro de cómo llegaron allí. Además de esas pocas citas tengo que agradecer lo aprendido de mis profesores de licenciatura, doctorado y curso de especialización, mis colegas, mis directores de tesis y, como digo, sobre todo, de mis alumnos.

A mis colaboradores más estrechos, en primer lugar, por la lectura crítica de diferentes capítulos, pero sobre todo por sus incansables mensajes de entusiasmo, corresponde también un importante agradecimiento. Y como no, a Judith por su paciencia.

Finalmente quiero aclarar que este texto no pretende ser de lectura obligada, ni para estudiantes ni para profesionales. En mi opinión, a los primeros, los estudiantes, les será útil para adquirir conocimientos que habitualmente les resultan abstractos en su lucha por superar la materia durante sus estudios. A los segundos creo que sólo les será útil en su iniciación a la especialización en valoración genética animal. Desconozco si tendrá algo de éxito en el futuro entre unos o entre otros pero yo ya soy consciente de su primera meta alcanzada: el haber calmado una prolongada inquietud de su autor.

Juan Pablo Gutiérrez

ÍNDICE

INTRODUCCIÓN GENERAL

Históricamente el hombre ha observado variabilidad en los rendimientos de los animales de los que ha venido obteniendo recursos para su propio bienestar. Asimismo ha observado también cómo los animales con mejores rendimientos tendían a transmitir esa superioridad a sus descendientes por lo que la selección artificial ha sido una realidad desde antiguo.

Otro hecho que no escapa a la observación humana es que los rendimientos de los animales se ven afectados por las condiciones en que se desarrollan, de manera que muchas decisiones de selección no se han fundamentado únicamente en el propio rendimiento del animal, sino que se ha tenido en cuenta si sus parientes más cercanos corroboraban la impresión que se tenía inicialmente de él.

Un avance más permite observar que determinados caracteres se ven menos afectados por el ambiente que otros. En concreto, los caracteres relacionados con el tamaño parecen ser menos dependientes del ambiente. Si un animal no se alimenta suficientemente, parece que tiende a ser más delgado pero si sus padres tenían mucha alzada, el animal mantiene esa tendencia a tener una elevada alzada. Por el contrario, por mucho que se alimente a un animal con tendencia a crecer poco, el animal engrosará pero no se le logrará hacer crecer más. Otro tipo de caracteres, sin embargo, dependen mucho más del manejo que se proporciona a los animales. Tales caracteres parecen tener que ver con la esfera reproductiva. El hombre concluye que en este tipo de caracteres necesitará mucha más información para estar convencido de la bondad de un determinado animal. Así por ejemplo, el hecho de que una hembra quede gestante en el primer apareamiento no garantiza que esto vaya a suceder en sus demás celos y los de sus parientes, pero si este comportamiento se repite de esta manera, probablemente se trata de un individuo genéticamente bueno.

Otra observación importante es que algunos criadores tratan mejor a sus animales que otros, y también la estacionalidad afecta a las

producciones, por lo que no es posible comparar el rendimiento de los animales que no comparten ambiente. La elección de los animales como reproductores debería hacerse entonces únicamente de entre los que comparten el mismo ambiente, o, en caso de necesitar escoger globalmente, habría que buscar un ajuste previo de efectos no genéticos que se pueden controlar.

De estas observaciones se deducen varios hechos importantes en valoración genética:

- La variabilidad es fundamental en valoración genética. Si todos los individuos rindiesen por igual, no tendría sentido la selección genética ni mucho menos, por tanto, la valoración genética animal.
- Los rendimientos están influidos por el ambiente por lo que el valor genético aproximado a partir de ellos siempre estará sujeto a error. Esto justifica la necesidad de conocimientos estadísticos en la valoración genética.
- La variabilidad que se observa en los rendimientos no sólo tiene origen genético sino que existe una influencia ambiental que no afecta por igual a todos los caracteres. Es necesario un parámetro que permita medir esta mayor o menor influencia ambiental.
- La información de parientes es importante para valorar genéticamente a los animales. La cantidad de información de parientes necesaria no será la misma para todo tipo de caracteres.
- Los rendimientos de los animales deben ajustarse para aquellos efectos que influyen sobre los mismos y que se pueden controlar, como por ejemplo, la ganadería, la estación o el año de producción.

Este texto trata exactamente esto. Se comienza con unos conocimientos estadísticos previos necesarios. A continuación se establecen conceptos fundamentales en la valoración genética, particularmente el concepto de heredabilidad. Posteriormente se introducen los modelos lineales, su definición, la resolución del modelo fijo, para entrar después en la valoración genética de los animales a través de la predicción de efectos aleatorios. Se incluyen después los índices de selección como método clásico de

valoración genética que, aunque en desuso en la práctica, tantísimo valor didáctico poseen. Y finalmente se desarrolla la metodología BLUP, hoy el método actual de valoración genética a nivel internacional. Se concluye la parte teórica de este libro con modelos concretos de valoración genética muy utilizados en la actualidad.

Concluye el libro con una pequeña colección de problemas que permiten al lector entrenarse en la utilización de herramientas de valoración genética.

PRIMERA PARTE

CONCEPTOS ESTADÍSTICOS ÚTILES EN MEJORA ANIMAL

RESUMEN

Repaso de algunos conceptos estadísticos necesarios en mejora genética animal. Se sientan los principios más generales de la estadística como la definición de variable aleatoria, las diferencias entre población y muestra y la utilización de los parámetros obtenidos a partir de ésta para hacer inferencias sobre aquélla. Se continúa con la familiarización con la distribución normal y la teoría de la estimación de parámetros. Por su importancia en este contexto se desarrollan con especial detalle las ideas relacionadas con la variabilidad y su medida, la varianza, la covarianza, el análisis de varianza y el análisis de regresión. Otros conceptos de especial interés que aparecen en este capítulo se relacionan con la utilización de modelos en estadística, la interacción entre efectos o la definición de factores como fijos o aleatorios.

La estadística podría definirse como una ciencia de probabilidades y de errores. Así, en el contexto de la estadística, nada puede afirmarse sin asumir una cierta incertidumbre.

En general las ciencias exactas hacen uso de modelos deterministas o determinísticos, llamados así porque determinan con exactitud el valor de sus componentes, como por ejemplo:

$$x + 3 = 7$$

Según este modelo x sólo puede tomar el valor $x = 4$.

En cambio, la estadística hace uso de modelos probabilísticos según los cuales el valor de sus componentes no es exacto sino que está sujeto a un término de error:

$$x + 3 = 7 + e$$

En este modelo el valor de x es aún desconocido, pudiendo únicamente establecer que su valor se encuentra en un intervalo que encierra $x = 4$, pero en el que el valor que x podría valer 4,1 o 3,9, y si estamos dispuestos a asumir mayor error podríamos también afirmar que vale por ejemplo 2 o 6, o incluso -1000 o 2000, siempre que estuviéramos dispuestos a asumir el correspondiente error.

Así pues, las conclusiones estadísticas que obtengamos van a depender de tres factores:

- Tamaño de la muestra (Nº de datos).
- Variabilidad de la variable.
- Error que estemos dispuestos a asumir.

I-1. Variable aleatoria

La definición formal de variable aleatoria, o simplemente variable, hace referencia a conceptos matemáticos demasiado complejos

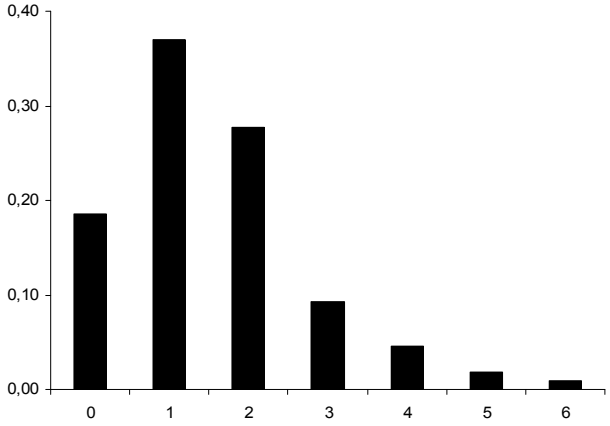
para lo que necesitamos saber en el ámbito de la mejora genética animal. La idea que subyace en este concepto es la existencia de una distribución de probabilidad de manera que el valor de una variable no es único sino que es el resultado del proceso de extracción de un valor de su distribución de probabilidad. Por ejemplo, el resultado de tirar una moneda no trucada al aire es una variable aleatoria ya que no se conoce el resultado; sin embargo se conoce su distribución de probabilidad, es decir, se sabe que en la mitad de las ocasiones saldrá cara y en la otra mitad saldrá cruz. Por tanto, de una variable sabemos:

- Que no conocemos el valor antes de medirla.
- Sí que conocemos su distribución de probabilidad, o al menos su tipo de distribución.

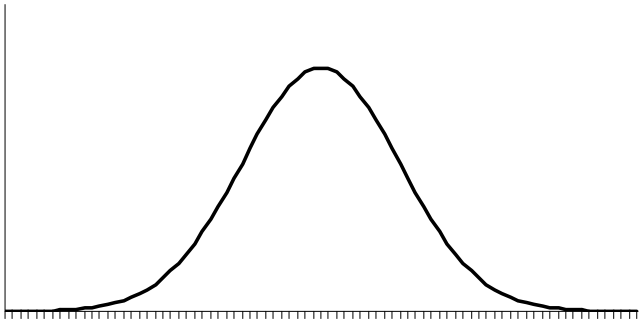
Una variable sería por ejemplo también el resultado de tirar un dado, desconozco lo que saldrá pero sé que hay 1/6 de posibilidades de que salga un 1; 1/6 de que salga un 2, etc. Pero también sería una variable aleatoria los valores procedentes de una distribución estadística típica como por ejemplo la distribución normal.

I-1.1. Tipos de variables aleatorias

- Discretas: Sólo pueden tomar valores enteros. Ejemplo: nº de corderos al parto de cada oveja. Una oveja podrá tener 0,1, 2, etc. corderos, pero no podrá tener uno y medio o 2,3. Su representación descriptiva típica es la función de cuantía que es más popularmente entendida como histograma de frecuencias relativas. Si conocemos la frecuencia con la que se dan los distintos tipos de parto, podemos dibujar un histograma que muestre la probabilidad de tener un determinado número de corderos por camada:



- Continuas**:** Pueden tomar cualquier valor dentro de su rango. Ejemplo: los litros de leche que da una cabra al día. Como en el caso anterior, habrá cabras que den 0 litros, 1, 2, etc., pero en este caso también es posible que obtengan valores intermedios: 1,5 3,4 4,35 litros, etc. En este caso no tiene sentido presentar una barra vertical para cada uno de los valores posibles ya que estos son infinitos, por lo que se representan mediante funciones de densidad:



En general trabajaremos con variables de tipo continuo cuya distribución es por tanto una función de densidad conocida como Distribución Normal o campana de Gauss, que quedará descrita con los dos parámetros que la definen como son el parámetro de posición (la media μ) y el parámetro de escala (la varianza σ^2), de manera que para decir que una variable x se distribuye

normalmente (o según una Normal) con media μ y varianza σ^2 se expresa:

$$x \sim N(\mu, \sigma^2)$$

Los valores que describen la ordenada (*f(x)*) de esta distribución se obtendrían simplemente dando valores a *x* en su rango (desde menos infinito a más infinito, aunque sólo toma valores significativos próximos a la media) en la expresión:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

I-2. Población y muestra

Una población está constituida por elementos, individuos, que poseen una característica común en la que estamos interesados. Por tanto, el interés del investigador es lo que define la población. Si estamos interesados en la estatura de los individuos españoles mayores de 18 años, un granadino de 17 años no pertenece a la población. Si estamos interesados en la estatura de los madrileños mayores de 18 años, un granadino de 20 años pertenecería a la población del primer ejemplo pero no del segundo. Asimismo, si estamos interesados en la estatura de los madrileños varones mayores de 18 años, entonces una madrileña de 20 años no pertenece a la población, pero sí a las dos poblaciones definidas en primer lugar.

Los parámetros que describen las poblaciones son fijos en un momento determinado, dado que contienen valores únicos. Así, aunque la estatura de los madrileños varones mayores de 18 años es variable (no todos miden lo mismo), los parámetros que miden su distribución sí lo son. Si contáramos con la estatura de todos los individuos de la población, bastaría sumarlas, dividir por el número de individuos en la población y obtendríamos la media. Este valor

es único y por tanto fijo. Los parámetros que describen las poblaciones suelen expresarse con letras griegas. La media se suele expresar como μ y la varianza como σ^2.

Desafortunadamente es habitual no poder contar con todos los elementos de la población en la que estamos interesados por lo que nos vemos obligados a estudiarlas a partir de un subconjunto de la población que se convierte así en una muestra aleatoria de la población. Para que las muestras puedan ser útiles en el estudio de las poblaciones, deben cumplir dos condiciones:

1- Deben ser aleatorias. Es decir, cualquier elemento de la población debe tener la misma probabilidad de entrar en nuestra muestra, para lo cual, deben escogerse al azar entre todos los de la población. Así, si para estudiar la cantidad de dinero disponible que tiene un alumno de veterinaria en el bolsillo en un momento cualquiera decidimos coger una muestra de alumnos, para lo cual nos situamos en la entrada de la cafetería y, tras comprobar que la persona que entra está matriculada en asignaturas de la licenciatura, le preguntamos la cantidad de dinero que lleva consigo. Obsérvese que no todos los alumnos de la población (estudiantes de veterinaria) tienen la misma probabilidad de entrar en la cafetería. Por ejemplo, aquéllos que no van a la facultad normalmente podrían estar trabajando al mismo tiempo, lo que implicaría que seguramente llevan más dinero en el bolsillo que una persona que no trabaje. Esta forma de proceder provocaría resultados sesgados.

2- Deben ser representativas. La muestra debe tener la misma representación que la población. Así, si en veterinaria el 30% de los alumnos son de primero, el 20% de segundo, el 20% de tercero , el 15% de cuarto, y el 15% de quinto, en nuestra muestra el número de individuos debe guardar la misma proporción. Esta condición suele cumplirse con una muestra suficientemente grande cuyos elementos se hayan escogido al azar.

La muestra, al igual que la población, tiene unos parámetros que la describen, como son su media ($\bar{x}$) y varianza (V). En general los parámetros de las muestras suelen expresarse con letras corrientes mientras que los parámetros de la población se expresan con letras griegas. Otra diferencia importante entre los parámetros muestrales y los poblacionales es que, mientras que los parámetros de la población son fijos, los parámetros muestrales son a su vez variables aleatorias. Esto es porque sólo hay una población posible mientras que existen infinidad de posibles muestras aleatorias de cada población, acompañando a cada una de ellas sus parámetros descriptivos, es decir, su media y su varianza. Se dice así que los parámetros muestrales no coinciden con los poblacionales por efecto de muestreo.

	Media	Varianza	Deviación Típica	Parámetros
Población	μ	σ^2	σ	Fijos
Muestra	$\bar{x}$	V	DT	Variables

A partir de los parámetros muestrales (o combinaciones de ellos) podemos inferir valores aproximados de los parámetros poblacionales. Este procedimiento se conoce con el nombre de estimación y está ligado con el concepto de valor esperado o Esperanza Matemática. La definición formal de esperanza matemática de una variable x es la siguiente:

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx$$

Es decir, es la media de los distintos valores de x ponderada por su probabilidad de aparición (*f(x)*), de manera que para distribuciones asumidas como simétricas, el valor de la esperanza matemática coincide con la media de la población cuando se trata de una variable aleatoria.

Algunas propiedades de la esperanza matemática:

- Esperanza de la suma. Es la suma de las esperanzas:

$$E(x_1+x_2) = E(x_1) + E(x_2)$$

- Esperanza de una constante. Es igual a la constante:

$$E(k) = k$$

- Esperanza de la media de la población. Es un caso particular de la anterior:

$$E(\mu) = \mu$$

- Esperanza de una constante por una variable. Es igual a la constante: por la esperanza de la variable:

$$E(kx) = k\,E(x)$$

I-3. Medidas de centralidad y medidas de dispersión

Para caracterizar la distribución de una variable suelen utilizarse dos medidas, una que refleja el valor en torno al cual se agrupan la mayoría de las observaciones y otro que refleja el grado de dispersión de los datos alrededor de ese valor. Para lo primero se utilizan las medidas de centralidad y para lo segundo las medidas de dispersión.

I-3.1. Medidas de centralidad

Las más conocidas son:

- Media aritmética o simplemente media**.** Se obtiene a partir de la suma de las observaciones dividida por el número de datos:

$$\bar{x} = \frac{\sum x_i}{n}$$

- Moda**:** Es aquél valor que aparece con mayor frecuencia, el valor más frecuente, el que está de moda.

- Mediana: Es el valor que queda en medio al ordenar los datos de menor a mayor. Si el número de datos es par, entonces quedan dos valores en medio; entonces sería la media aritmética de los dos valores centrales.

En distribuciones normales simétricas estas tres medidas coinciden en el mismo valor. Aunque hay otras medidas de centralidad, éstas son las más utilizadas.

I-3.2. Medidas de dispersión

La más utilizada es la varianza y combinaciones de la misma:

- Varianza (V cuando se refiere a la muestra o σ^2 cuando se refiere a la población): Se define como la media de los cuadrados de las desviaciones de los datos con respecto a la media y mide el grado de dispersión. El numerador de una varianza se conoce como suma de cuadrados (SC) y su cálculo puede llevarse a cabo sin necesidad de calcular la media y cada una de las desviaciones de los datos con respecto a la media

$$\sigma^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n} = \frac{SC}{n} = \frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n}$$

donde *S.C.* es la suma de cuadrados.

- Desviación típica (DT cuando se refiere a la muestra o σ cuando se refiere a la población). La varianza es una medida de variabilidad en el cuadrado de la unidad de medida de la variable. Por ello, para obtener una medida de variabilidad de la variable en las unidades del carácter, la varianza se transforma a desviación típica simplemente calculando su raíz cuadrada:

$$DT = \sigma = \sqrt{V} = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{SC}{n}} = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n}}$$

- Desviación típica, error típico, desviación estándar y error estándar son sinónimos. Típico y estándar son dos traducciones de la misma palabra. Existe también una explicación para la utilización de desviación o error. En ocasiones la variable de trabajo es un estimador de un parámetro. Por ejemplo, cuando uno desea estimar la media de la población (μ) a partir de una muestra, utiliza la media de la muestra ($\bar{x}$) como estimador de μ. Así como μ es fijo y tiene un valor único, $\bar{x}$ no lo es y tomará diferentes valores si se hacen diferentes muestras. Así, la variabilidad de $\bar{x}$ mide el error con el que se estima μ. Ésta es la razón por la que la desviación típica, cuando hace referencia a variables que son estimadores, se suele llamar error típico, dado que mide de alguna manera el error con que $\bar{x}$ estima μ. Volveremos más adelante sobre el valor de esta varianza de la media muestral.

- Coeficiente de variación. (CV). Tanto la varianza como la desviación típica presentan el inconveniente de presentar valores absolutos por lo que no es fácil interpretar si determinado valor se puede consultar alto o bajo. Para resolver este problema, ña variabilidad se suele expresar en relación a la media en forma de coeficiente de variación:

$$CV = \frac{\sigma}{\mu}\ ; \qquad \text{o} \qquad CV\ (\%) = \frac{\sigma}{\mu}100$$

I-3.3. Medidas de codispersión

- Covarianza (σ_{xy}): La covarianza entre dos variables *x* e *y*, mide la variación conjunta de dos variables y se calcula de manera muy similar a la varianza:

$$\sigma_{xy} = CoV(x,y) = CoV_{xy} = \frac{\sum (x_{i-}\bar{x})(y_i - \bar{y})}{n} =$$

$$= \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n}$$

Para su cálculo son necesarias parejas de datos, de manera que para cada valor de la variable *x* debe existir un valor para la variable *y*, siendo *n* el número de parejas de datos. Mide la relación que tienen las dos variables. Obsérvese que, al contrario que la varianza, la covarianza puede tomar valores negativos. Si las dos variables son independientes la covarianza entre ellas vale cero. Si la covarianza es positiva cuando una de ellas tiende a valer más la otra también, como por ejemplo el peso y la estatura. Del mismo modo, si la covarianza es negativa cuando una de ellas tiende a valer más la otra tiende a valer menos, como por ejemplo la cantidad de leche que produce una vaca y su porcentaje de grasa. Cuando dos variables son independientes su covarianza es cero. Obsérvese que la covarianza de una variable consigo misma (*x* = *y*) es igual a la varianza de la variable.

- Coeficiente de correlación de Pearson o simplemente correlación (*r* cuando se refiere a la muestra o ρ cuando se refiere a la población). Es un coeficiente que mide la relación existente entre dos variables a partir del cálculo de su variación conjunta. Por lo tanto, proporciona la misma información que la covarianza, pero en este caso el coeficiente es adimensional (presenta valores estandarizados entre -1 y 1), lo que facilita su interpretación. Al igual que la varianza, la covarianza presenta el problema de expresarse en términos absolutos por lo que no se sabe si determinado valor debe considerarse alto o bajo. El coeficiente de correlación resuelve este problema. Su expresión es la siguiente:

Correlación muestral: $r = \frac{CoV(x,y)}{\sqrt{V_x V_y}}$, o

poblacional: $\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Sus valores se encuentran por tanto en el rango entre 1 y -1. Los valores próximos a los extremos se obtienen en variables con mucha relación mientras que los valores próximos a cero sugieren muy poca relación. Así, cuando dos variables son independientes, tanto su covarianza como su correlación son igual a cero.

I-3.4. Propiedades de la varianza

Las propiedades que se presentan a continuación se deducen de la aplicación de las expresiones anteriores para cada uno de los casos.

- Varianza de la suma: Es la suma de las varianzas más el doble de la covarianza entre ellas:

$$V(x + y) = V(x) + V(y) + 2\,CoV(x,y)$$

o en notación poblacional:

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

NOTA: En general adoptaremos la notación muestral cuando afrontemos un desarrollo algebraico, aunque trabajemos con parámetros poblacionales.

- Varianza de la diferencia: Es la suma de las varianzas menos el doble de la covarianza entre ellas:

$$V(x - y) = V(x) + V(y) - 2\,CoV(x,y)$$

Obsérvese que cuando las dos variables son independientes su covarianza es nula ($CoV(x,y) = 0$) y en ese caso la varianza de la suma y de la diferencia son iguales entre sí, e iguales a la suma de las varianzas:

$$V(x + y) = V(x - y) = V(x) + V(y)$$

- Varianza de una constante: Una constante no tiene varianza por definición:

$$V(k) = 0$$

- Varianza de la media de la población: Recordemos que la media de la población es un parámetro fijo, luego su valor es constante:

$$V(\mu) = 0$$

- Varianza de una constante por una variable: Es igual a la constante al cuadrado multiplicada por la varianza de la variable:

$$V(kx) = k^2 V(x)$$

  Esta expresión es útil en los cambios de escala. Imaginemos por ejemplo la variable producción de leche en mililitros, y deseamos transformarla de manera que la consideremos medida en litros. Ya que la nueva variable es igual a la primera dividida por mil, la varianza de la nueva variable transformada sería igual a la varianza de la original dividida por un millón.

- Varianza de la media muestral: En cambio, la media de la muestra es una variable aleatoria. Podría estimarse la varianza de la media muestral a partir de diferentes medias obtenidas mediante la repetición del experimento. Sin embargo, el valor de esta varianza se puede deducir:

$$V\left(\bar{x}\right) = V\left(\frac{\sum x_i}{n}\right) = \frac{1}{n^2} V(x_1 + x_2 + \ldots + x_n)$$

En la expresión anterior se ha sacado de la varianza la constante 1/n al cuadrado. La varianza de la suma de varias variables sería igual a la suma de las distintas varianzas más el doble de las covarianzas entre cada dos de ellas. Si la muestra con la que se ha calculado la media está bien obtenida, todas las observaciones son independientes entre sí, y por tanto, todas las covarianzas son nulas. Además, la varianza de todas y cada una de ellas será igual a la varianza de la variable dado que se trata de realizaciones obtenidas de la misma distribución. En adelante, debe recordarse que la varianza de una realización de una variable es igual a la varianza de la distribución a la que pertenece:

$$V\left(\bar{x}\right) = \frac{1}{n^2}\left[V(x_1) + V(x_2) + \ldots + V(x_n)\right] =$$
$$= \frac{1}{n^2}\left(V_x + V_x + \ldots + V_x\right) = \frac{1}{n^2} n V_x = \frac{V_x}{n}$$

Es decir, no es necesario hacer varias muestras del mismo tamaño para obtener distintas medias y así calcular la varianza de la media, sino que ésta es igual a la varianza de la variable partido por el número de datos de la muestra.

- Error típico o error estándar de la media. Como se dijo en un apartado anterior, estos términos pueden considerarse sinónimos de los términos desviación estándar y desviación típica. Sin embargo, suelen emplearse estos cuando miden la variabilidad de medias muestrales dado que esta variabilidad informa sobre el error con el que se estima la media poblacional. Así, el error típico de la media muestral es su desviación típica:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

El error típico sería entonces igual a la desviación típica de la variable partido por el número de datos. Obsérvese que este error de estimación depende de dos de los tres factores que condicionan las conclusiones estadísticas, la variabilidad de la variable y el número de datos. Obsérvese también que a medida que aumentamos el número de datos de la muestra este error es cada vez más pequeño, y si dispusiéramos de los infinitos elementos de la población, entonces este error sería igual a cero, obviamente dispondríamos de la auténtica media poblacional, y esa, ya sabemos que tiene valor fijo.

I-4. La distribución normal de media cero y varianza uno: N(0,1)

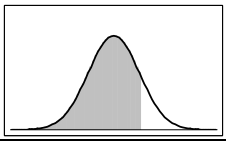
La distribución normal de media cero y varianza una ha sido ampliamente estudiada, de manera que se conoce perfectamente. Se sabe qué área de la curva encierra esta distribución desde cualquier valor del eje de abscisas hasta cualquier otro incluidos sus límites, los infinitos negativo y positivo. Se conoce también a la perfección la ordenada en cada valor del eje de abscisas y los dos valores (excepto en el punto máximo que lógicamente se encuentra en cero al ser éste el valor de la media) del eje de abscisas correspondientes a cada valor en el eje de ordenadas, valores que obviamente corresponden al mismo valor absoluto con distinto signo.

Así pues, existen tablas donde se pueden buscar los valores necesarios. La tabla que se presenta a continuación proporciona en las celdas interiores, el área (sombreado en la figura) que queda a la izquierda del umbral marcado por los valores de las coordenadas de ka celda, es decir, la suma de su fila y de su columna.

Así, si por ejemplo uno desea conocer los umbrales de un intervalo simétrico que encerrase el 95% de los datos, habría que pensar en dejar a la derecha del umbral de la figura el 2,5% de los datos, de manera que ese umbral con signo contrario dejaría el otro 2,5% de

las observaciones fuera del intervalo por el lado izquierdo. Así, el área que hay que buscar en el gráfico sería 0,95 + 0.025 = 0,975. En la tabla ese valor se encuentra en la fila 1,9 y columna 0,06. Así pues, se concluye que el 95% de los datos de una distribución normal de media cero y varianza uno se encuentran entre los valores -1,96 y 1,96. Evidentemente el usuario es libre de buscar otros intervalos caprichosos.

Tabla 1.1. Áreas bajo la curva normal estándar. Los valores representan la probabilidad de observar un valor menor o igual a z. La cifra entera y el primer decimal de z se buscan en la primera columna, y el segundo decimal en la cabecera de la tabla.



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

I-4.1. Tipificación de una variable

La ventaja de disponer de una distribución normal tabulada para valores estándar ($z_i \sim N(0, 1)$) nos permite tener puntos de referencia para otras. Así, el hecho de que una variable se distribuya normalmente nos proporciona una gran cantidad de información sobre la misma, ya que una variable que se distribuya normalmente con media μ y varianza σ^2 ($x_i \sim N(\mu, \sigma^2)$) puede ser fácilmente transformada en la primera. Para ello la variable debe tipificarse, lo que se consigue restando a cada dato la media y dividiendo por la desviación típica.

Así, si x se distribuye normalmente con media μ y varianza σ^2 ($x_i \sim N(\mu, \sigma^2)$), entonces:

$$\boxed{\frac{x_i - \mu}{\sigma} \sim N(0,1)}$$

Del mismo modo, una variable que se distribuya con media cero y varianza uno puede ser transformada a otra con parámetros caprichosos. Si por ejemplo la variable z se distribuye ($z_i \sim N(0, 1)$) y deseamos transformarla a otra con media μ y varianza σ^2, basta con hacer la operación inversa:

$$z_i\sigma + \mu \sim N(\mu, \sigma^2)$$

I-4.2. Intervalos de confianza

Una aplicación interesante de disponer de todos los valores tabulados de la distribución normal $N(0, 1)$, es la construcción de intervalos de confianza. Como ha sido ya repetidamente comentado, la media muestral $\bar{x}$ es una variable aleatoria distribuida normalmente, de manera que posee una media (la media de las infinitas posibles medias muestrales, es decir, μ) y una varianza (la varianza de la media muestral, es decir, σ^2/n), por lo

que puede ser tipificada. Así, $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ es una variable tipificada, y por lo tanto son aplicables en ella todos los valores tabulados de la distribución normal *N(0, 1)*. Por ejemplo, se sabe el 95% de las medias muestrales tipificadas estarán entre -1,96 y +1,96, o lo que es lo mismo, que la probabilidad de que la media tipificada se encuentre entre estos dos valores es del 95%:

$$P\left\{-1,96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1,96\right\} = 0,95.$$

En el otro 5% se encontrará fuera del intervalo, de manera que si afirmamos que la media tipificada se encuentra entre -1,96 y +1,96, estamos cometiendo un error del 5%. Este error se conoce como error de tipo I y se representa por α expresado en tanto por uno, en este caso α=0,05.

Si queremos generalizar esta expresión para cualquier error, y llamando a $z_{\alpha/2}$ al umbral que deja $\alpha/2$ de área a cada lado del intervalo, la expresión anterior se puede generalizar:

$$P\left\{-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1-\alpha$$

Haciendo unas sencillas operaciones en esta expresión, se puede transformar en:

$$P\left\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Es decir, que la probabilidad de que la verdadera media μ se encuentre entre esos dos valores es de 0,95, o lo que es lo mismo, el intervalo que deja fuera al 5% de las medias muestrales de tamaño n es:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Obsérvese que el intervalo depende de los tres factores que condicionaban las conclusiones estadísticas: el número de datos (n), la variabilidad de la variable (σ), y del error que estemos dispuestos a asumir (α).

Una cosa más antes de acabar este capítulo. El error α debe plantearse antes de comenzar el experimento. No es ortodoxo decidir el nivel de significación de las pruebas a la vista de los resultados. Además debe tenerse en cuenta que el error α no es el único error en el que debemos fijarnos. Es claro que este error es el que se comete al dejar fuera del intervalo algunos valores que corresponderían al verdadero valor, pero existe un error de tipo II o error β, que depende también de α, y que representa los valores que se aceptan por quedar dentro del intervalo cuando corresponden a elementos de una población diferente. No se tratará en más detalle aquí la utilización de los errores α y β aunque este capítulo tiene un elevado interés para determinar las pruebas de las potencias estadísticas.

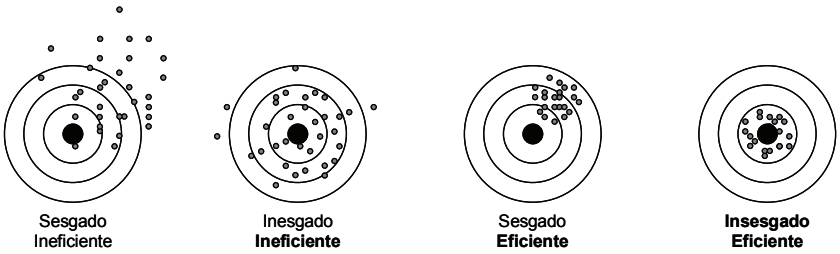
I-5. Estimación de parámetros

Pueden ser muchas las propiedades que se exijan a un estimador. En general todos los estimadores deben cumplir al menos las dos siguientes propiedades:

- Debe ser Insesgado. Un estimador insesgado es aquél cuyo valor esperado coincide con el verdadero valor del parámetro.

- Debe ser Eficiente. Es decir, debe tener mucha precisión, o dicho de otra manera, tener poco error. Recordemos que el error lo mide la dispersión, es decir, la varianza del estimador o su desviación típica, llamada con frecuencia para este propósito error típico.

Ilustraremos el significado de estas dos propiedades mediante un ejemplo. Supongamos que van dos amigos a una feria a intentar ganar un premio disparando a una diana. Uno de los amigos es un buen tirador (**eficiente**) porque sistemáticamente acierta cerca de donde marca la mirilla del rifle, mientras que el otro es un mal tirador (Ineficiente). Ambos prueban dos rifles distintos, uno de ellos trucado (Sesgado), que sistemáticamente apunta fuera del centro de la diana mientras que el otro está perfectamente calibrado (**insesgado**). El aspecto que presentan las dianas es el siguiente:



Obsérvese que lo ideal es que el estimador presente ambas propiedades pero que en ocasiones es mejor que no presente ninguna a que presente sólo una de ellas. En la figura se puede observar que en el primero de los casos (sesgado e ineficiente) el tirador acierta con el centro de la diana, mientras que en el tercer caso (sesgado pero **eficiente**), el tirador nunca acierta.

En general los estimadores se representan con el mismo símbolo que el parámetro coronados con un sombrero. Así, por ejemplo, el estimador de la media poblacional se representaría por $\hat{\mu}$.

I-5.1. Estimador de la media poblacional

El estimador de la media poblacional es habitualmente la media muestral, es decir, se asume $\hat{\mu} = \bar{x}$. Aunque no lo demostraremos aquí, se trata de un estimador insesgado porque $E\left(\bar{x}\right) = \mu$. En cuanto a la eficiencia, ésta se mide de forma inversa por su error, el cuál es: $\sigma_{\hat{\mu}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Otra vez, obsérvese que a medida que aumenta el número de datos el estimador es más eficiente.

I-5.2. Estimador de la varianza poblacional

El estimador de la varianza de la población no es la varianza de la muestra:

$$\widehat{\sigma^2} \neq V = \frac{\sum\left(x_i - \bar{x}\right)^2}{n}$$

V sería un estimador sesgado ($E(V) \neq \sigma^2$). La razón es que la expresión contiene información redundante. En el numerador se están utilizando los *n* datos de que disponemos junto con la media general, por lo que sólo *n-1* pueden considerarse verdaderamente independientes, al ser posible deducir el otro por disponer de la media. Se define así el concepto de grados de libertad (gl), como el número de observaciones verdaderamente independientes. Así, el estimador de la varianza de la población sería igual a la suma de cuadrados dividido por el número de grados de libertad:

$$\widehat{\sigma^2} = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}$$

Éste sí sería un estimador insesgado de la varianza de la población. Así pues, la diferencia en obtener una varianza dividiendo la suma de cuadrados por el número de datos o por este valor menos uno,

se define en el objetivo por el que se realiza la operación. Si se desea simplemente conocer la dispersión de una serie de valores, debe dividirse por *n*, pero si se desea conocer la variabilidad de una variable a partir de una muestra de datos, en realidad lo que se desea es estimar la varianza de la población, por lo que debe dividirse por *n-1*. Obsérvese también, que para muestras de tamaño razonablemente grande estos valores no difieren mucho; la razón es que una muestra de tamaño muy grande puede considerarse en sí misma una población.

I-6. Contraste de hipótesis

Cualquier análisis estadístico debe empezar con la definición de la hipótesis de partida. Normalmente el investigador sospecha sobre la influencia de determinado factor sobre determinada variable de manera que diseña un experimento que le permita comprobar esta hipótesis. Sin embargo, la forma de pensar es estadísticamente la contraria, es decir, la hipótesis que debe plantearse de partida hace referencia al desconocimiento que hay sobre la influencia del factor hasta antes de iniciar el experimento. En otras palabras, la hipótesis de partida nos dice que el factor NO tiene influencia sobre la variable o que NO hay diferencias entre las medias de los grupos, por lo que se llama Hipótesis Nula y se representa por H_0.

Así pues, el experimento se plantea siempre con la intención de refrendar la hipótesis nula, de manera que, concluido el experimento, si efectivamente se confirma esta hipótesis, éste queda acabado. Sin embargo, antes de comenzar el experimento debemos tener preparada una hipótesis alternativa a la que se desea refrendar por si tras el análisis de los datos tuviéramos que rechazar la H_0. Esta hipótesis se nombra exactamente de esa manera, Hipótesis Alternativa, se representa por H_A, y debe cubrir el resto de las situaciones que no contemple la H_0. Por ejemplo, si la H_0 dice que no hay diferencias entre dos medias, entonces la H_A dice que sí las hay. Sin embargo, si la H_0 dice que no hay diferencias entre más de de dos medias, la H_A no puede decir que sí las hay, ya que no sabríamos dónde encuadrar por ejemplo el caso en que hay dos no diferentes entre sí pero diferentes a su vez

de una tercera. Es claro que en este caso la H_A debe decir que "al menos una media es diferente de las demás". En este caso, tras aceptar la H_A, la práctica habitual es proceder a continuación a comprobar si hay diferencias entre las medias de dos en dos.

I-7. Análisis de la varianza

El análisis de varianza (ANOVA o ADEVA) es una técnica estadística que presenta dos posibles utilidades:

1. Comparar series de medias.
2. Descomponer la varianza total en una serie de componentes.

I-7.1. Introducción al análisis de varianza

Tenemos la sospecha de que determinado factor, como la edad, el sexo, la ganadería, puede tener influencia sobre el valor de un determinado carácter. Diseñamos un experimento en el que asignamos *J* unidades, individuos o elementos a cada uno de los *I* distintos niveles o grupos del factor y recogemos el valor del carácter. Cada clase diferente de un factor del modelo se conoce también como nivel del factor. Pongamos por ejemplo que tres ($I = 3$) tratamientos farmacológicos pueden estar influyendo sobre la fertilidad de las cerdas y sometemos a cuatro de ellas ($J = 4$) a cada uno de los tratamientos. Los datos que observamos son los siguientes:

	T_1	T_2	T_3
I grupos = 3 J datos por grupo = 4	5	6	4
	4	7	3
	1	2	1
	2	1	0
$x_{i.}$ =	**12**	**16**	**8**
$\bar{x}_{i.}$ =	**3**	**4**	**2**

$x_{..}$= 36 y $\bar{x}_{..}$= 3

NOTACIÓN:

- Cada observación será referida como x_{ij} donde el primer subíndice, *i*, indica el grupo o tratamiento en el que está, mientras que el segundo, *j*, corresponderá al orden de la observación dentro del tratamiento. Por ejemplo, la tercera observación del segundo tratamiento se indicaría como x_{23} y en este caso x_{23} = 2.
- La suma de las observaciones dentro de un tratamiento se identifican porque el segundo subíndice se sustituye por un punto, indicando el primero a qué tratamiento se refiere la suma: $x_{i.} = \sum_j x_{ij}$, existiendo tantas sumas parciales como grupos. Por ejemplo la suma de los datos del tercer grupo vale 8 se indica $x_{3.}$ = 8.
- La media de cada grupo o tratamiento se indica como la suma pero con una barra sobre la *x*: $\bar{x}_{i.} = \frac{\sum_j x_{ij}}{J}$, y nuevamente hay tantas medias parciales como grupos. Que la media del segundo grupo es 4 se indica $\bar{x}_{2.} = 4$. Obviamente las sumas y medias globales se señalan respectivamente como $x_{..}$ y $\bar{x}_{..}$.

I-7.2. Fundamentos del análisis de varianza

Como fue indicado en el capítulo 6, todo análisis estadístico parte de una hipótesis de partida o hipótesis nula (H_0). En el análisis de varianza dicha hipótesis nula establece que "no hay diferencias entre las medias de los grupos". Entonces, bajo la H_0, existen dos maneras distintas de obtener la varianza de la variable:

1. A partir de la varianza de los datos dentro de los grupos ("varianza interna" o "varianza dentro")
2. A partir de la varianza entre las medias de los grupos, la cuál, según el apartado I.3.4. sería equivalente a la varianza de la variable partido por el número de datos de cada muestra, en este caso *J*.

Las diferencias en las dos varianzas de la variable así obtenidas, sólo se deben a efecto de muestreo bajo la H_0, de manera que si la

varianza entre las medias superase el valor esperado bajo efecto de muestreo, habría que rechazar la H_0, y aceptar la H_A. Es importante destacar que en este caso la no puede establecer que "hay diferencias entre las medias de los grupos" porque debe cubrir todo el espectro no cubierto por la H_0, y no sabríamos a qué hipótesis correspondería el caso en el que, por ejemplo, dos medias no fueran diferentes entre sí pero si lo fueran de una tercera. La debe H_A establecer en este caso que "al menos una media es diferente de las demás".

Veamos ahora cómo se desarrolla en la práctica la comparación.

Varianza interna: Corresponderá al cociente entre la suma de cuadrados de los datos con respecto a la media de su grupo, y los grados de libertad correspondientes. Dado que en el numerador de esta componente se dan por conocidas todas las medias de los grupos, el número de grados de libertad será el número de datos por grupo (*J*) por el número de grupos (*I*), es decir, *IJ*, menos el número de medias utilizadas, es decir, el número de grupos (*I*), o lo que es lo mismo *J*-1 grados de libertad por cada uno de los *I* grupos:

$$\widehat{\sigma^2} = V_I = \frac{\sum\left(x_{ij} - \overline{x_{i.}}\right)^2}{IJ - I} = \frac{\sum\left(x_{ij} - \overline{x_{i.}}\right)^2}{I(J-1)}$$

Así pues, si la varianza interna (V_I) así calculada estima la varianza externa, entonces el valor esperado de la V_I será la varianza de la variable:

$$E(V_I) = \sigma^2$$

Varianza externa: Se trata de estimar una varianza entre medias utilizando las propias medias como si fueran datos. Es decir, corresponderá al cociente entre la suma de cuadrados de las I medias con respecto a la media global, y los grados de libertad correspondientes, que es el número de medias menos uno:

$$\widehat{\sigma^2_{\bar{x}}} = V_E = \frac{\sum\left(\overline{x_{i.}} - \overline{x_{..}}\right)^2}{I-1}$$

Estas dos componentes, bajo la hipótesis nula están entonces relacionadas de la siguiente manera: $\sigma^2_{\bar{x}} = \frac{\sigma^2}{J}$, de modo que $J\sigma^2_{\bar{x}} = \sigma^2$, o lo que es lo mismo, $JV_E = V_I$, y con infinitos datos el cociente $\frac{JV_E}{V_I} = 1$, pero bajo la hipótesis alternativa existe una componente de varianza adicional que la causa el agrupamiento que se hace de los datos, y entonces $V_E = \widehat{\sigma^2_{\bar{x}}} = \frac{\widehat{\sigma^2}}{J} + \widehat{\sigma^2_G}$, donde σ^2_G es la varianza de la variable que es consecuencia de que los datos pertenezcan a distintos grupos, o simplemente la varianza debida a grupos.

En términos de esperanza matemática, se puede entonces escribir:

$$E\left(V_E\right) = \sigma^2_{\bar{x}} = \frac{\sigma^2}{J} + \sigma^2_G$$

Dado que por efecto de muestreo no disponemos de los verdaderos valores de las varianzas sino de sus estimadores, debemos recurrir a una prueba estadística que nos permita obtener conclusiones, aunque siempre con un cierto margen de error. Este margen de error nos lo proporciona la distribución F que está tabulada para el cociente entre cuadrados medios con diferentes grados de libertad. Así para comprobar si el cociente $\frac{JV_E}{V_I}$ puede considerarse superior a 1 para los grados de libertad de que disponemos, comparamos el valor del cociente con el que se obtiene de las tablas:

$$F_c = \frac{JV_E}{V_I} \sim F^{I-1}_{IJ-I,\alpha}$$

Si $F^{I-1}_{IJ-I,\alpha} > \frac{JV_E}{V_I}$ SE ACEPTA $H_{0,}$ no hay diferencias entre los distintos tratamientos, y por tanto las diferencias, son consecuencia del efecto de muestreo, o lo que es lo mismo, se deben al azar.

Si $F^{I-1}_{IJ-I,\alpha} \leq \frac{JV_E}{V_I}$ SE RECHAZA H_o Y SE ACEPTA H_A, las diferencias no las explica únicamente el azar, por lo que hay variabilidad en los datos dependiendo del tipo de tratamiento, o, lo que es lo mismo, al menos una de las medias de los grupos es diferente del resto.

I-7.3. Ecuación del Modelo

En general, utilizaremos modelos lineales sencillos en los que la variable de trabajo o variable dependiente, se explique en función de la suma de otros efectos, factores o variables independientes. En este contexto usaremos los vocablos factor y efecto de forma indistinta para expresar los elementos del modelo de la parte derecha, es decir, aquéllos de los que depende el valor de la variable dependiente x_{ij}. La ecuación del modelo correspondiente a los datos proporcionados en la introducción sería la siguiente:

$$x_{ij} = \mu + T_i + e_{ij}$$

En la ecuación del modelo así planteada, se especifica que el valor de cualquier dato (x_{ij}) será aproximadamente la media del carácter (μ), pero no todos los datos valen lo mismo, sino que podrán tener una influencia originada por el grupo o tratamiento al que pertenecen (T_i), y no todos los datos de cada grupo son iguales tampoco, sino que existe una parte de la variabilidad no explicada por el modelo que agrupamos en un cajón de sastre, conocido como error o residuo (e_{ij}), y que agrupa cualquier causa de variabilidad no tenida en cuenta en el modelo, como por ejemplo,

otros efectos no considerados, errores de medida, variabilidad individual, etc.

I-7.4. Clasificación de análisis de varianza en función del diseño de los datos

I-7.4.1. Análisis de Varianza Jerárquico, Jerarquizado o anidado.

Los datos se encuentran organizados dentro de efectos con estructura de árbol.

I-7.4.1.1. Análisis de Varianza Jerárquico Simple

Se corresponde con el ejemplo de la introducción del apartado I.7.1.

Ecuación del modelo:

$$x_{ij} = \mu + T_i + e_{ij}$$

I-7.4.1.2. Análisis de Varianza Jerárquico Doble

Los datos están jerarquizados a un factor que a su vez está jerarquizado a otro. Supongamos por ejemplo que un número determinado de verracos se aparea cada uno con un número determinado de cerdas distintas, y se pesan los lechones de cada camada. La estructura de los datos sería de la siguiente manera:

♂1			♂2			♂3		
♀11	♀12	♀13	♀21	♀22	♀23	♀31	♀32	♀33
x_{111}	x_{121}	x_{131}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
x_{112}	x_{122}	x_{132}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
x_{113}	x_{123}	x_{133}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-

En este caso el número de verracos es el número de grupos (I = 3), el número de cerdas es el número de subgrupos dentro de cada grupo (J = 3), y el número de lechones por camada es el número de observaciones dentro de cada subgrupo (K=6).

NOTACIÓN:

- Cada observación es en este modelo referida como x_{ijk} donde el primer subíndice, *i*, indica el grupo o tratamiento en el que está, el segundo, *j*, indica el subgrupo dentro del grupo, y el tercero, *k*, corresponderá al orden de la observación dentro del subgrupo. Por ejemplo, la tercera observación de la segunda cerda del primer macho se indicaría como x_{123}.
- La suma de las observaciones dentro de un subgrupo se identifican porque el tercer subíndice se sustituye por un punto, indicando el primero y el segundo a qué subgrupo dentro de qué grupo se refiere la suma: $x_{ij.} = \sum_{k} x_{ijk}$, existiendo tantas sumas parciales con dos subíndices como subgrupos en total.
- La suma de las observaciones dentro de un grupo se identifican porque tanto el segundo como el tercer subíndice se sustituyen por un punto, indicando el primero a qué grupo se refiere la suma: $x_{i..} = \sum_{jk} x_{ijk}$, existiendo tantas sumas parciales de este tipo como grupos en total.
- Del mismo modo que en el jerárquico simple, las medias se expresan como las sumas con una barra superior, y habrá *IJ* medias distintas de subgrupos $\left(\overline{x}_{ij.} = \frac{\sum_{k} x_{ijk}}{K} = \frac{x_{ij.}}{K} \right)$, y I medias distintas de

grupos $\left(\overline{x}_{i..} = \dfrac{\sum_{jk} x_{ijk}}{JK} = \dfrac{x_{i..}}{JK} \right)$.

- Obviamente las sumas y medias globales se señalan respectivamente como $x_{...}$ y $\overline{x}_{...}$.

Ecuación del modelo:

$$x_{ijk} = \mu + V_i + C_{ij} + e_{ijk}$$

En esta ecuación del modelo el dato x_{ijk} corresponde al lechón que ocupa la *k*-ésima posición en la camada de la *j*-ésima cerda (C_{ij}) que se apareó con el *i*-ésimo verraco (V_i).

Obsérvese cómo los análisis de varianza jerárquicos pueden ser útiles en estudios genéticos, ya que las diferencias dentro de los subgrupos que no fueran debidas al azar serían originados por las hembras, e igualmente ocurriría con los grupos en relación a los machos, siendo por tanto esas diferencias de origen genético.

I-7.4.1.3. Análisis de Varianza Jerárquico Múltiple

No existe en la teoría ningún grado de jerarquía limitante, procediéndose en la notación con la misma lógica que los otros diseños jerarquizados previamente expuestos.

I-7.4.2. Análisis de Varianza Factorial

Mientras que en los análisis de varianza jerarquizados hay un factor principal, y puede haber otros de inferior jerarquía, en el análisis de varianza factorial existe más de un factor de igual jerarquía. Por ejemplo, imaginemos que en un ejemplo similar al anterior 3 cerdas se reprodujeran consecutivamente con todos los machos. En este caso no tendría sentido el segundo subíndice para distinguir por ejemplo entre la hembra ♀$_{12}$ y la ya ♀$_{22}$, dado que se

trataría de la misma hembra. El gráfico de este modelo sería más bien así:

	♂1	♂2	♂3
♀1	x_{111} x_{112} x_{113} -	x_{211} x_{212} x_{213} -	x_{311} x_{312} x_{313} -
♀2	x_{121} x_{122} x_{123} -	x_{221} x_{222} x_{223} -	x_{321} x_{322} x_{323} -
♀3	x_{131} x_{132} x_{133} -	x_{231} x_{232} x_{233} -	x_{331} x_{332} x_{333} -

Ecuación del modelo:

$$x_{ijk} = \mu + V_i + C_i + e_{ijk}$$

En esta ecuación del modelo el dato x_{ijk} corresponde al lechón que ocupa la *k*-ésima posición en la camada de la *j*-ésima cerda (C_i) que se apareó con el *i*-ésimo verraco (V_i), pero a diferencia del modelo jerarquizado, no es preciso un segundo subíndice para identificar con qué cerda se cruza cada macho ya que todas las hembras se prueban con todos los machos y viceversa, siendo ambos efectos de igual jerarquía.

NOTACIÓN:

- La suma de las observaciones dentro del efecto que agrupa los datos por columnas se identifica porque el segundo y el tercer subíndices se sustituyen por un punto $\left(x_{i..} = \sum_{jk} x_{ijk} \right)$**, mientras que la suma de los datos dentro del efecto que agrupa los datos por filas se identifica en este caso porque son el primer y el tercer subíndices los que se sustituyen por**

puntos: $\left(x_{.j.} = \sum_{ik} x_{ijk} \right)$.

- El resto de las notaciones es completamente deducible con la misma lógica empleada hasta aquí.

I-7.4.2.1. Análisis de Varianza Factorial con interacciones

En ocasiones, en los factores de igual jerarquía, determinados niveles o clases de un efecto no se comportan igual dependiendo del nivel del otro factor. Por ejemplo, podría ocurrir que determinada hembra produjera lechones de mayor tamaño cuando se cruza con determinado verraco, mientras que para otra hembra ocurriese justo al contrario. Esto se conoce con el nombre de interacción entre efectos y debe incluirse en los modelos correspondientes. Aunque trataremos más adelante esta cuestión, se adelanta la ecuación del modelo que incluyese la interacción entre verraco y cerda (VC_{ij}):

$$x_{ijk} = \mu + V_i + C_i + VC_{ij} + e_{ikj}$$

I-7.4.3. Análisis de Varianza Mixto

Los análisis de varianza mixtos combinan efectos jerarquizados y factoriales. No existe una limitación teórica sobre el número de factores a incluir ni sobre su diseño.

I-7.5. Tipos de Factores

La clasificación de los factores como fijos o aleatorios tendrá más trascendencia en adelante, cuando se desee utilizar metodología basada en modelos lineales para la valoración genética animal, aunque ya aquí, en el ANOVA, su diferenciación es clave en la parte final del análisis, cuando se pretenden extraer conclusiones.

En general, los factores de los que depende la variable de trabajo pertenecerán siempre a una de estas dos categorías, efectos fijos y efectos aleatorios.

- **Factores fijos:**

Son aquéllos que presentan pocos niveles en la población y todos están contemplados en nuestros datos. Un ejemplo clásico de efecto fijo es el sexo de los animales en la variable peso al nacimiento. Es un factor que sólo puede presentar dos niveles o categorías, sexo macho y sexo hembra, y en nuestros datos, aunque sólo son una muestra de todos los datos de la población, aparecen animales que pertenecen a uno de esos dos niveles, o son machos, o son hembras. Conceptualmente se asume que el pertenecer a una de las categorías del efecto representa una diferencia fija con respecto a los otros niveles del efecto, aunque esta diferencia es desconocida. Por ejemplo, se asume que las hembras pesarán sistemáticamente k kilogramos menos que los machos. Por esta razón se suelen llamar también efectos sistemáticos.

- **Factores aleatorios:**

Son aquéllos que presentan muchos niveles (hasta el punto de asumir conceptualmente que hay infinitos) en la población y en nuestra muestra sólo están contemplados una muestra representativa de los mismos, considerados como una representación aleatoria del resto. El ejemplo más sencillo es el efecto individuo cuando se hace un experimento para analizar cualquier variable. En la muestra no están representados los infinitos animales posibles sino que se escogen unos cuantos como unidades experimentales. Conceptualmente se asume que todos los niveles pertenecen a una distribución, generalmente normal, con media y varianza, pero el valor esperado de cada uno de los niveles no es un valor sistemático diferente de los otros niveles, sino que el valor esperado de todos ellos es el mismo, el de la media de su distribución.

La clasificación de un determinado efecto o factor como fijo o aleatorio no siempre es sencilla. Un ejemplo clásico es el efecto ganadero, efecto que aporta generalmente alta variabilidad a los

datos. Existen determinadas cuestiones que facilitan la clasificación:

a) Si al repetir el experimento cojo los mismos niveles se trata de un efecto fijo. Por ejemplo, si se desea comparar la velocidad de los caballos de Pepe con los caballos de Juan, al repetir el experimento también cogeremos caballos de Pepe y Juan, porque lo que queremos saber es los de quién corren más, pero no me interesa ningún otro ganadero. Sin embargo, si mi objetivo es conocer si el manejo influye en la velocidad de los caballos, en el 2º experimento puedo coger algún caballo de Juan, de Pepe, o de ninguno.

b) Si las inferencias se pretenden limitar a los niveles contemplados entonces es fijo, pero si pretendo extender las conclusiones que obtenga al resto, entonces es aleatorio.

La definición de los efectos como fijos o como aleatorios al realizar un ANOVA influye sobre el tipo de conclusiones que se obtendrán al final. Tal y como se dijo en la presentación del capítulo, esta técnica sirve para comparar series de medias o para descomponer la varianza total en una serie de componentes. Pues bien, cuando se considera como fijo el ANOVA sirve para comparar series de medias, utilidad que se le suele dar al experimentar con un nuevo fármaco o probar una nueva ración de crecimiento, o comparar varias ya existentes. Sin embargo, cuando se considera como aleatorio, el ANOVA sirve para descomponer la varianza total en una serie de componentes. Ésta será su utilización más frecuente en genética cuantitativa, materia en la que es fundamental conocer qué parte de la variabilidad que se observa es de origen genético de cara a obtener respuesta por selección artificial.

I-7.6. Tabla de un ANOVA jerárquico simple

Construir una tabla ANOVA consiste sencillamente en detallar y resumir la información necesaria para llevar a cabo las comparaciones descritas en el apartado 7.2. En otras palabras, se trata de reunir la información que permita obtener las componentes V_I y JV_E definidas en el apartado correspondiente para facilitar el análisis.

La primera columna del ANOVA identificará los efectos que son causa de variación en los datos. La ecuación del modelo jerárquico simple del apartado 7.4.1.1. muestra tres sumandos a la derecha de la igualdad, de los cuáles, uno, μ, es el mismo para todos los datos, y, por tanto, no aporta variación a los datos. De los otros, el identificado como T_i es fuente de variación y ha sido incluido en la tabla como fuente de variación "EXTERNA" pero es corriente indicar directamente en la tabla el nombre del efecto, como por ejemplo, "Tratamiento", "Sexo" o "Ganadería". El otro elemento del modelo, e_{ij}, es un elemento presente en todos los modelos, y siempre identifica la fuente de variación de inferior jerarquía, la que define lo que se conoce normalmente como varianza residual, por lo que suele expresarse indistintamente como σ_e^2 o como σ^2.

La siguiente columna representa el denominador de V_I y JV_E. Atendiendo a comentarios previos, los grados de libertad de la fuente de variación externa serán el número de medias menos mismo (I – 1), mientras que serán J-1 por cada grupo en la fuente de variación interna, lo que hace un total de $I(J-1)$.

El numerador de V_I y JV_E corresponde a la siguiente columna, es decir, son las sumas de cuadrados. Para ello haremos una descomposición de una desviación de un dato con respecto a la media general, en otras dos, una que corresponde a la desviación de cada dato con respecto a la media de su grupo y otra a la desviación de la media del grupo respecto a la media general, las cuáles identifican respectivamente las desviaciones interna y externa:

$$x_{ij} - \overline{x_{..}} = \left(x_{ij} - \overline{x_{i.}}\right) + \left(x_{i.} - \overline{x_{..}}\right)$$

Tomando sumatorios a ambos lados de la igualdad y elevando al cuadrado, después de reordenar la expresión en la que se eliminan todos los dobles productos se llega a separar la suma de cuadrados total en otras dos componentes:

$$\sum\left(x_{ij} - \overline{x_{..}}\right)^2 = \sum\left(x_{ij} - \overline{x_{i.}}\right)^2 + \sum\left(x_{i.} - \overline{x_{..}}\right)^2$$
$$SC_T \quad = \quad SC_I \quad + \quad SC_E$$

Conviene recalcar que todos los elementos de esta expresión entran en los sumatorios una vez por cada observación, por lo que cada suma de cuadrados externa (SC_E) entra en esta ecuación J veces, de manera que lo que se obtiene es JV_E.

Como se dijo al hablar del cálculo de la varianza, las sumas de cuadrados pueden llevarse a cabo sin necesidad de calcular las distintas medias y cada una de las desviaciones, de manera que, a partir de esta expresión, y siguiendo una sencilla regla mnemotécnica se llega a las expresiones que aparecen en la tabla. La regla consiste en a) eliminar las barras que aparecen en la expresión, convirtiendo todas las medias en sumas, b) introducir un sumatorio dentro de los paréntesis excepto cuando no tiene sentido, como es el caso de la suma global, y c) dividir cada expresión por el subíndice en mayúsculas que falta en el numerador al haber sido sustituido por un punto. Estas expresiones así convertidas son trasladadas a los lugares correspondientes en las tabla ANOVA. Así, las sumas de cuadrados externa (SC_E) e interna (SC_I) se podrían obtener:

$$SC_E = \sum\left(\overline{x_{i.}} - \overline{x_{..}}\right)^2 = \sum \frac{x_{i.}^2}{J} - \frac{x_{..}^2}{IJ}$$

$$SC_I = \sum\left(x_{ij} - \overline{x_{i.}}\right)^2 = \sum x_{ij}^2 - \sum \frac{x_{i.}^2}{J}$$

Los elementos de la columna etiquetada como cuadrados medios se obtienen sencillamente mediante el cociente entre las sumas de cuadrados y los grados de libertad correspondientes.

Finalmente, la última columna representa el valor esperado de los cuadrados medios calculados, y cuya explicación ha sido desarrollada en el apartado 7.2. Así, por ejemplo, $\widehat{\sigma^2_{\bar{x}}} = V_E = \frac{\sum\left(\overline{x_{i.}} - \overline{x_{..}}\right)^2}{I-1} = \frac{\widehat{\sigma^2}}{J} + \widehat{\sigma^2_G}$, o lo que es lo mismo, hablando en términos de esperanza o valor esperado de los cuadrados medios, y multiplicando por *J* de acuerdo con lo expuesto más arriba, $E\left(CM_E\right) = E\left(JV_E\right) = JE\left(V_E\right) = J\sigma^2_{\bar{x}} = \sigma^2 + J\sigma^2_G$, y con respecto a la varianza interna, $E\left(CM_I\right) = E\left(V_I\right) = \sigma^2$. Con todo lo anterior se construye la tabla ANOVA que será la siguiente:

Fuentes de varianza	Grados de libertad gl	Suma de cuadrados SC	Cuadrados medios CM	Esperanza de los cuadrados medios E(CM)
EXTERNA	*I*-1	$\frac{\sum x_{i.}^{\ 2}}{J} - \frac{x_{..}^{\ 2}}{IJ}$	$\frac{SC_E}{I-1}$	$\sigma^2 + J\sigma^2_G$
INTERNA	*I(J*-1*)*	$\sum x_{ij}^{\ 2} - \frac{\sum x_{i.}^{\ 2}}{J}$	$\frac{SC_I}{I\left(J-1\right)}$	σ^2
TOTAL	*IJ*-1	$\sum x_{ij}^{\ 2} - \frac{x_{..}^{\ 2}}{JK}$	$\frac{SC_T}{IJ-1}$	

I-7.6.1. Inferencias obtenidas por ANOVA cuando se pretenden comparar series de medias

En este caso el efecto clasificador, el efecto tratamiento o grupo, el causante de varianza externa se está considerando como efecto fijo, y se pretende concluir si la varianza aportada por este efecto,

σ_G^2, se puede considerar despreciable o no. Una inspección de la columna correspondiente a las E(CM) nos permite comprobar que el valor del cuadrado medio externo corresponde exactamente a lo mismo que el cuadrado medio interno más una componente que debería ser nula en el caso de no existir influencia del efecto que clasifica las observaciones por grupos, de manera que éste es exactamente el cociente que llamaremos F_c y cuyo valor debería ser superior al de las tablas F para poder rechazar la H_0, teniendo en cuenta los grados de libertad del numerador, denominador y error alfa que estemos dispuestos a asumir:

$$F_c = \frac{CM_E}{CM_I} \sim F_{IJ-I,\alpha}^{I-1}$$

- Hipótesis nula. Conviene recordar que el contraste de hipótesis se inicia dando por buena la hipótesis de partida o hipótesis nula, de manera que si $F_{IJ-I,\alpha}^{I-1} > F_c = \frac{CM_E}{CM_I}$, entonces el valor de σ_G^2 no es tan grande como para ser considerada diferente de cero, con lo que refrendamos nuestra hipótesis nula, concluimos que no hay diferencias entre las medias de los grupos no atribuibles al azar, y hemos finalizado el análisis.

- Hipótesis alternativa. Sin embargo, podría darse el caso en que $F_{IJ-I,\alpha}^{I-1} \leq F_c = \frac{CM_E}{CM_I}$. Entonces, el numerador es significativamente diferente del denominador y nos vemos obligados a rechazar la hipótesis nula, por lo que deberemos aceptar la hipótesis alternativa, según la cuál, al menos una de las medias es diferente de las demás. En este punto suele darse un segundo paso para discernir qué medias deben ser consideradas diferentes entre sí y cuáles no.

I-7.6.2. Inferencias obtenidas por ANOVA cuando se utiliza para descomponer la varianza total en una serie de componentes

Cuando el efecto es considerado aleatorio resulta útil descomponer la varianza de la variable en una componente debida al efecto en cuestión y el resto. Es importante hacer notar que en el caso del apartado 7.6.1 anterior, la varianza de la variable se asume que es la varianza interna o varianza residual, siendo el efecto clasificador causante de una varianza mayor entre los datos que es originada por el efecto sistemático que proporciona cada nivel del factor. Sin embargo, en este apartado la varianza de la variable es la varianza global, aunque la partición de esta varianza global sea discernible.

Los estimadores de las distintas componentes se obtienen en este caso simplemente igualando cada cuadrado medio a su esperanza, obteniendo así los estimadores de los distintos componentes. Así, después de despejar obtenemos los valores de cada componente y la varianza fenotípica total σ_P^2.

$$\widehat{\sigma^2} = CM_I$$

$$\widehat{\sigma_G^2} = \frac{CM_E - CM_I}{J}$$

$$\widehat{\sigma_P^2} = \widehat{\sigma_G^2} + \widehat{\sigma^2}$$

I-7.7. Tabla de un ANOVA jerárquico doble

Siguiendo la misma lógica que para el ANOVA jerárquico simple, y asumiendo que el lector podría deducir todo lo que sigue en este apartado, se exponen a continuación las particularidades de un ANOVA jerárquico doble en el que se consideran *K* observaciones dentro de cada uno de los *J* subgrupos en que se encuentra subdividido cada uno de los *I* grupos.

Descomposición de una desviación del dato respecto de la media global:

$$x_{ijk} - \overline{x_{...}} = \left(x_{ijk} - \overline{x_{ij.}}\right) + \left(\overline{x_{ij.}} - \overline{x_{i..}}\right) + \left(\overline{x_{i..}} - \overline{x_{...}}\right)$$

Descomposición de la suma de cuadrados total:

$$\sum\left(x_{ijk} - \overline{x_{...}}\right)^2 = \sum\left(x_{ijk} - \overline{x_{ij.}}\right)^2 + \sum\left(\overline{x_{ij.}} - \overline{x_{i..}}\right)^2 + \sum\left(\overline{x_{i..}} - \overline{x_{..}}\right)^2$$
$$SC_T \quad = \quad SC_I \quad + \quad SC_S \quad + \quad SC_G$$

Tabla ANOVA:

Fuentes de varianza	Grados de libertad gl	Suma de cuadrados SC	Cuadrados medios CM	Esperanza de los cuadrados medios E(CM)
GRUPOS	*I*-1	$\frac{\sum x_{i..}^{\ 2}}{JK} - \frac{x_{...}^{\ 2}}{IJK}$	$\frac{SC_G}{I-1}$	$\sigma^2 + J\sigma_S^2 + JK\sigma_G^2$
SUBGRUPOS	*I(J-1)*	$\frac{\sum x_{ij.}^{\ 2}}{K} - \frac{\sum x_{i..}^{\ 2}}{JK}$	$\frac{SC_S}{I(J-1)}$	$\sigma^2 + J\sigma_S^2$
INTERNA	*IJ(K-1)*	$\sum x_{ijk}^{\ 2} - \frac{\sum x_{ij.}^{\ 2}}{K}$	$\frac{SC_I}{IJ(K-1)}$	σ^2
TOTAL	*IJK*-1	$\sum x_{ijk}^{\ 2} - \frac{x_{...}^{\ 2}}{IJK}$	$\frac{SC_T}{IJK-1}$	

Y consecuentemente, hay dos posibles contrastes de hipótesis si el efecto se considera fijo, y otra componente de varianza si se considera aleatorio.

I-7.8. Coeficiente de correlación intraclase

Mide el grado de parecido de las observaciones dentro de los grupos en relación al total de observaciones. Nótese que si los grupos se realizan de acuerdo a familias (hijos del mismo padre, de la misma madre o de los mismos padre y madre), el parecido de las familias será un parecido genético, por lo que este coeficiente es muy importante en el contexto de la genética cuantitativa para la estimación de un parámetro tan importante como es la heredabilidad.

En un análisis de varianza jerárquico simple, en el que la varianza total se descompone de esta manera: $\sigma^2_{Total} = \sigma_G{}^2 + \sigma^2$, el coeficiente de correlación intraclase *t* se calcula:

$$t = \frac{\sigma_G{}^2}{\sigma^2_{Total}} = \frac{\sigma_G{}^2}{\sigma_G{}^2 + \sigma^2}$$

Obsérvese que cuando todos los individuos dentro de los grupos son iguales, la varianza interna es cero ($\sigma^2 = 0$), toda la varianza es debida a grupos, $\sigma^2_{Total} = \sigma^2_G$, y el coeficiente de correlación intraclase alcanza su valor máximo, $t = 1$

Si el modelo es jerárquico doble, estaremos en disposición de obtener dos coeficientes de correlación intraclase, el coeficiente de correlación intra-grupos (t_1) y el coeficiente de correlación intra-subgrupos (t_2).

$$t_1 = \frac{\sigma_G{}^2}{\sigma_G{}^2 + \sigma_{SG}{}^2 + \sigma^2} \qquad \text{y} \qquad t_2 = \frac{\sigma_G{}^2 + \sigma_{SG}{}^2}{\sigma_G{}^2 + \sigma_{SG}{}^2 + \sigma^2}$$

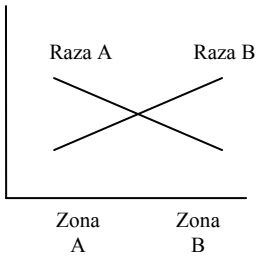
El primero mide el grado de parecido de las observaciones dentro de los grupos (si $\sigma^2 = 0$ y $\sigma^2_{SG} = 0$, todas son iguales y $t_1 = 1$), y el segundo mide el grado de parecido de las observaciones dentro de

los subgrupos (si $\sigma^2 = 0$ aunque pueda darse que $\sigma^2_{SG} \neq 0$, todas las observaciones dentro de los subgrupos son iguales y $t_2 = 1$).

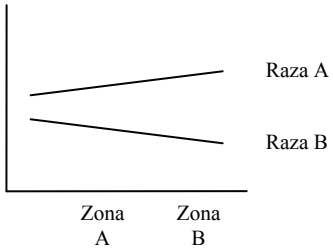
I-7.9. Interacción entre efectos

En ocasiones dos efectos del mismo nivel de jerarquía interaccionan de manera que el efecto que producen los niveles de uno de ellos sobre la variable no es el mismo en todos los niveles del otro efecto. Se dice entonces que existe una interacción entre efectos.

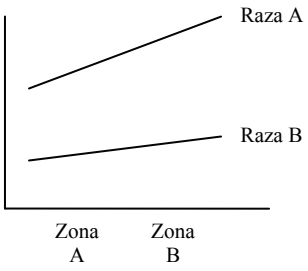
Imaginemos por ejemplo dos razas autóctonas perfectamente adaptadas al ambiente en el que registran su producción. Un investigador decide comprobar cuál de las dos razas produce más y en cuál de los dos ambientes se produce más, de manera que estudia el rendimiento medio de cada una de las razas en cada uno de los ambientes. La representación de los rendimientos en función de las razas y ambientes se observa en el siguiente gráfico:



Cada una de las razas está adaptada a su ambiente, de manera que no hay una raza mejor, sino que dependiendo de la zona en la que se encuentren la mejor será una de ellas. Del mismo modo, tampoco hay una zona preferible para una mayor producción, sino que depende de la raza de la que dispongamos. Este efecto de interacción es muy extremo pero existen otros tipos de interacciones más difíciles de advertir. Por ejemplo, puede existir realmente una raza mejor en todos los ambientes, pero, mientras que para una de ellas una zona es mejor, para la otra raza es otra zona la que es mejor:



Pero incluso existe también un tipo de interacción en el que existe una raza mejor que la otra, y también para ambas hay una zona que es mejor. Aún así, la influencia de uno de los efectos sobre el otro depende del nivel de uno de ellos. En el siguiente ejemplo la raza A es siempre mejor y la zona B también es siempre mejor, pero está claro que el pasar de la zona A a la B favorece más a la raza A que a la B:



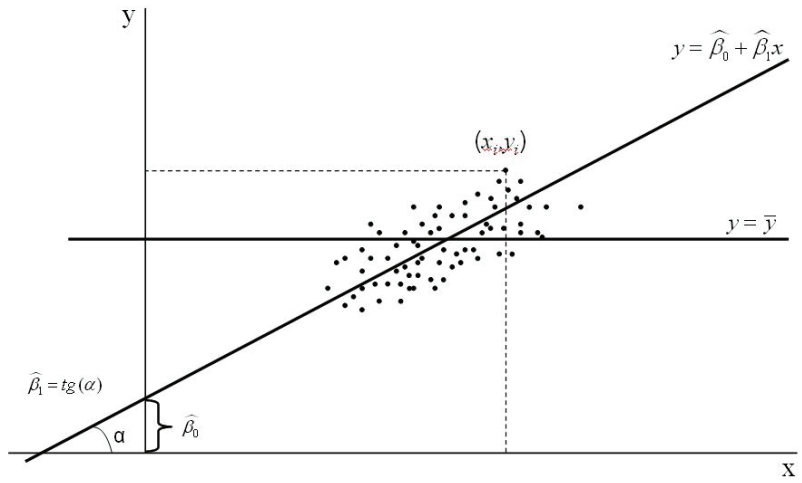
I-7.9.1. Modelos con interacción entre efectos

Aunque no se va a profundizar aquí, conviene resaltar que los modelos con interacciones deben ajustar al mismo nivel de jerarquía los dos efectos entre los que hay interacción. Además, para poder separar la interacción de la componente de varianza interna, en cada combinación de niveles de los dos efectos debe existir más de una observación.

I-8. Regresión

El análisis de regresión es utilizado para medir la relación de dos variables, pero, a diferencia del coeficiente de correlación que mide la variación conjunta, la regresión asume la posibilidad de predecir una un función de la otra. Así, mientras que el coeficiente de correlación no establece diferencias entre las dos variables, en el coeficiente de regresión, una de las variables se asume medida sin error y se conoce como variable independiente o regresora, mientras que la otra variable se conoce como variable dependiente. Para su cálculo se precisan también parejas de datos para cada una de ellas.

Si se representan los valores de cada pareja de datos en un eje de coordenadas se obtendrá una nube de puntos que aparecen en la figura:



Obtener una recta de regresión consistiría entonces en obtener los parámetros b_0 y b_1 de una recta de regresión que se podrían denominar también $\hat{\beta}_0$ y $\hat{\beta}_1$, al asumirse que existe una relación de dependencia entre las variables x e y de manera que se podría obtener y_i a partir de x_i mediante la expresión:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

En esta ecuación ε_i es el residuo del modelo, originado por la variabilidad residual o interna, o en otras palabras, por el hecho de que no para todas las parejas de datos con un mismo valor x_i, se corresponda un mismo valor y_i.

NOTACIÓN:

Obsérvese que el modelo de regresión asume un verdadero valor poblacional desconocido de los parámetros β_0 y β_1 que definen la recta de regresión $y = \beta_0 + \beta_1 x$, razón por la que se usan letras griegas en la notación. Sin embargo, la recta que construiremos tendrá unos parámetros conocidos obtenidos a partir de nuestra muestra de datos, por lo que su representación no se llevará a cabo con letras griegas: $y = b_0 + b_1 x$. Sin embargo, cuando se pretenda extrapolar la ecuación obtenida como estimador de la verdadera, entonces la ecuación se escribirá $y = \widehat{\beta_0} + \widehat{\beta_1} x$. Por otro lado, y del mismo modo, se asume que existe un verdadero valor del residuo (ε_i) que en caso de conocerse permitiría predecir el valor de cualquier observación de la variable dependiente al conocerse también el valor de la variable independiente utilizando la educación $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Así, por considerarse verdadero el valor del residuo de este modelo, debe escribirse también con letras griegas.

El coeficiente β_1 se conoce como coeficiente de regresión de y sobre x, y tiene varias interpretaciones interesantes:

- A la vista de la ecuación de regresión $y = \beta_0 + \beta_1 x$ se observa que β_1 representa el incremento que se da en la variable y cuando x aumenta una unidad. Por ejemplo, si la variable dependiente fuera el sueldo de una persona y la variable independiente el año, representaría la cantidad que aumenta el sueldo de una persona por cada año que pasa.

- Representa también la pendiente de la recta de regresión. Al ser el cociente entre el incremento de y y el incremento de x,

($\beta_1 = \frac{\Delta y}{\Delta x}$) también es la tangente del ángulo α de la figura, es decir, el ángulo que forma la recta de regresión con el eje *x* o eje de abscisas. Así, en caso de no existir una dependencia lineal de *y* sobre *x*, el valor de *y* sería constante para cualquier valor de *x*, y por tanto, la recta de regresión sería paralela al eje de abscisas, α sería igual a cero y también lo sería su tangente.

El coeficiente β_0 se conoce como ordenada en el origen, ya que representa el valor que adquiere la variable dependiente *y* en el eje de ordenadas, es decir, cuando la variable independiente *x* vale cero. En algunos textos aparece con el nombre de intercepto.

I-8.1. Estimación del coeficiente de regresión y de la ordenada en el origen.

Entre todos los posibles métodos de estimación utilizaremos el método de ajuste por mínimos cuadrados. Su fundamento es el siguiente. Según la ecuación $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, los valores obtenidos por la recta de regresión ($y = \beta_0 + \beta_1 x_i$) para la variable *y* a partir de cada valor de la variable *x*, se desviarán del verdadero valor en el error o residuo ε. Por tanto, un buen ajuste será aquél en el que todos estos errores tengan un valor absoluto muy pequeño, o lo que es lo mismo que la suma de los cuadrados de los errores sea lo más pequeña posible.

Así pues, es preciso encontrar una expresión que explique la suma de cuadrados de los residuos y que dependa de los parámetros de la recta. A continuación hay que encontrar los valores de estos parámetros que hagan mínima la expresión, lo que se logra igualando a cero la primera derivada de la expresión con respecto a cada uno de los parámetros.

DERIVACIÓN DE LOS COEFICIENTES DE LA RECTA DE REGRESIÓN

1) Expresión de la suma de cuadrados de los errores en función de los parámetros de la recta. De la ecuación de la recta

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ se puede despejar el residuo,

$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$, y sobre esta expresión se puede obtener la función que expresa la suma de cuadrados de los residuos:. Obsérvese que el cuadrado del polinomio es la suma de los cuadrados de cada elemento más el doble producto de cada dos, y que sumar *n* veces un mismo valor lleva al producto de *n* por ese valor.

$$\sum (\varepsilon_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2 =$$
$$= \sum y_i^2 + n\beta_0^2 + \beta_1^2 \sum x_i^2 - 2\beta_0 \sum y_i - 2\beta_1 \sum x_i y_i + 2\beta_0 \beta_1 \sum x_i$$

2) Derivada de la expresión con respecto a cada uno de los parámetros a estimar. Obsérvese que en la expresión hay muchos elementos que son independientes del parámetro para el que se deriva, por lo que son constantes al efecto y su derivada vale cero:

$$\frac{\delta\left[\sum (\varepsilon_i)^2\right]}{\delta\beta_0} = 2n\beta_0 + -2\sum y_i + 2\beta_1 \sum x_i$$

$$\frac{\delta\left[\sum (\varepsilon_i)^2\right]}{\delta\beta_1} = 2\beta_1 \sum x_i^2 - 2\sum x_i y_i + 2\beta_0 \sum x_i$$

3) Se iguala la primera derivada a cero y se despejan los valores de los estimadores de β_0 y β_1. Obsérvese que en este punto los parámetros son ya estimadores y por tanto deben escribirse con la notación apropiada, por lo que llevan ya sombrero:

- Derivación de la ordenada en el origen:

$$2n\widehat{\beta_0} + -2\sum y_i + 2\widehat{\beta_1} \sum x_i = 0$$

$$n\widehat{\beta_0} = \sum y_i - \widehat{\beta_1} \sum x_i \qquad \Rightarrow$$

$$\widehat{\beta_0} = \frac{\sum y_i}{n} - \widehat{\beta_1} \frac{\sum x_i}{n} = \overline{y} - \widehat{\beta_1}\overline{x}$$

Así pues, la ordenada en el origen depende del coeficiente de regresión.

- Derivación del coeficiente de regresión:

$$2\widehat{\beta}_1 \sum x_i^2 - 2\sum x_i y_i + 2\widehat{\beta}_0 \sum x_i = 0$$

$$\widehat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \left(\frac{\sum y_i}{n} - \widehat{\beta}_1 \frac{\sum x_i}{n} \right) \sum x_i =$$

$$= \sum x_i y_i - \left(\frac{\sum x_i \sum y_i}{n} - \widehat{\beta}_1 \frac{\left(\sum x_i\right)^2}{n} \right)$$

$$\widehat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} + \widehat{\beta}_1 \frac{\left(\sum x_i\right)^2}{n}$$

$$\widehat{\beta}_1 \sum x_i^2 - \widehat{\beta}_1 \frac{\left(\sum x_i\right)^2}{n} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$\widehat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{CoV(x,y)}{V_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

En la última línea se ha incluido la covarianza entre las variables en el numerador, así como la varianza de la variable regresora en el denominador, ya que tanto en numerador como denominador se había obtenido las correspondientes sumas de cuadrados. Obsérvese también la doble notación muestral o poblacional que utilizaremos según nuestro interés se centre en obtener los valores para una muestra concreta o se pretenda extrapolar los resultados a la población.

Por tanto, los coeficientes de regresión y ordenada en el origen se pueden obtener directamente de los datos:

$$\widehat{\beta}_1 = b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{CoV(x,y)}{V_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\widehat{\beta}_0 = b_0 = \overline{y} - b_1 \overline{x}$$

I-8.2. Análisis de regresión

Como en todas las pruebas estadísticas, la estimación de un parámetro no determina su valor con exactitud, sino que podría no considerarse diferente de cero.

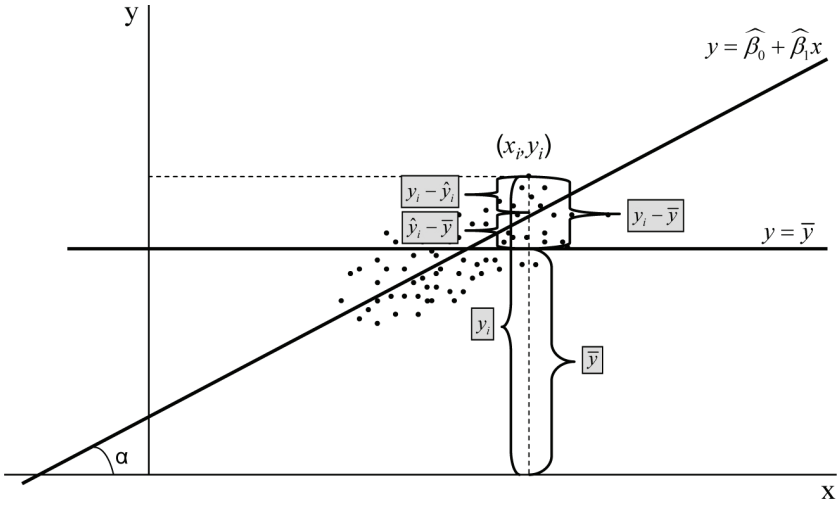
Nuevamente se plantea un contraste de hipótesis, con el establecimiento de la hipótesis nula (H_0) , que como en otras ocasiones establece el conocimiento que se tiene hasta el momento sobre el parámetro: "el coeficiente de regresión no es diferente de cero". Y también como siempre, se establece una hipótesis alternativa (H_A) que deberá aceptarse en caso de no poder sostenerse la nula después del análisis, y que indica lo contrario de ella, es decir, "el coeficiente de regresión es diferente de cero".

Necesitamos ahora determinar si la variabilidad explicada por la recta de regresión es suficientemente importante en relación a la variabilidad total de la variable dependiente como para que sea considerada diferente de cero.

Descomponemos entonces cada desviación de un dato de la variable dependiente con respecto a su media en otras dos:

$$y_i - \overline{y} = \left(y_i - \widehat{y_i} \right) + \left(\widehat{y_i} - \overline{y} \right)$$

Una explicación gráfica de esta descomposición se puede obtener de la siguiente gráfica:



En la expresión anterior $\widehat{y}_i$ representa el punto proporcionado por la recta para cada x_i. Si este punto se aleja mucho de la media en relación a lo que se aleja el verdadero valor del dato de $\widehat{y}_i$, entonces la regresión tiene mucha importancia relativa. En la gráfica se observa cómo, si tiene sentido la existencia de una recta de regresión que relacione las dos variables, la parte de la desviación del dato con respecto a la media que corresponde a la recta de regresión, es importante en relación al total de la desviación.

Se toman sumatorios de las desviaciones al cuadrado y se establecen así las distintas sumas de cuadrados TOTAL, debida a REGRESIÓN, y RESIDUAL:

$$\sum (y_i - \bar{y})^2 = \sum \left(y_i - \widehat{y}_i\right)^2 + \sum \left(\widehat{y}_i - \bar{y}\right)^2$$

$$SC_T = SC_{Regresión} + SC_{Residual}$$

$$SC_T = SC_{Regresión} + SC_{Residual}$$

No se detallará aquí el desarrollo de las expresiones que conducen a la tabla del análisis de regresión. Por el contrario, se expone a continuación la tabla correspondiente al análisis de regresión, a partir de la cual se deberá proceder como en aquel caso:

Fuentes de variación	Grados de libertad gl	Suma de cuadrados SC	Cuadrados medios CM	Esperanza de los cuadrados medios E(CM)
Regresión	1	$\widehat{\beta^2}\sum\left(x_i-\overline{x}\right)^2$	$SC_{Regresión}$	$\sigma^2+\beta_1^2\sum\left(x_i-\overline{x}\right)^2$
Residual	*n-2*	$\sum\left(y_i-\overline{y}\right)^2-$ $-\widehat{\beta^2}\sum\left(x_i-\overline{x}\right)^2$	$\frac{SC_{Residual}}{n-2}$	σ^2
TOTAL	*n*-1	$\sum\left(y_i-\overline{y}\right)^2$		

A la vista de los valores esperados de los cuadrados medios se observa que el correspondiente a regresión contiene lo mismo que el debido al residuo más una parte que depende del coeficiente de regresión. Así pues, el cociente entre ambos cuadrados medios debe compararse con el valor de las tablas F para decidir si se acepta o se rechaza la hipótesis nula, o, en otras palabras, si el coeficiente de regresión debe considerarse respectivamente diferente de cero o no.

El análisis de regresión no debería darse por concluido aquí ya que es frecuente que la regresión no sea diferente de cero pero el ajuste lineal no sea el más apropiado. Tales pruebas tampoco van a ser detalladas aquí.

CONCEPTOS CLAVE

- ¿Qué diferencias hay entre población y muestra? ¿y entre parámetros muestrales y poblacionales?
- Interpretación del concepto de Esperanza Matemática
- Relación entre las diferentes medidas de dispersión: varianza, desviación típica, error típico y coeficiente de variación.
- Interpretación de los posibles valores del coeficiente de correlación. ¿Cuál es su rango? ¿Qué significado tiene un coeficiente de correlación nulo? ¿Y uno negativo?
- La media muestral como variable aleatoria. ¿Cuánto vale su varianza? ¿Cómo se interpreta su la varianza de la media muestral como medida de precisión de las estimación de la media poblacional?
- ¿Para qué sirve una distribución normal de media cero y varianza uno? ¿Cómo se tipifica una variable? ¿Cómo se puede expresar una variable en una media y varianza arbitrarias?
- Propiedades de los estimadores. Sesgo y varianza
- ¿Cuál es el fundamento del Análisis de Varianza? ¿Si los datos dentro de los grupos son muy parecidos se apreciarán diferencias entre los grupos?
- ¿Qué miden los coeficientes de correlación intraclase?
- ¿Qué es la interacción entre efectos?
- ¿En qué se parecen y en qué se diferencian los coeficientes de regresión y de correlación?

SEGUNDA PARTE

ALGUNOS CONCEPTOS GENERALES DE GENÉTICA CUANTITATIVA

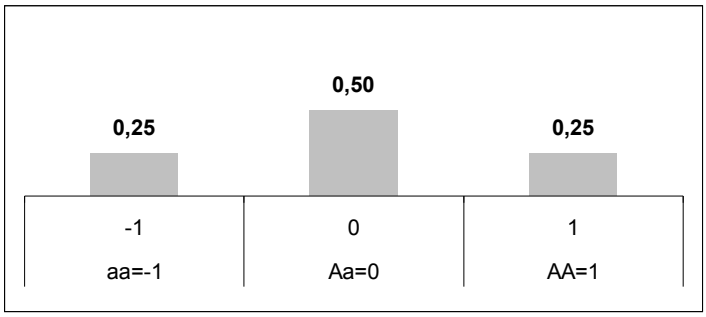
RESUMEN

Se desarrollan con el mínimo detalle los conceptos clásicos de la genética cuantitativa que resultan necesarios en las valoraciones genéticas. Se justifica la utilización de la distribución normal en mejora genética, la partición del fenotipo en genotipo y ambiente, así como la varianza fenotípica en componentes similares. Se define un parámetro imprescindible como la heredabilidad, porcentaje de la variabilidad de los datos que se debe a variabilidad genética, así como otros parámetros como el ambiente permanente y la repetibilidad. Finalmente se describen someramente aquellos factores de los que depende la respuesta a la selección, tanto cuando se realiza directamente a partir de la información que proporciona el carácter que se desea mejorar, como la respuesta correlacionada que se obtiene al utilizar la información de otro carácter con el que tiene relación.

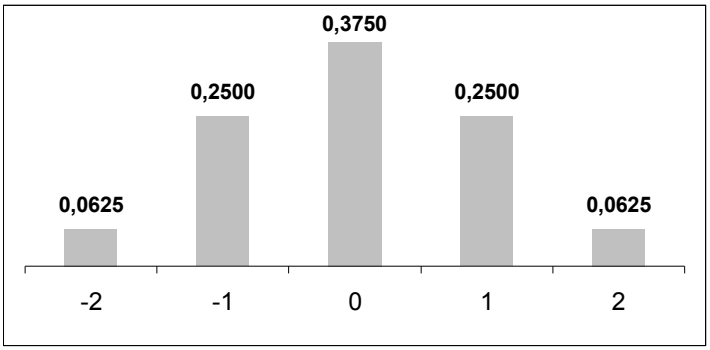
II-1. Partición del fenotipo
II-2. Partición del genotipo
II-3. Partición de la varianza fenotípica
II-4. Heredabilidad
II-5. Ambiente permanente y Repetibilidad
II-6. Selección
    II-6.1. Respuesta a la selección
        II-6.1.1. Selección indirecta y respuesta correlacionada

Varias son las diferencias que existen entre la genética cuantitativa y la genética cualitativa. Mientras que en ésta los caracteres se asumen determinados por uno o unos pocos genes de gran efecto, en la genética cuantitativa los caracteres se asumen determinados por infinitos genes de efecto infinitesimal cada uno de ellos. Así, mientras que en la genética cualitativa los fenotipos son discernibles, en la genética cuantitativa existe un espectro continuo que impide distinguir las diferentes clases.

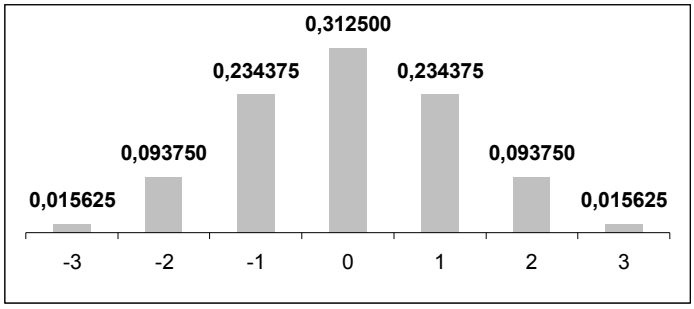
Un carácter determinado por un solo *locus*, con sólo dos alelos posibles, *A* y *a*, y representación equilibrada, presenta una frecuencia de cada uno de sus alelos igual a 0,5. Si la reproducción es al azar se dice que la población se encuentra en equilibrio Hardy-Weinberg, y en este caso, la frecuencia de los tres genotipos posibles está definida directamente mediante el producto de las frecuencias, siendo 0,25 la frecuencia de los individuos con genotipo *AA* ($0,5^2 = 0,25$), 0,25 la frecuencia de los individuos con genotipo *aa* ($0,5^2 = 0,25$), y 0,50 la frecuencia de los individuos heterocigotos con genotipo *Aa* ($2 \times 0,5^2 = 0,5$), en este caso multiplicada por 2 porque deben contarse el caso en el que el alelo dominante lo proporciona uno de los padres y el caso en el que lo proporciona el otro. Si la posesión de un alelo *A* supone un incremento de una unidad del carácter con respecto a la media, mientras que un alelo *a* supone una unidad menos que la media, se puede representar las frecuencias con que aparecen los individuos de los distintos valores de la siguiente manera:



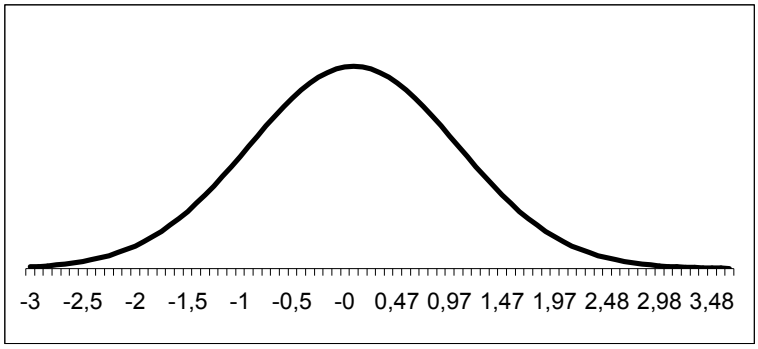
Si en lugar de un solo *locus* se asume la existencia de dos *loci* en las mismas condiciones, entonces los posibles genotipos son cinco, el rango va de -2 a +2 y la distribución es la siguiente:



Se puede extender la explicación para tres *loci*, con 7 clases posibles y rango de -3 a +3, pero los valores extremos empiezan a ser ya muy poco representativos.



Así que con un pequeño puñado de genes, el número de genotipos posibles se va incrementando, y no es descabellado asumir un modelo infinitesimal, en el que el espectro de valores es continuo y debemos estudiar los caracteres mediante una función de densidad:



Esta distribución de datos se ajusta perfectamente a la distribución normal, por lo que en genética cuantitativa no trataremos con proporciones sino con parámetros estadísticos como medias y varianzas. Obsérvese que esta distribución se ha construido partiendo como base de genes con únicamente dos alelos y herencia mendeliana simple, de manera que este tipo de herencia no contradice la herencia de los caracteres de clase, sino que trasciende a la estadística por asumir que los caracteres se determinan a partir de un elevado número de genes.

Los genes individuales que forman parte de este modelo se llaman también poligenes y en este modelo pueden presentar todo tipo de influencias de unos sobre otros como modificación, refuerzo, disminución o supresión de la expresión determinada por otro, presentando igualmente los valores genéticos globales una distribución normal. Se habla así del modelo infinitesimal según el cuál el fenotipo, y también el valor gético aditivo, son el resultado de la suma del efecto de infinitos alelos con efecto infinitesimal cada uno de ellos.

A continuación se va a describir la partición del fenotipo y de su varianza para desembocar en el parámetro más importante en genética cuantitativa, la heredabilidad. Algunos conceptos sobre selección artificial son también de utilidad para contextualizar apropiadamente la valoración genética animal.

II-1. Partición del fenotipo

Llamamos fenotipo a la expresión de un carácter determinado, al menos en parte, genéticamente. El resto del carácter es la consecuencia de un cúmulo de influencias no genéticas que agrupamos en el término ambiente. Se define así la primera base de la herencia en genética cuantitativa:

Fenotipo = Genotipo + Ambiente

$$\mathbf{P = G + E}$$

Así por lo tanto, el genotipo no se expresará en su totalidad sino que se verá modificado por el ambiente. Por ejemplo, una vaca puede tener una enorme capacidad genética para producir leche, pero si no se la alimenta suficientemente, no la producirá. Sin embargo, es posible que un animal adaptado a un ambiente no rinda adecuadamente en un segundo ambiente contrapuesto, mientras que un animal adaptado a este segundo ambiente no rendiría convenientemente en el primero. Este ejemplo no es sino un caso de interacción entre genotipo y ambiente que debería incluirse en la ecuación anterior:

$$\mathbf{P = G + E + GxE}$$

La interacción Genotipo x Ambiente es muy difícil de tener en cuenta en los modelos de valoración genética por lo que suele incorporarse a la parte no genética del fenotipo definida inicialmente como ambiente. El nuevo componente ambiental es el que existía más la propia interacción: $\mathbf{E} = \mathbf{E} + \mathbf{GxE}$.

II-2. Partición del genotipo

Como se vio en la introducción de este capítulo, en el marco de la genética cuantitativa se asume que el genotipo de un individuo es consecuencia de multitud de genes que participan del genotipo de distintas formas:

- Valor genético Aditivo *A*: Es la consecuencia de la suma del efecto de todos los alelos presentes en todos los *loci* que participan en el carácter.
- Interacción de la Dominancia *D*. Se debe a la influencia en el mismo *locus* que un alelo ejerce sobre el otro.
- Interacción Epistática *I*. Es originada por la influencia que un alelo o un genotipo de un *locus* ejerce sobre otro *locus* diferente.

Con esta descomposición, el fenotipo queda expresado de la siguiente manera:

$$\mathbf{P = A + D + I + E}$$

II-3. Partición de la varianza fenotípica

Como fue comentado en la introducción de este capítulo, la varianza es el material de trabajo de las valoraciones genéticas. El hecho de que haya individuos que produzcan más que otros nos proporciona una oportunidad de mejora, más aún si esta variabilidad es de origen genético. Así pues, la varianza de los fenotipos, conocida como varianza fenotípica (σ_p^2) puede descomponerse en varianzas de diferente origen. Si asumimos que las covarianzas entre los distintos elementos de la expresión anterior son nulos, la varianza fenotípica puede descomponerse de la siguiente manera:

$$\sigma_p^2 = \sigma_G^2 + \sigma_e^2 = \sigma_u^2 + \sigma_D^2 + \sigma_I^2 + \sigma_e^2$$

En esta expresión $\sigma_G^2, \sigma_e^2, \sigma_u^2, \sigma_D^2$ y σ_I^2 son, respectivamente, las varianzas genotípica, residual, genética aditiva, dominante y epistática. En ella se han escrito en minúsculas los subíndices de las componentes de varianza fenotípica, genética aditiva y residual, por destacar las que serán trabajadas más en profundidad durante los temas relacionados con la valoración genética.

II-4. Heredabilidad

Se ha comentado que la variabilidad del carácter es imprescindible para llevar a cabo la selección, y se ha mostrado que esta variabilidad de los datos o variabilidad fenotípica presenta varios orígenes, siendo únicamente de interés la que tiene base genética. El conocimiento de la proporción de la variabilidad que es de origen genético es entonces un parámetro de mucho interés y tiene dos posibles acepciones. Se definen así los siguientes parámetros:

- Heredabilidad en sentido amplio (H^2). Es el porcentaje de la variabilidad fenotípica que es de origen genético:

$$H^2 = \frac{\sigma_u^2 + \sigma_D^2 + \sigma_I^2}{\sigma_p^2}$$

- Heredabilidad en sentido estricto (h^2). Es el porcentaje de la variabilidad fenotípica que es de origen genético aditivo:

$$h^2 = \frac{\sigma_u^2}{\sigma_p^2}$$

Obsérvese que sólo la parte aditiva se hereda de forma directa dado que las otras componentes son interacciones y precisarían combinaciones específicas de los genes procedentes del mismo individuo (epistasia), o de de ambos padres (epistasia y

dominancia). Por ello, el concepto realmente utilizado es el de heredabilidad en sentido estricto o sencillamente heredabilidad.

De su definición se desprende que los valores de la heredabilidad se encuentran en el rango de cero a uno. En los extremos, si la heredabilidad vale uno, toda la variabilidad observada sería de origen genético de forma que el rendimiento de los individuos sería una medida directa de su valor genético. En esta situación la valoración genética no tendría sentido ya que bastaría escoger como padres de la siguiente generación a los animales con mejores rendimientos. En el otro extremo, si la heredabilidad vale cero, los valores genéticos de los animales son todos idénticos, las diferencias observadas se deberían al ambiente y en este no tendría sentido la selección. En la medida en que la heredabilidad es más baja será necesaria una mayor cantidad de información de parientes para valorar genéticamente al animal; en otras palabras, un buen o mal rendimiento del individuo no dirá mucho de su valor genético a menos que se repita sistemáticamente este rendimiento en sus parientes.

Aunque la heredabilidad no es específica de especies ni de caracteres ni de poblaciones, existen valores comunes de heredabilidad en función del tipo de carácter:

- Heredabilidad alta (mayor de 0,40). Caracteres relacionados con el tamaño como por ejemplo la alzada a la cruz.

- Heredabilidad moderada (de 0,15 a 0,40). Son los valores de heredabilidad más comunes como por ejemplo, la heredabilidad de la producción de leche.

- Heredabilidad baja (menor de 0,15). Caracteres relacionados con la esfera reproductiva como por ejemplo la prolificidad.

Aunque no es parte del contenido de este texto la estimación de parámetros genéticos como la heredabilidad, ésta debe llevarse a cabo en cada población como paso previo a la valoración genética.

Dicha estimación ha sido llevada a cabo tradicionalmente a partir de métodos estadísticos aplicados sobre diseños preparados *ad hoc*, como el análisis de varianza en el que los grupos los forman hermanos o medios hermanos, la regresión de la producción de las hijas sobre la de sus madres, o la correlación entre los rendimientos de hermanos.

II-5. Ambiente permanente y Repetibilidad

Se define el ambiente permanente como la parte no genética de un individuo que se repite en todos sus registros. Así, un mismo individuo aportará a todos sus datos dos efectos, su valor genético y su ambiente permanente.

Dando un paso más allá, interesa conocer el porcentaje de la variabilidad que se observa que es de origen genético más ambiental permanente. Se define así la repetibilidad y se calcula como:

$$R = \frac{\sigma_u^2 + \sigma_{ep}^2}{\sigma_p^2}$$

En esta expresión σ_{ep}^2 es la varianza ambiental permanente. La repetibilidad toma valores entre cero y uno y su valor es siempre mayor o igual al de la heredabilidad. Tiene un particular interés por medir la correlación entre dos datos de un mismo individuo y por tanto sólo es de aplicación cuando se trabaja con datos repetidos de un mismo individuo como la producción de leche de una vaca, el resultado de una prueba deportiva de un caballo, el tamaño de camada o caracteres que puedan medirse de forma repetida sobre el mismo individuo como el peso de los dos jamones de un cerdo. Así, si llamamos y_i y y_j a los dos datos de un individuo, se da la siguiente igualdad:

$$r_{y_i y_j} = \frac{\sigma_{y_i y_j}}{\sigma_p^2}$$

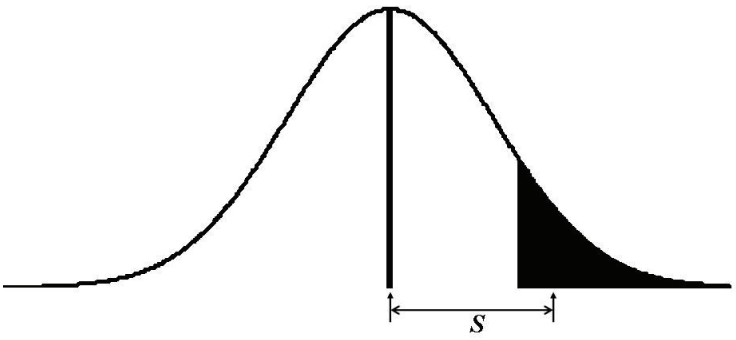
II-6. Selección

En una población que se encuentra en equilibrio la reproducción se produce totalmente al azar de manera que todos los individuos tienen la misma probabilidad de actuar como padres de la siguiente generación. Se puede definir la selección como la aparición preferente de determinados animales como padres de la siguiente generación en detrimento de otros.

Hablaremos en el contexto de este libro de selección haciendo referencia a la selección artificial, es decir, la llevada a cabo por el hombre con el fin de obtener rendimientos deseados, en contraste con la selección natural que es la llevada a cabo por la propia naturaleza que favorece la reproducción de animales que presentan alguna ventaja con respecto a los otros posibles en el ambiente en el que se desarrolla la población.

II-6.1. Respuesta a la selección

En el desarrollo de la respuesta a la selección se asumirá por simplicidad que se trata de selección por truncamiento, es decir, que se marcará un umbral sobre la distribución de fenotipos a partir del cual se escogerían como reproductores todos los que tengan igual valor o superior:



La respuesta a la selección es el cambio que se produce en la media del carácter como consecuencia de la selección, o lo que es lo mismo, la diferencia entre la media de la población seleccionada y la de la población de la que se parte. Llamaremos diferencial de selección S a esta diferencia entre las medias de la población seleccionada y la población original.

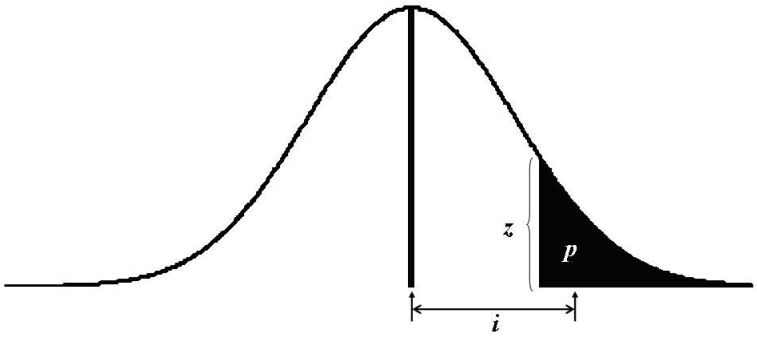
Obviamente, si se selecciona directamente sobre los valores genéticos de los animales, la respuesta a la selección será directamente el valor del diferencial de selección:

$$R = S$$

Sin embargo, la selección se realiza siempre sobre la base de los datos por lo que se produce una reducción en la respuesta, tanto mayor cuanto menor sea la heredabilidad. La respuesta a la selección es entonces la siguiente:

$$R = h^2 S$$

El diferencial de selección se expresa en unidades del carácter. Depende de la proporción seleccionada y de la variabilidad del carácter. La proporción seleccionada se mide como intensidad de selección y es el valor medio que poseen los individuos seleccionados si la variable presenta una distribución normal de media cero y varianza uno, aunque estos valores dependen también del tamaño de la población. Para poblaciones de tamaño suficientemente grande pueden utilizarse los valores de la distribución para poblaciones de tamaño infinito. Estos valores se obtienen mediante el cociente entre el valor de la ordenada en el umbral que delimita la población seleccionada y la propia proporción seleccionada:



Siendo $i = \frac{z}{p}$. Por ejemplo, 1,96 se corresponde con el umbral que delimita un intervalo de confianza del 95%, es decir, que dejaría por el lado derecho el 2,5%. Por tanto, para esta proporción seleccionada z = 1,96 y p = 0,025, y por tanto i = 2,338. Como esta distribución se encuentra tabulada, se pueden conocer con exactitud los distintos valores de intensidad de selección para distintos umbrales. A continuación se presenta una tabla que relaciona la intensidad de selección con la proporción seleccionada para distintas proporciones posibles:

Tabla 2.1. Valores de la intensidad de selección correspondientes a distintas proporciones seleccionadas.

Proporción (%)	*i*	Proporción (%)	*i*	Proporción (%)	*i*
0,01	3,960	1,2	2,603	21	1,372
0,02	3,790	1,4	2,549	22	1,346
0,03	3,687	1,6	2,502	23	1,320
0,04	3,613	1,8	2,459	24	1,295
0,05	3,554	2,0	2,421	25	1,271
0,06	3,507	2,2	2,386	26	1,248
0,07	3,464	2,4	2,353	27	1,225
0,08	3,429	2,6	2,323	28	1,202
0,09	3,397	2,8	2,295	29	1,180
0,10	3,367	3,0	2,268	30	1,159
		3,2	2,243	31	1,138
0,12	3,317	3,4	2,219	32	1,118
0,14	3,273	3,6	2,197	33	1,097
0,16	3,234	3,8	2,175	34	1,078
0,18	3,201	4,0	2,154	35	1,058
0,20	3,170	4,2	2,135	36	1,039
0,22	3,142	4,4	2,116	37	1,020
0,24	3,117	4,6	2,097	38	1,002
0,26	3,093	4,8	2,080	39	0,984
0,28	3,070	5,0	2,063	40	0,966
0,30	3,050			41	0,948
0,32	3,030	5,5	2,023	42	0,931
0,34	3,012	6,0	1,985	43	0,913
0,36	2,994	6,5	1,951	44	0,896
0,38	2,978	7,0	1,918	45	0,880
0,40	2,962	7,5	1,887	46	0,863
0,42	2,947	8,0	1,858	47	0,848
0,44	2,932	8,5	1,831	48	0,830
0,46	2,918	9,0	1,804	49	0,814
0,48	2,905	9,5	1,779	50	0,798
0,50	2,892	10	1,755		
0,55	2,862	11	1,709		
0,60	2,834	12	1,667		
0,65	2,808	13	1,627		
0,70	2,784	14	1,590		
0,75	2,761	15	1,554		
0,80	2,740	16	1,521		
0,85	2,720	17	1,489		
0,90	2,701	18	1,458		
0,95	2,683	19	1,428		
1,00	2,665	20	1,400		

Así pues, para obtener el diferencial de selección *S*, se busca en la tabla el valor de la intensidad de selección para la proporción seleccionada y se multiplica por la desviación típica fenotípica del carácter: $S = i\,\sigma_p$. La respuesta a la selección queda entonces:

$$R = ih^2\sigma_p$$

Finalmente, dado que es deseable conocer la respuesta a la selección por unidad de tiempo, esta expresión se puede expresar de esta manera:

$$R = \frac{ih^2\sigma_p}{T}$$

En esta expresión *T* es el intervalo generacional que se define como la edad que tienen los padres de los reproductores el momento del nacimiento de éstos.

II-6.1.1. Selección indirecta y respuesta correlacionada

Dos caracteres están correlacionados genéticamente cuando los valores genéticos de los individuos para estos dos caracteres también están correlacionados. Esta correlación puede tener dos orígenes:
-Pleiotropía: Se trata de la correlación existente al depender ambos caracteres de algún gen común. Por ejemplo, el color oscuro de ojos y del cabello son dos caracteres que están correlacionados por estar ambos influidos por los mismos genes relacionados con la pigmentación.
- Ligamiento: Se debe a la proximidad que poseen los respectivos genes en el genoma del individuo, de manera que durante la meiosis es poco probable que se hereden de forma separada.

Generalmente la correlación genética entre caracteres suele tener origen pleiotrópico.

En ocasiones se selecciona para mejorar un carácter utilizando como información las medidas de otro con el que está

correlacionado. Este tipo de selección se conoce como selección indirecta y es el caso en el que el carácter objetivo de selección es difícil de medir, siendo en cambio más abordable el registro de otro con el que tiene elevada correlación. Particularmente interesante es otro caso en el que el carácter objetivo de selección tiene menor heredabilidad siendo la correlación genética elevada entre ambos.

Conocer la respuesta esperada por selección indirecta es entonces de interés de cara a conocer su eficacia. La respuesta correlacionada obtenida en el carácter *x* cuando se selecciona a partir de la información del carácter *y* (RC_y), se mide mediante la regresión del valor genético de *y* sobre el de *x* (*b*), multiplicada por la respuesta obtenida sobre el propio carácter *x*, llamada en este contexto respuesta directa(R_x):

$$RC_y = bR_x = \frac{\sigma_{u_x u_y}}{\sigma^2_{u_x}} R_x = \frac{\sigma_{u_x u_y}}{\sigma^2_{u_x}} \frac{\sigma_{u_y}}{\sigma_{u_y}} R_x = \frac{\sigma_{u_x u_y}}{\sigma_{u_x}\sigma_{u_y}} \frac{\sigma_{u_y}}{\sigma_{u_x}} R_x = r_G \frac{\sigma_{u_y}}{\sigma_{u_x}} R_x$$

En esta expresión $\sigma_{u_x u_y}, \sigma^2_{u_x}, \sigma^2_{u_y}$ y r_G son, respectivamente, la covarianza genética entre *x* e *y*, las correspondientes varianzas genéticas y la correlación genética entre los dos. Operando sobre esta expresión de forma sencilla se puede poner en función de las raíces cuadradas de las respectivas heredabilidades:

$$RC_y = r_G h_x h_y i_x \sigma_y$$

CONCEPTOS CLAVE

- ¿Es razonable el modelo infinitesimal? ¿Se ajusta a la realidad? ¿Por qué se puede imaginar que funcionará en la práctica?
- ¿Cómo es la partición del fenotipo? ¿Cómo es la partición del genotipo?
- ¿Qué partes del fenotipo y del genotipo se deben a la interacción entre los efectos de los genes?
- ¿Qué es la heredabilidad en sentido amplio y en sentido estricto?
- ¿Mide la heredabilidad el grado en que se hereda un carácter?
- ¿Qué tipo de caracteres tienen heredabilidades altas y bajas?
- ¿Qué parámetro mide la correlación entre varias medidas del mismo carácter en el mismo individuo? ¿Qué relación tiene este parámetro con la heredabilidad?
- ¿De qué factores depende la respuesta a la selección?
- ¿La intensidad de selección será mayor o menor cuanto mayor sea la proporción seleccionada? ¿Y la respuesta a la selección?
- ¿Qué son la pleiotropía y el ligamiento?

TERCERA PARTE

LOS MODELOS LINEALES EN LA VALORACIÓN GENÉTICA ANIMAL

RESUMEN

Probablemente éste es el capítulo de mayor importancia. Se inicia haciendo una pequeña introducción de la metodología de los modelos lineales justificando su utilización por su sencillez de manejo y extraordinaria adaptación en el manejo de datos. Se recogerán observaciones de la variable de trabajo y su variabilidad será explicada en función de una serie de factores, considerados unos como fijos y otros como aleatorios, los primeros como continuos o discontinuos. Con la ayuda de un ejemplo numérico sencillo se define un modelo, denominado mixto por incorporar ambos tipos de factores, y se muestra en qué consiste esta definición del mismo, detallando sus tres etapas, la ecuación del modelo, la definición de las esperanzas y varianzas de los elementos que lo componen y las asunciones, restricciones y limitaciones del mismo. Se exponen las implicaciones que supone la definición de un factor como fijo o aleatorio en el momento de definir el modelo.

La mejora genética se fundamenta en la variabilidad presente en los datos de los individuos. Así, si todos ellos presentaran idéntico rendimiento, no existiría la posibilidad de decidir cuáles serían seleccionados como padres de las futuras generaciones. Esta variabilidad se debe a múltiples causas, por lo que realizar una modelización de la variabilidad desentrañando sus causas supone un reto importante antes de pasar a analizar la importancia de cada una de ellas. En este sentido podemos distinguir varias etapas:

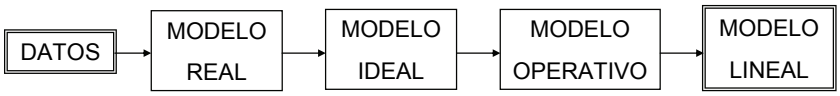
MODELO REAL: En el modelo real se conocen todos los factores o efectos causantes de la variabilidad, así como la manera en que influyen sobre el rendimiento. Así, bajo este modelo no sería preciso medir el rendimiento de los individuos si conocemos los factores que lo afectan, sino que bastaría aplicar el modelo real para obtener su valor. Aunque en cierto modo solemos tener algunas ideas sobre el modelo real, por desgracia, en la mayoría de las situaciones estamos lejos de conocerlo por lo que debemos renunciar al mismo.

MODELO IDEAL: Este modelo es el que formula el investigador intentando aproximar al máximo el modelo real. En principio sería deseable utilizar este modelo, pero en la práctica normalmente no existe suficiente información sobre el mismo.

MODELO OPERATIVO: Este modelo es el que queda después de intentar poner en práctica el modelo ideal, y es el que se llevará a la práctica. En el modelo operativo ya no están aquellos factores que sabemos que influyen en los datos pero que no van a ser tenidos en cuenta, normalmente porque no se ha registrado información sobre los mismos o porque pensamos que su influencia es poco importante.

MODELO LINEAL: De todos los modelos operativos posibles trabajaremos con los modelos lineales dado que producen un extraordinario ajuste de la variabilidad de los datos siendo sencillos de utilizar. El concepto de linealidad supone expresar un dato

como función lineal de los efectos, es decir, simplemente sumar los valores de los efectos ocasionalmente multiplicados por algún factor. Esto es evidentemente mucho más sencillo que buscar otras relaciones matemáticas más complejas como exponenciales, trigonométricas, etc.



III-1. Conceptos generales

Se presentan a continuación algunos conceptos que serán útiles después en la definición de un modelo lineal.

III-1.1. Observaciones o datos

Las observaciones son las medidas del carácter que deseamos mejorar. Son siempre tratadas como variables aleatorias con su distribución estadística por lo que se suele utilizar la palabra "variable" para referenciarlas. Es precisamente la variabilidad la propiedad que explotan los modelos lineales. Las variables pueden ser de ambos tipos, discretas o continuas. Las variables discretas, también llamadas categóricas, sólo toman valores enteros, y precisan normalmente un tratamiento más complejo por lo que inicialmente se presentará el modelo asumiendo que la variable de trabajo es de tipo continuo, es decir, la que puede tomar cualquier valor dentro de su rango.

III-1.2. Factores o efectos

Factor y efecto serán vocablos usados como sinónimos en este contexto. Un factor o efecto es una causa de variabilidad que ha sido identificada y que será incluida en el modelo como parte de la explicación de la variabilidad de un carácter. Por ejemplo, el sexo del individuo, el ser macho o hembra, implica variabilidad en el

peso adulto dado que normalmente los machos pesan más que las hembras. En todos los modelos se ajustarán alguno o algunos efectos en el modelo con el objeto de realizar inferencias sobre el mismo o sobre los mismos. Por ejemplo, el ajuste del sexo como factor en el modelo podría hacerse para intentar conocer las diferencias en el peso adulto que hay entre machos y hembras. Sin embargo, será frecuente el ajuste de factores o efectos causantes de variación, únicamente con el objeto de reducir al máximo la parte de variabilidad no explicada por el modelo. Así, por ejemplo, si se desea conocer el valor genético de los individuos para su peso adulto, ajustaremos el efecto genético en el modelo para hacer inferencias sobre el mismo, pero también incluiremos el efecto sexo. Obsérvese que si no se incluyese este efecto los machos aparecerían sistemáticamente mejor valorados que las hembras al confundirse este efecto con el genético.

Aunque ya fue tratado en el capítulo relacionado con la estadística, se repite aquí por su importancia la diferenciación en efectos fijos y aleatorios. En general, los factores de los que depende la variable de trabajo pertenecerán siempre a una de estas dos categorías, efectos fijos y efectos aleatorios.

- **Factores fijos:**

Son aquéllos que presentan pocos niveles (clases o categorías) en la población y todos aparecen en nuestros datos. Un ejemplo de efecto fijo es el sexo de los animales en la variable peso al nacimiento. Es un factor que sólo puede presentar dos niveles o categorías, sexo macho y sexo hembra, y en nuestros datos, aunque sólo son una muestra de todos los datos de la población, aparecen animales que pertenecen a uno de esos dos niveles, o son machos, o son hembras. Conceptualmente se asume que el pertenecer a una de las categorías del efecto representa una diferencia fija con respecto a los otros niveles del efecto, aunque esta diferencia es desconocida. Por ejemplo, se asume que las hembras pesarán sistemáticamente k kilogramos menos que los machos. Por esta razón se suelen llamar también efectos sistemáticos. Se dice que se va a *estimar* efectos fijos cuando se pretende obtener valores aproximados de los efectos fijos. Los factores fijos pueden ser a su vez continuos o discontinuos:

- Efectos fijos continuos. El efecto o factor es una variable y por tanto puede tomar cualquier valor dentro de su rango. Un ejemplo sería la edad en meses en relación al peso al destete. La relación entre la variable de trabajo o variable dependiente y el efecto fijo continuo o variable independiente es fácil de explicar, dado que es de esperar que los individuos de más edad presenten mayores pesos al destete. La relación entre ambos es proporcional y suele emplearse un coeficiente de proporcionalidad entre ambas que coincide con el coeficiente de regresión. Por ello, los efectos fijos continuos se llaman también variables regresoras o covariables.
- Efectos fijos discontinuos. El efecto fijo agrupa las observaciones en clases de manera que una observación puede pertenecer a una clase o a otra. El sexo es un ejemplo claro al poder pertenecer el dato a un macho o a una hembra, sin que exista una relación de proporcionalidad entre ambas. Es el tipo de efecto que se define al llevar a cabo un análisis de varianza en el que se pretende comparar series de medias de forma simultánea. Un efecto fijo discontinuo presente en todos los modelos lineales es la media general (μ) que agrupa a todas las observaciones dentro de su único nivel o categoría. Como veremos después, los modelos que presentan algún efecto discontinuo además de la media general son modelos de rango no completo y no presentan solución única.

- **Factores aleatorios:**

Presentan muchos niveles (hasta el punto de asumir conceptualmente que hay infinitos) en la población y en nuestra muestra sólo están contemplados una muestra representativa de los mismos, considerados como una representación aleatoria del resto. El ejemplo más interesante en nuestro contexto es el efecto genético aditivo identificado por el individuo. En la muestra no están representados los infinitos animales posibles sino que se escogen sólo unos cuantos. A diferencia de los efectos fijos, los efectos aleatorios no presentan valor constante único desconocido sino que son en sí mismos una distribución estadística. Conceptualmente se asume que todos los niveles pertenecen a una distribución, generalmente normal, con media y varianza, pero el valor esperado de cada uno de los niveles no es un valor

sistemático diferente de los otros niveles, sino que el valor esperado de todos ellos es el mismo, el de la media de su distribución. Obsérvese que en los efectos aleatorios no tiene sentido la estimación, es decir, el conocer el valor exacto del efecto, ya que éste, como valor exacto, no existe en la realidad. Por el contrario, sólo se pretende conocer el valor de algunos de los elementos de la distribución, lo que conceptualmente supondría aproximar su valor antes de hacer una extracción aleatoria de los mismos. Por ejemplo, si la variable aleatoria fuera el resultado de tirar un dado, para muchos apostantes sería muy interesante ser capaz de acertar el valor que proporcionará. Así, el hecho de conocer el valor aproximado de los efectos aleatorios se conoce con el nombre de ***predecir***, quedando la estimación circunscrita a los efectos fijos.

La clasificación de un determinado efecto o factor como fijo o aleatorio fue suficientemente tratada en los temas relacionados con las generalidades estadísticas y no será desarrollado aquí en profundidad.

III-1.2.1. Tipos de modelos en función de los factores

Todos los modelos lineales tienen al menos un efecto fijo y un efecto aleatorio:

- Efecto fijo común: media general (μ). Aparece en la ecuación de todas las observaciones y tiene un único nivel.

- Efecto aleatorio común: error o residuo (e_i). Tiene tantos niveles como datos existiendo un único nivel independiente para cada observación. Se suele asumir para el mismo una distribución normal de media cero.

Los modelos lineales utilizados en mejora genética animal suelen contemplar otros efectos tanto fijos como aleatorios adicionales. Si no es así se definen:

- Modelo aleatorio: El único efecto fijo que incluyen es la media general (μ).

- Modelo fijo: El único efecto aleatorio que incluyen es el residuo (e_i).

- Modelo mixto: Además de la media general y el residuo (μ y e_i) incluyen al menos otro efecto fijo y otro efecto aleatorio.

III-2. Definición de un modelo lineal mixto

Afrontaremos a continuación la definición de un modelo lineal mixto de forma general.

Utilizaremos para desarrollar la teoría un ejemplo sin importancia práctica pero de fácil interpretación. En el mismo se pretende evaluar genéticamente un grupo de individuos para tamaño de pie, utilizando como fuente de información el número del zapato y ajustando la información para el efecto sexo dado que los hombres suelen calzar un número superior al de las mujeres. La información recogida es la siguiente:

Individuo	Sexo	Número de zapato
1	♀	37
2	♀	38
3	♂	42
4	♀	36
5	♂	44

La definición de un modelo lineal mixto supone necesariamente el establecimiento de tres apartados:

- La ecuación del modelo
- Esperanzas y varianzas del modelo
- Asunciones restricciones y limitaciones

III-2.1. Ecuación del modelo

Comenzaremos escribiendo las ecuaciones para cada uno de los números de zapatos registrados de acuerdo con el planteamiento general, según el cuál, todos los registros valdrán aproximadamente la media del carácter (μ), efecto que tal y como fue indicado más arriba, será considerado como efecto fijo con una única clase o nivel.

Es de esperar que las diferencias en el número del zapato se deban, en primer lugar, al hecho de pertenecer a un varón o a una mujer. Por tanto, consideraremos un segundo efecto, el efecto sexo (denotado con la letra *s*), con dos niveles o clases, el varón al que adjudicaremos arbitrariamente el primer nivel (s_1) y la mujer al que asignaremos el segundo nivel (s_2). Obsérvese que podríamos haberlos asignado exactamente al revés. El índice de nivel o clase no indica por tanto nada y sólo sirve para distinguir unas clases de otras, lo que indica que se trata de un efecto discontinuo que clasifica los datos como pertenecientes a una u otra clase del efecto. Asimismo, obsérvese que no existen otras posibles clases de sexo por lo que debemos considerar el efecto como fijo.

Dado que no todos los datos dentro de cada clase del efecto sexo son iguales, debe quedar por explicar aún parte de la variabilidad de los datos. Es lógico pensar que habrá individuos que de nacimiento tengan tendencia a tener un determinado tamaño de pie, y que esta tendencia pueda ser genética, es decir, que los individuos con un valor genético alto para tamaño de pie, tengan parientes con la misma tendencia. Se puede definir entonces un efecto genético que llamaremos efecto genético aditivo que será denotado con la letra *u*, y que tendrá tantos niveles como individuos pertenezcan a nuestra población (u_1, u_2, u_3, u_4 y u_5). Obsérvese que el análisis se puede realizar con cada grupo de individuos de manera que un investigador que utilizase el mismo modelo incluiría el efecto sexo con los mismos dos niveles (varones y mujeres), pero su efecto genético aditivo presentaría niveles distintos. El primero consideraría por ejemplo el efecto genético aditivo de Juan y de María, mientras que en el segundo estos individuos no estarían sino que por ejemplo el efecto

contemplaría los valores genéticos aditivos de Luis y Marta. Este hecho convierte este efecto en efecto aleatorio.

Finalmente, tal y como fue comentado anteriormente, se incluye un efecto aleatorio residual con tantos residuos como datos. Este efecto es un cajón de sastre ya que incluye todas las causas de variabilidad no contempladas en el resto de los factores o efectos del modelo. Se incluyen así, factores que no han sido registrados aunque se sabe de su influencia, errores de medición, factores cuya influencia desconocemos, etc.

Obsérvese que la inclusión de un efecto en el modelo no siempre se realiza con el objeto de hacer inferencias sobre él, sino que su inclusión permite reducir la variabilidad de los residuos, detectando mejor las diferencias entre los niveles de otros efectos sobre los que se centra nuestro interés. Por ejemplo, si se desea detectar los individuos con el mayor valor genético para tamaño de pie, será necesario incluir el efecto sexo para que el modelo pueda eliminar de los datos la variabilidad que produce este efecto. De no incluirse este efecto, se vería confundido con el que nos interesa. En resumen, todos los individuos con mayor valor genético según ese modelo serían varones, enmascarándose el valor genético alto de los individuos que fueran mujeres.

Por tener efectos fijos y aleatorios además de la media general y el residuo, se trata de un modelo mixto cuya ecuación genérica para cada número de zapatos y_i, en álgebra de escalares sería:

$$y_i = \mu + s_j + u_k + e_i$$

y cuyas ecuaciones detalladas para cada uno de los datos son las siguientes:

$$
\begin{array}{lllllllllll}
37 & = & \mu & & + s_2 & + u_1 & & & & & + e_1 \\
38 & = & \mu & & + s_2 & & + u_2 & & & & + e_2 \\
42 & = & \mu & + s_1 & & & & + u_3 & & & + e_3 \\
36 & = & \mu & & + s_2 & & & & + u_4 & & + e_4 \\
44 & = & \mu & + s_1 & & & & & & + u_5 & + e_5
\end{array}
$$

Estas ecuaciones pueden ser tratadas conjuntamente en una sola expresión mediante notación matricial:

GENERALIDADES SOBRE ÁLGEBRA DE MATRICES

- Una matriz es un conjunto de elementos ordenados en forma de *f* filas y *c* columnas que se identifica normalmente con una letra mayúscula en negrita (**M**).
- Los elementos de una matriz **M** se representan con dos subíndices que proporcionan las coordenadas del elemento dentro de la matriz. Así, m_{ij} sería el elemento de la matriz **M** que se encontraría en la fila *i* y en la columna *j*.
- Una matriz con una sola fila y una sola columna no es en realidad una matriz y se denomina escalar.

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & ..... & m_{1n} \\ m_{21} & m_{22} & m_{23} & ..... & m_{2n} \\ ... & ... & ... & ..... & ... \\ ... & ... & ... & ..... & ... \\ m_{n1} & m_{n2} & m_{n3} & ..... & m_{nn} \end{bmatrix}$$

- Como excepción, existe un particular tipo de matrices que se llaman vectores y que se representan con letras minúsculas (**v**); son las que presentan una única fila o una única columna, tratándose siempre de un vector columna si no se especifica nada al respecto.
- Una matriz cuadrada es una matriz con idéntico número de filas y columnas.
- La diagonal principal (o simplemente diagonal) de una matriz **M**, es el conjunto de elementos que poseen idéntico subíndice de fila y columna (m_{ii}).
- Una matriz diagonal es aquélla con todos los elementos nulos fuera de la diagonal principal.
- La matriz con todos sus elementos igual a 0 excepto los de la diagonal en la que todos son 1, se llama matriz identidad o matriz unidad, se representa por **I** y es el elemento neutro del producto de matrices.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

En esta expresión **y** es el vector de datos que contiene los números de zapatos de los individuos, **b** es el vector de efectos fijos que contiene las tres incógnitas fijas del modelo (μ, s_1 y s_2), y **u** es el vector de efectos aleatorios que contiene las cinco incógnitas aleatorias de este modelo (u_1, u_2, u_3, u_4, y u_5).

Las matrices **X** y **Z** se llaman matrices de incidencia o matrices diseño, **X** para los efectos fijos y **Z** para los aleatorios. En el caso

de factores discontinuos, como los presentados aquí, este tipo de matrices están compuestas de ceros y unos. El número de filas de estas matrices se corresponde con el número de datos mientras que el número de columnas se corresponde con el número de incógnitas fijas (**X**) o aleatorias (**Z**). La construcción de estas matrices se lleva a cabo poniendo en cada posición x_{ij} un 1 si se incluye la incógnita de la columna *j* al registro de la fila *i*, y un 0 en caso contrario.

$$\begin{matrix} & & & \mu & s_1 & s_2 & & & u_1 & u_2 & u_3 & u_4 & u_5 & & \\ \begin{bmatrix} 37 \\ 38 \\ 42 \\ 36 \\ 44 \end{bmatrix} & = & \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} & \begin{bmatrix} \mu \\ s_1 \\ s_2 \end{bmatrix} & + & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} & + & \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} \\ \mathbf{y} & = & \mathbf{X} & \mathbf{b} & + & & \mathbf{Z} & & \mathbf{u} & + & \mathbf{e} \end{matrix}$$

Obsérvese que la primera columna de la matriz de incidencia de los efectos fijos está compuesta de unos y se corresponde a la suma de las otras dos que son mutuamente excluyentes. Esto conllevará alguna dificultad en el momento de resolver el modelo que será comentada después. Así, si además existiese otro efecto fijo, por ejemplo, la provincia de procedencia, en la que el primer individuo fuera de Barcelona, los dos siguientes de Madrid, y los dos últimos de Santander, este problema se daría por partida doble, siendo en este caso **Xb**:

$$\mathbf{Xb} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

La matriz de incidencia de los efectos aleatorios, Z, ha resultado una matriz de identidad. Obsérvese que esto es así porque todos los individuos poseen un único dato, y que no tendría esta forma en caso de que alguno de ellos tuviera más de un dato o no presentase registro.

OPERACIONES SENCILLAS EN ÁLGEBRA DE MATRICES

- La transpuesta de una matriz es otra en la que cada columna se corresponde con cada fila de la original. La matriz simétrica de **A** se representa por **A'**, y entre los elementos de ambas se da la relación $a_{ij}=a_{ij}'$.
- Una matriz **S** es simétrica si **S'** = **S**.
- La suma de dos matrices **A** y **B** sólo podrá llevarse a cabo si ambas son de las mismas dimensiones y será otra matriz **S** en la que $s_{ij} = a_{ij} + b_{ij}$.
- Del mismo modo, la resta de dos matrices **A** y **B** sólo podrá llevarse a cabo si ambas son de las mismas dimensiones y será otra matriz **R** en la que $r_{ij} = a_{ij} - b_{ij}$.
- El producto de dos matrices **A** (de *f* filas y *c* columnas) y **B** de (de *g* filas *d* columnas) sólo podrá llevarse a cabo si el número de columnas de **A** es igual al número de filas de **B** (si $c = g$) y dará lugar a una matriz **P** (**P** = **A B**) con *f* filas y *d* columnas de tal modo que cada elemento de **P** se obtendrá como

$$p_{jk} = \sum_{i=1}^{c} a_{ji} b_{ik} = a_{j1} b_{1k} + a_{j2} b_{2k} + ... + a_{jc} b_{ck} .$$

- Cualquier matriz multiplicada por la matriz identidad es igual a la propia matriz. **A I** = **A**
- La transpuesta de un producto de matrices es igual al producto de las transpuestas en orden inverso: (**A B C**)' = **C' B' A'**

Obsérvese cómo todas y cada una de las ecuaciones para cada uno de los zapatos se obtiene de la ecuación del modelo en álgebra matricial:

$$\begin{bmatrix} 37 \\ 38 \\ 42 \\ 36 \\ 44 \end{bmatrix} = \begin{bmatrix} 1\mu + 0s_1 + 1s_2 \\ 1\mu + 0s_1 + 1s_2 \\ 1\mu + 1s_1 + 0s_2 \\ 1\mu + 0s_1 + 1s_2 \\ 1\mu + 1s_1 + 0s_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} \mu + s_2 + u_1 + e_1 \\ \mu + s_2 + u_2 + e_2 \\ \mu + s_1 + u_3 + e_3 \\ \mu + s_2 + u_4 + e_4 \\ \mu + s_2 + u_5 + e_5 \end{bmatrix}$$

III-2.1.1. Ajuste de efectos fijos continuos

Así como los modelos con efectos fijos discontinuos son los mismos que se emplean en los análisis de varianza, los modelos con efectos fijos discontinuos se corresponden con los modelos de regresión. Efectivamente, los efectos fijos continuos no clasifican las variables, sino que se asume una relación lineal entre la variable dependiente y el valor del efecto fijo continuo que se denomina habitualmente como variable regresora o covariable y se asume medida sin error. En el ejemplo utilizado en el texto podríamos por ejemplo ajustar la edad de la persona en meses argumentando que tal vez en ese período el pie crece con la edad. Lo que se desea obtener en este caso es el coeficiente de regresión, o lo que es lo mismo, lo que incrementa la variable en estudio cuando la covariable aumenta en una unidad.

Habitualmente, con el fin de aislar la influencia de la covariable en el modelo ésta se suele introducir desviada con respecto a su media, asumiendo así que el individuo que se encuentre en el valor medio de la covariable no incrementará ni disminuirá su valor. Así por ejemplo, si se quiere ajustar la edad en meses, es necesario previamente calcular la media de los datos que en el ejemplo resulta ser igual a 381, y desviar los datos con respecto a la media:

Individuo	Sexo	Edad en meses	Covariable edad	Número de zapato
1	♀	385	385 - 381 = 4	37
2	♀	380	380- (-381) = -1	38
3	♂	392	392 - 381 = 11	42
4	♀	376	376 - 381 = -5	36
5	♂	372	372 -381 = -9	44

Las ecuaciones presentadas anteriormente se ajustan ahora incluyendo una nueva incógnita, el coeficiente de regresión (b_0) de la covariable edad en meses:

$$
\begin{array}{lllllllllll}
37 & = & \mu & & + s_2 & + u_1 & & & & + 4\, b_e & + e_1 \\
38 & = & \mu & & + s_2 & & + u_2 & & & - 1\, b_e & + e_2 \\
42 & = & \mu & + s_1 & & & & + u_3 & & + 11\, b_e & + e_3 \\
36 & = & \mu & & + s_2 & & & & + u_4 & - 5\, b_e & + e_4 \\
44 & = & \mu & + s_1 & & & & & + u_5 & - 9\, b_e & + e_5
\end{array}
$$

Y en la ecuación del modelo en álgebra matricial, la matriz de incidencia o matriz diseño de los efectos fijos incluiría coeficientes distintos de unos o ceros:

$$
\begin{bmatrix} 37 \\ 38 \\ 42 \\ 36 \\ 44 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 4 \\ 1 & 0 & 1 & -1 \\ 1 & 1 & 0 & 11 \\ 1 & 0 & 1 & -5 \\ 1 & 1 & 0 & -9 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \\ b_e \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}
$$

Tal y como está desarrollado hasta este punto todas las relaciones entre la covariable y la variable en estudio es lineal, es decir, la representación gráfica a que daría lugar sería una línea recta. En ocasiones la evolución de una variable con respecto a la otra no es lineal, sino que puede variar. Por ejemplo, el crecimiento en función de la edad puede considerarse aproximadamente lineal si los datos se recogen en períodos muy estrechos de tiempo, pero no será así si el periodo de recogida de datos es más amplio. Así, es razonable pensar que los individuos crecen más en sus edades más jóvenes, disminuyendo este crecimiento después. Para ello basta ajustar la misma covariable de forma polinómica (cuadrática, cúbica, etc.). Por ejemplo, un ajuste cuadrático se realizaría sencillamente usando como covariable el cuadrado de la edad en lugar de la propia edad. Existen muchas covariables que se ajustan frecuentemente de forma polinómica, siendo los más comunes el ajuste lineal y el ajuste lineal junto con el cuadrático.

III-2.2. Esperanzas y varianzas del modelo

En este apartado se definen qué efectos son fijos y aleatorios, así como las relaciones entre los distintos elementos aleatorios del modelo.

Para los efectos fijos se asumirá que existe un verdadero valor del efecto que siempre será desconocido pero que se desea estimar.

Para los efectos aleatorios se asumirá que no existe un verdadero valor del efecto sino una distribución con media y varianza, por lo que no se desea conocer su verdadero valor sino predecir el comportamiento de uno de los niveles de la distribución sobre el que se posee algún tipo de información. Por ejemplo, para el residuo se asumirá que la variable se distribuye de forma normal con media cero y una componente de varianza conocida (σ_e^2), lo que expresado en álgebra de escalares se representa como $e_i : N(0, \sigma_e^2)$. De esta asunción se deriva que el momento de primer orden, es decir, la esperanza (o valor esperado) de cada residuo es la media de su distribución (0), y el momento de segundo orden, la varianza de una realización de esta distribución (o lo que es lo mismo, uno de los elementos que pertenecen a la distribución), será asimismo la varianza de su distribución (σ_e^2). Sin embargo, la necesidad de expresar todas las relaciones entre todos los niveles de otros efectos aleatorios nos fuerza a expresar este apartado de la definición del modelo en álgebra matricial.

III-2.2.1. Esperanzas de los elementos del modelo

Como ha sido comentado el valor esperado de cada residuo es la media de su distribución, es decir, 0 ($E(e_i) = 0$), y por tanto, el valor esperado del vector de residuos es un vector de ceros:

$$E\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$, lo que puede ser abreviado como $E(\mathbf{e}) = \mathbf{0}$

El mismo razonamiento cabría para el vector de efectos genéticos aditivos si éstos perteneciesen a la misma distribución. Sin embargo es preciso distinguir los individuos fundadores (sin padres conocidos) del resto, ya que el proceso de selección conllevará una evolución, tanto en sus medias como en sus varianzas. Por tanto, la definición del efecto genético aditivo como $u_i \sim N(0, \sigma_u^2)$ sólo es válida para los individuos fundadores, debiendo tener en cuenta para el resto las relaciones de parentesco entre individuos. Ello será tenido en cuenta en el subapartado siguiente en el que se consideran las relaciones entre individuos. Para las esperanzas sigue siendo válido $E(\mathbf{u}) = \mathbf{0}$.

Para los efectos fijos no se asume una distribución sino su propio valor. Así, el valor esperado de la media general es la propia media general ($E(\mu) = \mu$), y lo mismo cabe definir para el resto de los niveles de efectos fijos ($E(s_1) = s_1$ y $E(s_2) = s_2$) lo que extendido al vector completo de efectos fijos se representaría como $E(\mathbf{b}) = \mathbf{b}$.

Finalmente el valor esperado del vector de datos se deduce con ayuda de las propiedades de la esperanza matemática:

$$E(\mathbf{y}) = E(\mathbf{Xb} + Z\mathbf{u} + \mathbf{e}) = E(\mathbf{Xb}) + E(\mathbf{Zu}) + E(\mathbf{e}) =$$
$$= \mathbf{X}E(\mathbf{b}) + \mathbf{Z}E(\mathbf{u}) + \mathbf{0} = \mathbf{Xb}$$

Obsérvese que **Xb** es la parte fija del modelo, es decir, $\mu + s_1$ si el número del zapato pertenece a un varón y $\mu + s_2$ si pertenece a una mujer. Es decir, que el valor esperado del número de zapato antes de medirlo sería el de la media más el efecto de ser de un sexo o de otro, o lo que es lo mismo, la media de los varones si se trata de un varón y la media de las mujeres si se tratara de una mujer. Obsérvese también que de existir un segundo efecto fijo

discontinuo, como por ejemplo la comunidad autónoma de procedencia, el valor esperado de un dato sería la media de los números de zapatos de los individuos del mismo sexo nacidos en la misma comunidad autónoma.

Finalmente, se pueden agrupar las esperanzas matemáticas en una única expresión:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

III-2.2.2. Varianzas de los elementos del modelo

Este punto consiste en definir las varianzas de cada uno de los elementos del modelo. Así, por ejemplo, si asumimos que los residuos se distribuyen normalmente según la definición $e_i \sim N(0,\sigma_e^2)$ vista más arriba, entonces cualquier realización de esta distribución tendrá como varianza σ_e^2, la varianza de su distribución de origen, es decir, $\mathrm{Var}(e_1) = \sigma_{e_1}^2 = \mathrm{Var}(e_i) = \sigma_e^2$. Asumir el mismo valor para todos y cada uno de los residuos es lo que se conoce como homogeneidad de la varianza residual.

Cuando se trata de definir la varianza de un vector (Var(**e**)), además de la varianza de cada elemento ($\sigma_{e_1}^2$), también hay que definir cada $\mathrm{Cov}(e_i, e_j)$, es decir la covarianza entre cada dos elementos del vector, que también puede representarse como $\sigma_{e_i e_j}$, por lo que es preciso recurrir a la notación matricial:

$$Var(\mathbf{e}) = \mathbf{R} = Var\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_1e_2} & \sigma_{e_1e_3} & \sigma_{e_1e_4} & \sigma_{e_1e_5} \\ \sigma_{e_2e_1} & \sigma^2_{e_2} & \sigma_{e_2e_3} & \sigma_{e_2e_4} & \sigma_{e_2e_5} \\ \sigma_{e_3e_1} & \sigma_{e_3e_2} & \sigma^2_{e_3} & \sigma_{e_3e_4} & \sigma_{e_3e_5} \\ \sigma_{e_4e_1} & \sigma_{e_4e_2} & \sigma_{e_4e_3} & \sigma^2_{e_4} & \sigma_{e_4e_5} \\ \sigma_{e_5e_1} & \sigma_{e_5e_2} & \sigma_{e_5e_3} & \sigma_{e_5e_4} & \sigma^2_{e_5} \end{bmatrix}$$

Obsérvese que todas las matrices de varianzas y covarianzas de un vector son matrices simétricas ya que $\sigma_{e_ie_j} = \sigma_{e_je_i}$, y por tanto, **R'**=**R**.

Asumir homogeneidad de la varianza residual permite simplificar la diagonal de la matriz **R**, pero es preciso establecer una asunción para las covarianzas entre los distintos errores o residuos. Recordemos que una covarianza representa la variación conjunta de dos variables. Así, si una de ellas tiende a ser elevada o reducida cuando la otra respectivamente también lo es, la varianza será positiva. Por el contrario, si una de las variables tiende a tener valores reducidos cuando la otra es elevada, o viceversa, entonces la covarianza será negativa. Cuando ambas variables son independientes, entonces el hecho de que una de ellas tome un valor elevado o reducido no influirá sobre la magnitud de la otra, y en este caso la covarianza entre ambas será nula. Dado que los residuos son la parte de la variabilidad del modelo no explicada por los efectos ajustados, parece lógico asumir que el hecho de que uno de los residuos sea elevado o reducido, no influirá en absoluto sobre la magnitud de los otros residuos, por lo que podemos asumir independencia entre residuos, es decir, $\sigma_{e_ie_j} = 0$ para todas las covarianzas de la matriz **R**:

$$Var(\mathbf{e}) = \mathbf{R} = Var\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \sigma_e^2 = \mathbf{I}\sigma_e^2$$

El otro efecto aleatorio de nuestro modelo es el efecto genético aditivo. Habrá que definir también, por tanto, la estructura de la matriz de varianzas y covarianzas de este efecto, matriz que llamaremos G y que tendrá la siguiente estructura:

$$Var(\mathbf{u}) = \mathbf{G} = Var\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_1u_2} & \sigma_{u_1u_3} & \sigma_{u_1u_4} & \sigma_{u_1u_5} \\ \sigma_{u_2u_1} & \sigma_{u_2}^2 & \sigma_{u_2u_3} & \sigma_{u_2u_4} & \sigma_{u_2u_5} \\ \sigma_{u_3u_1} & \sigma_{u_3u_2} & \sigma_{u_3}^2 & \sigma_{u_3u_4} & \sigma_{u_3u_5} \\ \sigma_{u_4u_1} & \sigma_{u_4u_2} & \sigma_{u_4u_3} & \sigma_{u_4}^2 & \sigma_{u_4u_5} \\ \sigma_{u_5u_1} & \sigma_{u_5u_2} & \sigma_{u_5u_3} & \sigma_{u_5u_4} & \sigma_{u_5}^2 \end{bmatrix}$$

Sería interesante simplificar igualmente esta matriz asumiendo homogeneidad de varianza genética aditiva e independencia entre niveles del efecto.

La homogeneidad de la varianza genética aditiva es coherente para todos los individuos fundadores pero, como fue indicado anteriormente, no para los individuos con ascendencia conocida. Algunos de estos individuos podrían haber heredado el mismo alelo tanto por la vía paterna como por la materna por tener ambos padres algún antepasado en común del que hubiesen heredado el mismo alelo. Esto originaría consanguinidad en el individuo y la

varianza de este individuo no sería entonces la de la distribución de origen.

La independencia entre los valores genéticos de diferentes individuos tampoco es coherente si en la realidad existe parentesco entre al menos algunos de ellos. Es razonable pensar que aquellos individuos con un elevado número de pie tendrán parientes con tendencia a tener también un elevado número de pie, y el mismo razonamiento se da para los individuos de pie pequeño. No parece lógico pensar en covarianzas nulas entre individuos emparentados, por lo que tampoco podemos asumir independencia entre valores genéticos aditivos.

Aunque la relación entre parientes constituye una dificultad en la definición del modelo, esta información es vital en la valoración genética de los individuos al permitir predecir el comportamiento de un individuo como reproductor.

Para la definición del modelo asumiremos que existe una matriz que contiene las relaciones entre los valores genéticos de los individuos, la matriz de relaciones aditivas que denotaremos por **A**, de manera que en caso de no existir relaciones de parentesco entre individuos se daría **A** = **I**. De acuerdo con esto, y asumiendo que los valores genéticos aditivos de los fundadores se distribuyen según $u_i \sim N(0, \sigma_u^2)$, la matriz de varianzas y covarianzas de los efectos genéticos aditivos se definiría:

$$Var(\mathbf{u}) = \mathbf{G} = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_1u_2} & \sigma_{u_1u_3} & \sigma_{u_1u_4} & \sigma_{u_1u_5} \\ \sigma_{u_2u_1} & \sigma^2_{u_2} & \sigma_{u_2u_3} & \sigma_{u_2u_4} & \sigma_{u_2u_5} \\ \sigma_{u_3u_1} & \sigma_{u_3u_2} & \sigma^2_{u_3} & \sigma_{u_3u_4} & \sigma_{u_3u_5} \\ \sigma_{u_4u_1} & \sigma_{u_4u_2} & \sigma_{u_4u_3} & \sigma^2_{u_4} & \sigma_{u_4u_5} \\ \sigma_{u_5u_1} & \sigma_{u_5u_2} & \sigma_{u_5u_3} & \sigma_{u_5u_4} & \sigma^2_{u_5} \end{bmatrix} =$$

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \sigma^2_u = \mathbf{A}\sigma^2_u$$

Dejaremos para más adelante la forma de obtener los distintos elementos de la matriz **A** pero anticiparemos que un elemento a_{jk} de esta matriz representa el porcentaje de genes que comparten los individuos *j* y *k*. Así, si estos individuos no son parientes este coeficiente valdrá cero.

También deben definirse las covarianzas entre elementos aleatorios de vectores distintos. De una forma global habría que definir las varianzas del modelo teniendo en cuenta todos estos elementos:

$$Var\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = Var\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_1u_2} & \sigma_{u_1u_3} & \sigma_{u_1u_4} & \sigma_{u_1u_5} & \sigma_{u_1e_1} & \sigma_{u_1e_2} & \sigma_{u_1e_3} & \sigma_{u_1e_4} & \sigma_{u_1e_5} \\ \sigma_{u_2u_1} & \sigma^2_{u_2} & \sigma_{u_2u_3} & \sigma_{u_2u_4} & \sigma_{u_2u_5} & \sigma_{u_2e_1} & \sigma_{u_2e_2} & \sigma_{u_2e_3} & \sigma_{u_2e_4} & \sigma_{u_2e_5} \\ \sigma_{u_3u_1} & \sigma_{u_3u_2} & \sigma^2_{u_3} & \sigma_{u_3u_4} & \sigma_{u_3u_5} & \sigma_{u_3e_1} & \sigma_{u_3e_2} & \sigma_{u_3e_3} & \sigma_{u_3e_4} & \sigma_{u_3e_5} \\ \sigma_{u_4u_1} & \sigma_{u_4u_2} & \sigma_{u_4u_3} & \sigma^2_{u_4} & \sigma_{u_4u_5} & \sigma_{u_4e_1} & \sigma_{u_4e_2} & \sigma_{u_4e_3} & \sigma_{u_4e_4} & \sigma_{u_4e_5} \\ \sigma_{u_5u_1} & \sigma_{u_5u_2} & \sigma_{u_5u_3} & \sigma_{u_5u_4} & \sigma^2_{u_5} & \sigma_{u_5e_1} & \sigma_{u_5e_2} & \sigma_{u_5e_3} & \sigma_{u_5e_4} & \sigma_{u_5e_5} \\ \sigma_{e_1u_1} & \sigma_{e_1u_2} & \sigma_{e_1u_3} & \sigma_{e_1u_4} & \sigma_{e_1u_5} & \sigma^2_{e_1} & \sigma_{e_1e_2} & \sigma_{e_1e_3} & \sigma_{e_1e_4} & \sigma_{e_1e_5} \\ \sigma_{e_2u_1} & \sigma_{e_2u_2} & \sigma_{e_2u_3} & \sigma_{e_2u_4} & \sigma_{e_2u_5} & \sigma_{e_2e_1} & \sigma^2_{e_2} & \sigma_{e_2e_3} & \sigma_{e_2e_4} & \sigma_{e_2e_5} \\ \sigma_{e_3u_1} & \sigma_{e_3u_2} & \sigma_{e_3u_3} & \sigma_{e_3u_4} & \sigma_{e_3u_5} & \sigma_{e_3e_1} & \sigma_{e_3e_2} & \sigma^2_{e_3} & \sigma_{e_3e_4} & \sigma_{e_3e_5} \\ \sigma_{e_4u_1} & \sigma_{e_4u_2} & \sigma_{e_4u_3} & \sigma_{e_4u_4} & \sigma_{e_4u_5} & \sigma_{e_4e_1} & \sigma_{e_4e_2} & \sigma_{e_4e_3} & \sigma^2_{e_4} & \sigma_{e_4e_5} \\ \sigma_{e_5u_1} & \sigma_{e_5u_2} & \sigma_{e_5u_3} & \sigma_{e_5u_4} & \sigma_{e_5u_5} & \sigma_{e_5e_1} & \sigma_{e_5e_2} & \sigma_{e_5e_3} & \sigma_{e_5e_4} & \sigma^2_{e_5} \end{bmatrix}$$

Todos los elementos de la matriz completa de varianzas y covarianzas han sido ya desarrollados excepto las covarianzas entre

cada valor genético aditivo y cada residuo. Parece lógico pensar que el hecho de que un individuo posea un determinado valor genético, bueno o malo, no implique nada sobre la magnitud del residuo que se ajusta en su ecuación. Sin embargo, si parte de este residuo es el trato diferencial del ganadero hacia cada individuo, esta asunción no es tan lógica, ya que los ganaderos tratarán mejor a sus mejores animales, lo que supondrá la existencia de interacción genotipo-ambiente. Sin embargo, por simplicidad se asume siempre esta independencia, de manera que se suele asumir para las varianzas:

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

PROPIEDADES DE LA VARIANZA EN ÁLGEBRA DE MATRICES

- Para la definición de la covarianza entre dos vectores, el segundo de ellos entrará en forma transpuesta. Así, si definimos la matriz **C** como la covarianza entre dos vectores **a** y **b**, el segundo participará en forma de vector fila:
C = Cov (**a**, **b'**), y su transpuesta: **C'** = Cov(**b**, **a'**).
- Varianza de la suma de dos vectores:
Var (**a** + **b**) = Var(**a**) + Var(**b**) + Cov (**a**, **b'**) + Cov(**b**, **a'**)
- Varianza de la diferencia de dos vectores:
Var (**a** - **b**) = Var(**a**) + Var(**b**) - Cov (**a**, **b'**) - Cov(**b**, **a'**)
- Varianza de un vector de constantes **b**: Var(**b**) = **0**.
- Varianza del producto de una matriz de constantes por una variable:
Var(**Ky**) = **K** Var(**y**) **K'**. Si llamamos **V** a la matriz de varianzas y covarianzas del vector **y**, entonces Var(**Ky**) = **K V K'**.
- Al producto que se obtiene multiplicando una matriz o un vector por una segunda matriz y nuevamente por la primera en forma transpuesta se le llama forma cuadrática, de manera que **K V K'** es una forma cuadrática.

Aunque no es estrictamente necesario, se puede incluir el vector de observaciones en la definición de las varianzas de los elementos del modelo. Para ello será preciso conocer:

- Var(**y**) = Var (**Xb** + **Zu** + **e**) =
= Var(**Zu**) + Var(**e**) + Cov(**Zu**, **e'**) + Cov(**e**, **u'Z'**) =
= **Z** Var(**u**) **Z'** + Var(**e**) + **Z** Cov(**u**, **e'**) + Cov(**e**, **u'**) **Z'** = **ZGZ'** + **R**.
- A la matriz de varianzas y covarianzas del vector de datos la denotaremos como **V**: **V** = Var(**y**)
- Cov(**y**, **e'**) = Cov(**Xb** + **Zu** + **e**, **e'**) = Cov(**Zu**, **e'**) + Cov(**e**, **e'**) = Var(**e**) = **R**

- $Cov(\mathbf{e}, \mathbf{y'}) = Cov(\mathbf{y}, \mathbf{e'})' = \mathbf{R'} = \mathbf{R}$.
- $Cov(\mathbf{y}, \mathbf{u'}) = Cov(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \mathbf{u'}) = Cov(\mathbf{Zu}, \mathbf{u'}) + Cov(\mathbf{e}, \mathbf{u'}) =$
  $= \mathbf{Z}\, Cov(\mathbf{u}, \mathbf{u'}) =$
  $= \mathbf{Z}\, Var(\mathbf{u}) = \mathbf{ZG}$
- $Cov(\mathbf{u}, \mathbf{y'}) = Cov(\mathbf{y}, \mathbf{u'})' = (\mathbf{ZG})' = \mathbf{G'Z'} = \mathbf{GZ'}$

Podemos ahora agrupar todas las definiciones de varianzas de elementos del modelo en una única expresión:

$$Var\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{ZGZ'+R} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ'} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{bmatrix}$$

III-2.3. Asunciones, restricciones y limitaciones del modelo

Este apartado de la definición del modelo describe lo que no es analizado por el mismo.

Esta tercera parte en la definición del modelo se refiere a cuestiones que no aparecen en las dos primeras, como por ejemplo, información acerca de la forma en que fueron recogidos los datos, si fueron obtenidos al azar entre todos los posibles, o si fueron escogidos según algún criterio.

Si un efecto no es incluido en el modelo habrá que asumir que no influye, restringir el análisis a algún nivel de este efecto, o advertir sobre la limitación del modelo en este sentido. Por ejemplo, si nuestro modelo para el número de zapatos no incluyese el efecto sexo habría que comentar acerca del mismo una de las siguientes opciones:

- Asunción: el modelo no incluye el efecto sexo porque se asume que no hay diferencias en el número de zapato entre mujeres y varones.
- Restricción: el modelo sólo incluye datos de mujeres, de manera que no se harán inferencias sobre el número de zapatos de los varones.

- Limitación: el modelo no incluye el efecto sexo porque no se ha recogido esta información ya que sólo se dispone de un código de identificación y el número de zapato siendo imposible conocer si pertenece a un varón o a una mujer.

En esta parte del modelo se reflejan las diferencias que hay entre el modelo ideal y el modelo operativo. Aunque por diversas razones esta parte de la descripción del modelo es a veces ignorada, contiene la información más importante para poder valorar el análisis.

En el ejemplo que nos ocupa se podría por ejemplo asumir que no hay diferencias en el número de zapatos de los individuos nacidos en distintas comunidades autónomas.

CONCEPTOS CLAVE

- ¿Puede considerarse la edad como una observación o como un factor?
- ¿Qué diferencia hay entre factores o efectos fijos y aleatorios?
- La media general, el sexo, la edad, el individuo, la ganadería, ¿cómo se clasifican? ¿como fijos o como aleatorios?
- ¿Se pueden estimar efectos aleatorios y predecir efectos fijos?
- ¿Qué representa una matriz de incidencia?
- ¿Cuántas etapas han de llevarse a cabo en la definición de un modelo lineal?
- ¿En qué momento de la definición de un modelo lineal se representan las diferencias entre el modelo ideal y el modelo operativo?
- ¿Qué tipo de efectos nos conducen a modelos de rango no completo?
- ¿Por qué no se incluyen los efectos fijos en la definición de las varianzas del modelo?
- ¿Qué estructura tiene la matriz de varianzas y covarianzas de un vector de efectos aleatorios cuando se asume homogeneidad de varianza e independencia entre los elementos del vector?
- ¿Cómo se suelen considerar las covarianzas entre efectos genéticos y los residuos? ¿Es razonable esta consideración?

CUARTA PARTE

RESOLUCIÓN DEL MODELO FIJO

RESUMEN

En este capítulo se trabaja con detalle la metodología de la estimación, es decir, la resolución del modelo fijo. Se comienza con su definición, lo que supone una simplificación del capítulo precedente. Se propone su resolución mediante el clásico método de ajuste por mínimos cuadrados y se resuelve un pequeño ejemplo numérico. Se incluye también un apartado para comentar la solución de variables regresoras o covariables o variables fijas continuas. Se discute con detalle la problemática de la dependencia lineal entre incógnitas fijas y se muestra cómo debe abordarse. Finalmente se comentan otros métodos de estimación de efectos fijos concretando sus propiedades estadísticas y se muestran los puntos en común de las diferentes metodologías

IV-1. El modelo fijo

Un modelo fijo es aquél que incluye como único efecto aleatorio el residuo (e_i). Siguiendo con el ejemplo del texto ajustaremos a los datos un modelo sin efectos aleatorios. El efecto genético aditivo descrito en el modelo mixto previo pasará a formar parte de ese cajón de sastre que llamamos error o residuo. Si en la ecuación del modelo mixto previa llamamos al residuo e_i*, presentaría la siguiente forma:

$$\mathbf{y = Xb + Zu + e^*}$$

El nuevo vector de residuos incluirá ahora el efecto genético aditivo: $\mathbf{e = Zu + e^*}$.

A continuación haremos una definición del modelo para nuestros datos y posteriormente se procederá a su resolución.

IV-2. Definición del modelo fijo

Se define a continuación un modelo proporcionando toda la información imprescindible. Detalles de esta definición pueden ser estudiados en el capítulo anterior.

- Ecuación del modelo:

$$\mathbf{y = Xb + e}$$

En esta ecuación, **y** representa el vector de datos que incluye los números de zapato de los individuos, **b** es el vector de incógnitas fijas que, además de la media general, incluye el efecto sexo con dos niveles, varones y mujeres, **X** es la matriz de incidencia de los efectos fijos y **e** es el vector de residuos.

- Esperanzas y varianzas de los elementos del modelo:

  - Esperanzas:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\end{bmatrix}$$

  - Varianzas

$$Var(\mathbf{y})=\mathbf{V}=Var(\mathbf{e})=\mathbf{R}=\mathbf{I}\sigma_e^2$$

En esta expresión σ_e^2 es la varianza residual. Obsérvese que **V** es la matriz de varianzas y covarianzas del vector de datos, mientras que **R** es la matriz de varianzas y covarianzas del vector de residuos, y que en el modelo fijo ambas matrices coinciden al ser el residuo el único efecto aleatorio del modelo.

- Asunciones, restricciones y limitaciones:

Aunque pensamos que van a existir diferencias entre los valores genéticos de los individuos para el tamaño del pie, no disponemos de información de parentesco entre los individuos, por lo que no podremos separar el valor genético de cada individuo del residuo del modelo. Esto supone una limitación del modelo.

No disponemos de individuos de otros países por lo que sólo haremos inferencias para los individuos españoles. Esto supondrá una restricción del modelo.

Asumimos también que todos los individuos calzan el mismo número de pie en el lado derecho y en el izquierdo.

IV-3. Resolución del modelo fijo

Resolver el modelo fijo consiste en aproximar el verdadero valor de los efectos fijos por lo que la resolución de efectos fijos se conoce como estimación de efectos fijos.

Desarrollaremos la resolución del modelo fijo sobre el método de ajuste por mínimos cuadrados ordinarios y posteriormente haremos una extrapolación a otros métodos.

IV-3.1. Método de ajuste por mínimos cuadrados ordinarios

Este método se basa en la idea de encontrar una solución de las incógnitas del modelo que proporcionen residuos lo más pequeños posibles para cada uno de los datos. Dado que los residuos tienen media cero (los hay positivos y negativos), y han de tenerse en cuenta todos los residuos en una sola expresión, se buscará una expresión que dependa de las incógnitas (**b**) y que proporcione la suma de cuadrados de los residuos para pasar a continuación a buscar los valores de **b** que hagan mínima esta expresión.

- Expresión para la suma de cuadrados de los residuos.

SUMA DE CUADRADOS DE LOS ELEMENTOS DE UN VECTOR

Como caso particular de las formas cuadráticas descritas en un punto anterior, la suma de cuadrados de los elementos de un vector se obtiene sencillamente multiplicando el vector en forma de vector fila por él mismo en forma de vector columna. Obsérvese que el producto de un vector de dimensiones (1 x n) por otro de dimensiones (n x 1), es un producto conformable que da lugar a un escalar (1 x 1) y que se corresponde con la suma de cuadrados de los elementos del vector. Por ejemplo:

$$\mathbf{e'e} = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 & e_5 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = (e_1)^2 + (e_2)^2 + (e_3)^2 + (e_4)^2 + (e_5)$$

Los residuos los podemos despejar de la ecuación del modelo:

$$\mathbf{e = y - Xb}$$

Función F que expresa la suma de cuadrados de los residuos y que depende de las incógnitas:

$$F = \mathbf{e'e} = (\mathbf{y - Xb})' (\mathbf{y - Xb}) = (\mathbf{y' - b'X'}) (\mathbf{y - Xb}) =$$
$$= \mathbf{y'y - y'Xb} - \mathbf{b'X'y} + \mathbf{b'X'Xb}$$

Obsérvese que todos los elementos de esta expresión son productos que comienzan con un vector fila (1 x n) y termina por un vector columna (n x 1), luego todos ellos dan lugar a escalares. En concreto **y'Xb** es el transpuesto de **b'X'y**, y por tanto son además el mismo número. De este modo, la función F puede expresarse como:

$$F = \sum e_i^{\,2} = \mathbf{y'y - 2\, b'X'y + b'X'Xb}$$

- Minimización de $\sum e_i^{\,2}$

La forma de obtener el máximo o el mínimo de una expresión es igualar su derivada primera a cero. Se deberá entonces derivar F con respecto a las incógnitas e igualar la expresión a cero.

$$\frac{\partial(\mathbf{y'y} - 2\,\mathbf{b'X'y} + \mathbf{b'X'Xb})}{\partial \mathbf{b}} = 0$$

DERIVACIÓN DE LAS ECUACIONES DE RESOLUCIÓN DEL MODELO FIJO

$$\frac{\partial(\mathbf{y'y} - 2\,\mathbf{b'X'y} + \mathbf{b'X'Xb})}{\partial \mathbf{b}} =$$

$$= \frac{\partial(\mathbf{y'y})}{\partial \mathbf{b}} - \frac{\partial(2\,\mathbf{b'X'y})}{\partial \mathbf{b}} + \frac{\partial(\mathbf{b'X'Xb})}{\partial \mathbf{b}} = 0$$

Se trata de la derivada de un polinomio de orden 2 en álgebra de matrices. Así, **y'y** es constante con respecto a b, por lo su derivada será cero. En 2 **b'X'y**, 2 **X'y** es la constante que multiplica a la incógnita, luego será el resultado de derivar este elemento. Finalmente, **b'X'Xb** es una forma cuadrática en la que **X'X** es la parte constante, luego debe derivarse disminuyendo en uno el exponente de la incógnita:

$$\frac{\partial(\mathbf{y'y})}{\partial \mathbf{b}} - \frac{\partial(2\,\mathbf{b'X'y})}{\partial \mathbf{b}} + \frac{\partial(\mathbf{b'X'Xb})}{\partial \mathbf{b}} = 0 - 2\,\mathbf{X'y} + 2\mathbf{X'X}\hat{\mathbf{b}} = 0$$

Obsérvese que después de derivar, la notación corresponde ya al estimador en lugar de al verdadero valor por lo que debe llevar sombrero. Ordenando esta expresión se obtiene:

$$(\mathbf{X'X})\hat{\mathbf{b}} = \mathbf{X'y}$$

La solución minimocuadrática del modelo fijo se obtiene resolviendo las ecuaciones:

$$(\mathbf{X'X})\hat{\mathbf{b}} = \mathbf{X'y}$$

En esta expresión **X'X** es lo que se conoce como matriz de coeficientes, $\hat{\mathbf{b}}$ es el vector de las incógnitas, y **X'y** es el vector del lado derecho o vector de términos independientes. Es la forma matricial de resolver un sistema de ecuaciones. Por ejemplo, las siguientes ecuaciones:

$$2x + y = 5$$
$$3x - 2y = 4$$

se escribirían de la siguiente manera en álgebra matricial:

$$\begin{bmatrix} 2 & 1 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

En nuestro caso, la matriz de coeficientes puede obtenerse mediante el producto de la transpuesta de **X** por sí misma. Asimismo el vector del lado derecho puede obtenerse mediante los correspondientes productos entre **X'** y el vector de datos **y**:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix};$$

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 37 \\ 38 \\ 42 \\ 36 \\ 44 \end{bmatrix} = \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix}$$

Obsérvese que la matriz **X** tiene tantas filas como registros (5) y tantas columnas como incógnitas en el vector de efectos fijos (3). Sin embargo la matriz **X'X** es una matriz cuadrada de tamaño 3. Nótese que si el número de registros fuera elevado, la matriz **X** y su transpuesta tendrían elevadas dimensiones pero **X'X** seguiría siendo una matriz cuadrada de tamaño 3. Lo mismo ocurre con el vector **X'y**: independientemente del número de registros su tamaño será igual al número de incógnitas, en este caso 3. Sin embargo, tanto **X'X** como **X'y** se pueden construir directamente sin necesidad de hacer productos de matrices. **X'X** se construye rellenando una tabla con tantas filas y columnas como número de

incógnitas: en cada celda se anotará el número de observaciones que en su modelo presentan simultáneamente ambas incógnitas. Por ejemplo, en el cruce de μ y s_1 se pondrá el número de observaciones que presentando la media en su ecuación (todas), presentan también s_1 (las correspondientes a chicas, es decir, 2). **X'y** será un vector columna en el que se anota sencillamente la suma de las observaciones que en su ecuación presentan cada una de las incógnitas. Por ejemplo, la segunda posición corresponderá a la suma de los números de zapatos de los varones:

X'X =		μ	s_1	s_1
	μ	5	2	3
	s_1	2	2	0
	s_1	3	0	3

X'y =		
	μ	197
	s_1	86
	s_1	111

La expresión $(\mathbf{X'X})\hat{\mathbf{b}} = \mathbf{X'y}$ queda finalmente de la siguiente manera:

$$\begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \end{bmatrix} = \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix}$$

El producto de cada fila por el vector de incógnitas igualado al término del lado derecho se interpreta como una ecuación para la incógnita que aparece en esa posición en el vector de incógnitas. Así, mediante el producto de la matriz de coeficientes por el vector de incógnitas se comprueba que se trata de un sistema de 3 ecuaciones con 3 incógnitas:

$$5\hat{\mu} + 2\hat{s}_1 + 3\hat{s}_2 = 197$$
$$2\hat{\mu} + 2\hat{s}_1 = 86$$
$$3\hat{\mu} + 3\hat{s}_2 = 111$$

La forma de resolver la expresión $(\mathbf{X'X})\hat{\mathbf{b}} = \mathbf{X'y}$ es despejar $\hat{\mathbf{b}}$, lo que requiere la inversión de la matriz de coeficientes.

INVERSA DE UNA MATRIZ

- La inversa de una matriz cuadrada **A** se denota como $\mathbf{A}^{-1}$ y es una matriz tal que $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{I}$.
- El cálculo de la inversa de una matriz requiere varias etapas:

1- En primer lugar es necesario el cálculo del determinante de la matriz.

- El determinante de una matriz **A**, es un escalar que resulta de obtener todos los productos posibles de una matriz de acuerdo a una serie de restricciones, siendo denotado como $|A|$.
- El determinante de una matriz de tamaño 2 resulta sencillamente de multiplicar los dos elementos de la diagonal principal y restarle el producto de los elementos de la diagonal secundaria.
- El determinante de matrices de dimensiones superiores puede obtenerse desarrollando una fila o columna de la matriz original por menores complementarios.
- Dada una matriz cuadrada **A** de orden n, se denomina menor complementario a cada una de las matrices de orden (n - 1) que se obtienen al suprimir la fila y la columna donde se encuentra un elemento (a_{ij}) de la matriz original.
- Para una matriz cuadrada **A** de orden n se llama adjunto A_{ij} del elemento a_{ij} al valor del menor complementario de dicho elemento multiplicado por (-1) elevado a i más j.
- El valor de un determinante puede desarrollarse a partir de los adjuntos de los elementos de su matriz correspondiente. Así, dada una matriz **A**, el valor de su determinante $|A|$ es igual a la suma de los productos de cada uno de los elementos de una de sus filas o sus columnas por los adjuntos respectivos de dichos elementos. Así el determinante de una matriz puede desarrollarse en función de otros de orden inferior. Puede desarrollarse por filas o columnas. Desarrollado por filas sería:

$$|\mathbf{A}| = a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13} + ... + a_{1n}A_{1n}$$

- Una matriz con determinante nulo no posee inversa.

2- En segundo lugar, debe construirse la matriz de adjuntos **M**.

3- Se obtiene la transpuesta de la matriz **M**.

4- Se dividen todos los elementos de M por el determinante de la matriz.

- La inversa de una matriz diagonal es la única inversa sencilla de realizar. Para ello basta con invertir uno a uno todos los elementos de la diagonal.

La solución a las ecuaciones $(\mathbf{X'X})\hat{\mathbf{b}} = \mathbf{X'y}$ se obtiene premultiplicando a ambos lados de la igualdad por la inversa de la matriz de coeficientes:

$$(\mathbf{X'X})^{-1}(\mathbf{X'X})\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\hat{\mathbf{b}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$$

Y en nuestro caso:

$$\begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix}$$

Al afrontar la inversa de la matriz de coeficientes observamos que no se puede obtener dado que su determinante es nulo.

DEPENDENCIA LINEAL E INVERSAS GENERALIZADAS

Una matriz que posee determinante nulo no tiene inversa. Se dice de la matriz que es de rango no completo. Esta situación se da cuando existe una dependencia lineal entre filas o entre columnas de la matriz, o lo que es lo mismo, que una de ellas se puede obtener como combinación lineal de las otras (se puede obtener una de ellas sumando las otras multiplicadas por los coeficientes apropiados). Cuando se trata de una matriz de coeficientes, el sistema se llama también de rango no completo y no tiene solución única, sino que tiene infinitas soluciones, y es la consecuencia de tener más incógnitas que ecuaciones. Pongamos como ejemplo el siguiente sistema de ecuaciones:

$$x + y = 2$$
$$2x + 2y = 4$$

Este sistema de ecuaciones se puede representar en álgebra matricial de la siguiente manera:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

La matriz de coeficientes tiene determinante cero y el sistema no tiene solución única. La realidad es que no tenemos dos ecuaciones y dos incógnitas sino que sólo tenemos una ecuación, la segunda es la primera multiplicada por 2. Y efectivamente, en la matriz de coeficientes hay una dependencia lineal, la segunda fila es la primera multiplicada por 2, o, visto por columnas, la segunda es la primera multiplicada por 1, ambas son iguales.

Este sistema no tiene solución única, tiene infinitas soluciones. Por ejemplo, son soluciones de este sistema las siguientes: ($x = 1$ e $y = 1$), ($x = 3$ e $y = -1$), ($x = 0$ e $y = 2$), o ($x = -5$ e $y = 3$).

- Una inversa generalizada $\mathbf{A}^-$ de una matriz de rango no completo **A** es aquélla que cumple:

$$\mathbf{A^- A A^- = A^-}$$
$$\mathbf{A A^- A = A}$$
$$\mathbf{A^- A \neq I}$$
$$\mathbf{A A^- \neq I}$$

Obsérvese que las dos primeras condiciones las cumplen también las matrices inversas de matrices de rango completo, pero no las dos últimas.

Como existen infinitas soluciones a los sistemas de rango no completo, también existen infinitas inversas generalizadas. Una forma sencilla de obtener una inversa generalizada consiste en igualar a cero una de las incógnitas. El proceso es como sigue:

1. Se escoge una de las incógnitas implicadas en la combinación lineal y se elimina la fila y la columna correspondiente. A esta incógnita se le asigna valor nulo.
2. Una vez eliminada la ecuación de esta incógnita se ha eliminado la dependencia lineal, así que el resto se puede invertir. Se invierte entonces la matriz reducida.
3. Se orla de ceros las posiciones donde estaban los elementos de la ecuación que se eliminó. La matriz resultante es una inversa generalizada de la matriz original.

La imposibilidad de resolver el sistema de ecuaciones no es una casualidad. Este hecho nos informa de la ilógica de intentar obtener al mismo tiempo estimaciones para la media y para cada una de las dos clases de sexo. Así, un modelo que incluye el efecto sexo no permite estimar la media general ya que el estimador va a depender del número de individuos de cada sexo, el cuál no tiene por qué ser en la misma proporción en nuestra muestra que en la población. Asimismo, no parece tener sentido la incógnita sexo varón ya que este nivel se define por el hecho de existir mujeres. En cambio, sí tendría sentido saber cuánto calzan más los hombres que las mujeres, o cuánto produce más un animal de un ganadero que un animal de otro.

Recurrimos entonces a obtener una inversa generalizada de la matriz de coeficientes para obtener una de las infinitas posibles soluciones. Escogemos como ecuación a eliminar la correspondiente a la media, dado que al eliminarla, la matriz reducida que nos queda es una matriz diagonal, y por tanto muy sencilla de invertir. La solución que obtendremos no será única y la denotaremos con un circulito en superíndice ($\hat{\mathbf{b}}^{\mathbf{o}}$):

$$\hat{\mathbf{b}}^{o} = \left(\mathbf{X'X}\right)^{-} \mathbf{X'y}$$

$$\hat{\mathbf{b}}^{o} = \begin{bmatrix} \hat{\mu}^{o} \\ \hat{s}_1^{o} \\ \hat{s}_2^{o} \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}^{-} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 \\ 86/2 \\ 111/3 \end{bmatrix} = \begin{bmatrix} 0 \\ 43 \\ 37 \end{bmatrix}$$

El valor obtenido para la media general es cero porque así lo decidimos para poder resolver el sistema. Las soluciones para los otros dos efectos son las correspondientes en este caso. Obsérvese que los valores obtenidos para estos dos efectos se corresponden con las medias, precisamente porque termina dividiéndose la suma de los datos dentro de cada nivel del efecto fijo por el número de datos que hay en ese nivel. Los resultados parecen razonables pero se ha empleado una herramienta demasiado compleja para terminar calculando medias aritméticas sencillas. Esto es así porque el modelo es sencillo. Sin embargo, la herramienta no difiere mucho cuando el modelo se complica, como veremos después.

Obsérvese que de haber considerado un segundo efecto fijo discontinuo como la provincia definida anteriormente, la matriz X'X sería la siguiente:

$$\mathbf{X'X} = \begin{bmatrix} 5 & 2 & 3 & 1 & 2 & 2 \\ 2 & 2 & 0 & 0 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{bmatrix}$$

En esta matriz hay dos dependencias lineales. La primera coincide con la vista en el ejemplo anterior al ser igual la suma de las filas correspondientes a los dos niveles del efecto sexo (la segunda y la

tercera), a la primera fila. Pero también las tres filas correspondientes a las tres provincias (las tres últimas), suman lo mismo que la primera. Por tanto, en este caso, sería necesario hacer cero una incógnita más.

IV-3.1.1. Las ecuaciones del modelo fijo con covariables

Cuando el modelo incluye variables fijas continuas, la matriz de incidencia de los efectos fijos incluirá valores diferentes de ceros y unos. En el ejemplo desarrollado más atrás que ajustaba la edad en meses, dicha matriz era la siguiente:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 4 \\ 1 & 0 & 1 & -1 \\ 1 & 1 & 0 & 11 \\ 1 & 0 & 1 & -5 \\ 1 & 1 & 0 & -9 \end{bmatrix}$$

Cuando el modelo incluye covariables, en lugar del total de observaciones dentro de los niveles de efectos fijos, la matriz de coeficientes incluirá las sumas de las covariables dentro de dichos efectos fijos. Obviamente, el valor que se incluirá en el cruce de la incógnita media con la covariable será cero, ya que ésta será la suma de la covariable para todos los datos si en el modelo se incluye la covariable desviada respecto de su media. En el cruce de la covariable consigo misma se incluirá la suma de cuadrados de la misma:

$$\sum (x_i - \bar{x})^2 = 4^2 + (-1)^2 + 11^2 + (-5)^2 + (-9)^2 = 244$$

En cuanto al lado derecho, la posición correspondiente a la covariable se rellenará con la suma del producto de la variable de trabajo con la covariable para todos los individuos, en este caso será:

$$\sum (x_i - \bar{x}) y_i = 4x37 + (-1)x38 + 11x42 + (-5)x36 + (-9)x44 = -4$$

X'X =		μ	s_1	s_1	b_e
	μ	5	2	3	0
	s_1	2	2	0	2
	s_1	3	0	3	-2
	b_e	0	2	-2	244

X'y =		
	μ	197
	s_1	86
	s_1	111
	b_e	-4

Eliminado la ecuación correspondiente a la media como en el caso anterior las soluciones del modelo serían:

$$\hat{\mathbf{b}}^{\mathbf{o}} = \begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \\ \hat{b}_e \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 & 0 \\ 2 & 2 & 0 & 2 \\ 3 & 0 & 3 & -2 \\ 0 & 2 & -2 & 244 \end{bmatrix}^{-} \begin{bmatrix} 197 \\ 86 \\ 111 \\ -4 \end{bmatrix} = \begin{bmatrix} 0 \\ 43,0665 \\ 36,9557 \\ -0,0665 \end{bmatrix}$$

La interpretación del resultado es sencilla. Por cada mes de edad el individuo tenderá a tener un número de pie de 0,0665 números menos. Así por ejemplo, un individuo con 10 meses más tendrá un número de pie 0,665 números menos.

Una cuestión más conviene resaltar aquí, y es que las soluciones para los efectos fijos continuos son únicas. En otras palabras, y como introducción al punto que se desarrolla a continuación, los coeficientes de regresión son funciones estimables.

IV-3.2. Funciones estimables

La solución obtenida anteriormente, aunque coherente, no deja de ser únicamente una de las infinitas posibles.

Las soluciones que se obtienen cuando se eliminan las otras incógnitas son las dos siguientes:

$$\hat{\mathbf{b}}^{\mathbf{o}} = \begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}^{-} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} =$$

$$= \begin{bmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 5/6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} = \begin{bmatrix} 37 \\ 6 \\ 0 \end{bmatrix}$$

$$\hat{\mathbf{b}}^{\mathbf{o}} = \begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}^{-} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} =$$

$$= \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 5/6 \end{bmatrix} \begin{bmatrix} 197 \\ 86 \\ 111 \end{bmatrix} = \begin{bmatrix} 43 \\ 0 \\ -6 \end{bmatrix}$$

Si observamos las tres soluciones de forma conjunta, efectivamente son completamente diferentes, pero también se intuye que existen parecidos entre ellas:

$$\begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \end{bmatrix} = \begin{bmatrix} 0 \\ 43 \\ 37 \end{bmatrix} \qquad \begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \end{bmatrix} = \begin{bmatrix} 37 \\ 6 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} \hat{\mu}^{\mathbf{o}} \\ \hat{s}_1^{\mathbf{o}} \\ \hat{s}_2^{\mathbf{o}} \end{bmatrix} = \begin{bmatrix} 43 \\ 0 \\ -6 \end{bmatrix}$$

Se define así Función Estimable como una combinación lineal de las incógnitas cuyo valor es único independientemente de la inversa generalizada que se utilice para resolver el modelo.

Siempre son funciones estimables:

a) Las diferencias entre niveles del mismo efecto fijo. Por ejemplo, la diferencia en el número de zapatos entre varones y mujeres ($s_1 - s_2$). También cuánto calzan más las personas de Cataluña que las de Madrid, o cuánto produce más un animal nacido en el 2009 que uno nacido en el

2008, o cuánto pesa más un animal de Pepe que otro de Juan, o cuánta leche da más un animal que pare en primavera que uno que lo hace en verano. Como vemos, aunque la solución del sistema no es única para cada incógnita, sí que lo son las soluciones de las funciones estimables, las cuáles, por otro lado, son en realidad las que nos interesan.

b) La media más un nivel de cada efecto fijo. Así, en el modelo que no incluye la provincia, la media más el nivel 2 del efecto sexo ($\mu + s_2$) sería una función estimable, es decir, sería estimable la media de los números de zapatos de las chicas. Sin embargo, esta combinación lineal no sería función estimable en un modelo que incluyese el efecto provincia de origen. En este modelo sí que sería estimable por ejemplo la media de los números de zapatos de los varones que nacen en Santander ($\mu + s_1 + p_3$). Obsérvese que estas combinaciones lineales son las contempladas en cada fila de **Xb**.

PRUEBA DE ESTIMABILIDAD DE UNA COMBINACIÓN LINEAL

Para saber si una combinación lineal es una función estimable hay que buscar un vector fila **k'**, de forma que **k'b** exprese la combinación lineal de la que deseo conocer su estimabilidad. Los elementos del vector **k'** se corresponden con los coeficientes de las incógnitas en la función estimable. Así, por ejemplo, para $s_1 - s_2$ el vector **k'** sería [0 1 -1], para $\mu + s_2$ el vector **k'** sería [1 0 1], y para la media (μ) el vector **k'** sería [1 0 0]. Obsérvese cómo **k'b** corresponde en los tres casos con dichas combinaciones lineales:

$$\mathbf{k'b} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \end{bmatrix} = s_1 - s_2$$

$$\mathbf{k'b} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \end{bmatrix} = \mu + s_2$$

$$\mathbf{k'b} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \end{bmatrix} = \mu$$

Sabemos a la vista de las tres soluciones posibles anteriores que las dos primeras combinaciones lineales son funciones estimables, mientras que la media general no lo es. Una vez definido **k'**, la combinación lineal definida por **k'b** será función estimable si se cumple la siguiente expresión:

$$\mathbf{k'(X'X)^{-}X'X = k'}$$

Efectivamente, por ejemplo $s_1 - s_2$ es función estimable ya que se cumple:

$$\begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$$

Se puede probar la estimabilidad de más de una combinación lineal incorporando cada vector k' como una fila de una matriz **K'**. Las filas de **K'** que permanezcan con el mismo valor en $\mathbf{K'(X'X)^{-}X'X}$ corresponderán a funciones estimables, y el resto no lo serán:

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 5 & 2 & 3 \\ 2 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

IV-3.2.1. Errores de estimación

Al contrario que los verdaderos valores de los efectos fijos, sus estimadores no poseen valores únicos sino que presentan variabilidad que puede ser medida por su varianza que representa de algún modo la medida de su error. Por tanto, cuanto menor es su varianza menor es su error. Así por ejemplo, el estimador de la media poblacional, $\hat{\mu}$, la media muestral ($\bar{x}$), presenta como varianza:

$$\sigma_{\bar{x}}^{\ 2} = \frac{\sigma^2}{n}$$

Obsérvese que cuanto más elevado es el número de datos con el que se estima la media poblacional, menor es el error que posee la media muestral como estimador de la media poblacional.

Calcularemos entonces la varianza del estimador:

$$Var\left(\hat{\mathbf{b}}^{\mathbf{o}}\right) = Var[\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y}] = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'Var(\mathbf{y})\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}$$

En esta expresión $\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ es un producto de matrices de constantes por lo que la varianza de $\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y}$ debe desarrollarse como una forma cuadrática, es decir, un producto con la parte constante en primer lugar, luego la varianza de la variable y luego la constante transpuesta. Para el segundo paso recuérdese que la transpuesta de un producto es el producto de las transpuestas en orden inverso, y que la matriz **X'X** es una matriz simétrica por lo que su transpuesta es igual a sí misma. Finalmente, a continuación se tendrá en cuenta lo estipulado en la segunda etapa de la definición del modelo fijo en la que $\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{I}\sigma_e^2$**:**

$$Var\left(\hat{\mathbf{b}}^{\mathbf{o}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-} = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{I}\sigma_e^2\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-} =$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\sigma_e^2$$

Y a continuación aprovecharemos la propiedad de las inversas generalizadas según la cual: $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$. Dado que esta propiedad también es válida para una inversa, el desarrollo es válido también para modelos de rango no completo:

$$Var\left(\hat{\mathbf{b}}^{\mathbf{o}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-}\sigma_e^2 \qquad Var\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma_e^2$$

Por tanto, los errores de los estimadores se encuentran siempre en el producto de la inversa de la matriz de coeficientes por la varianza residual del modelo. Dado que en el modelo fijo toda la varianza fenotípica es varianza residual $\sigma^2 = \sigma_e^2$, utilizando la

inversa generalizada con la que se obtenían las medias de los números de zapatos de varones y mujeres tenemos que:

$$Var\left(\hat{\mathbf{b}}^{\mathbf{o}}\right) = Var\begin{bmatrix}\hat{\mu}^{\mathbf{o}}\\ \hat{s}_1^{\mathbf{o}}\\ \hat{s}_2^{\mathbf{o}}\end{bmatrix} = \begin{bmatrix}\sigma^2_{\hat{\mu}^{\mathbf{o}}} & \sigma_{\hat{\mu}^{\mathbf{o}},\hat{s}_1^{\mathbf{o}}} & \sigma_{\hat{\mu}^{\mathbf{o}},\hat{s}_2^{\mathbf{o}}}\\ \sigma_{\hat{s}_1^{\mathbf{o}},\hat{\mu}^{\mathbf{o}}} & \sigma^2_{\hat{s}_1^{\mathbf{o}}} & \sigma_{\hat{s}_1^{\mathbf{o}},\hat{s}_2^{\mathbf{o}}}\\ \sigma_{\hat{s}_2^{\mathbf{o}},\hat{\mu}^{\mathbf{o}}} & \sigma_{\hat{s}_2^{\mathbf{o}},\hat{s}_1^{\mathbf{o}}} & \sigma^2_{\hat{s}_2^{\mathbf{o}}}\end{bmatrix} =$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-}\sigma_e^2 = \begin{bmatrix}0 & 0 & 0\\ 0 & \frac{1}{2} & 0\\ 0 & 0 & \frac{1}{3}\end{bmatrix}\sigma^2 = \begin{bmatrix}0 & 0 & 0\\ 0 & \frac{\sigma^2}{2} & 0\\ 0 & 0 & \frac{\sigma^2}{3}\end{bmatrix}$$

Los errores de los estimadores se encuentran en la diagonal de la matriz de varianzas y covarianzas de los estimadores. En este caso sólo se obtienen los errores de las dos incógnitas del efecto sexo ya que la media no se ha estimado. Las soluciones que se obtuvieron para estas dos incógnitas fueron la media de los números de zapatos de los varones y de las mujeres, y las varianzas de cada una de estas medias han resultado ser la varianza de la variable partida por el número de datos que se emplearon en el cálculo de cada media. El resultado es por tanto el esperado, aunque puede resultar desalentador el esfuerzo empleado en ello. Y así es, con un modelo sencillo no se precisa una herramienta tan potente. Sin embargo, la herramienta es prácticamente la misma cuando el modelo se complica. Se puede dar entonces la solución de las funciones estimables con cada uno de sus errores que se corresponderán con su desviación típica, conocida normalmente como error estándar de la media y que es igual a la raíz cuadrada de su varianza:

$$\hat{\mu} + \hat{s}_1 = 43 \pm \frac{\sigma}{\sqrt{2}}$$

$$\hat{\mu} + \hat{s}_2 = 37 \pm \frac{\sigma}{\sqrt{3}}$$

IV-3.3. Otros métodos de estimación

El modelo fijo anterior es el más sencillo desde el punto de vista teórico. Imaginemos que disponemos de varias medias de números de zapatos de varones y mujeres y se desea emplear todas ellas para obtener una media global, pero dando mayor peso a las que son más fiables. Cada una de las medias se comporta en el nuevo modelo como un dato del modelo anterior, pero, dado que la fiabilidad depende de la varianza del estimador, ésta es distinta para cada una de ellas, por lo que no puede ser asumida homogeneidad de varianzas.

Puede ocurrir igualmente que exista un parentesco conocido entre individuos que participan de más de una media, incluso que algunos de ellos participen directamente en más de una media, por lo que existiría una covarianza entre datos y tampoco podríamos asumir independencia entre residuos. Esta covarianza podría ser calculada para cada dos datos, es decir, dispondríamos de una matriz **V**, pero, al no existir homogeneidad de varianza residual ni independencia entre residuos, esta matriz **V** no es de la estructura definida previamente, y parte de la definición del modelo debe ser modificada. En concreto, ahora:

$$\text{Var}(\mathbf{y}) = \mathbf{V} = \text{Var}(\mathbf{e}) = \mathbf{R} \neq \mathbf{I}\sigma_e^2$$

En esta situación no puede utilizarse el método de mínimos cuadrados ordinarios que daba el mismo peso a cada residuo e ignoraba la covarianza entre observaciones, y debemos recurrir a métodos alternativos.

IV-3.3.1. Estimación por el método de Mínimos Cuadrados Generalizados (MCG)

Como vimos anteriormente, el vector de residuos puede despejarse del modelo como

$$\mathbf{e} = (\mathbf{y} - \mathbf{Xb})$$

La suma de cuadrados definida anteriormente como **e'e** asume que se dará el mismo peso al error de cada observación en la función. Ahora deseamos dar distinto peso a cada observación en función de su fiabilidad. Dado que el error es medido por la varianza del estimador, una ponderación adecuada para cada dato sería el inverso de su varianza definida en **V**, de manera que la nueva función F a minimizar será en este caso:

$$F = \mathbf{e'}\,\mathbf{V}^{-1}\,\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'\,\mathbf{V}^{-1}\,(\mathbf{y} - \mathbf{Xb})$$

La ponderación por el inverso de la varianza tiene en cuenta también la información repetida que aparece en las diferentes observaciones, información que es medida por la covarianza.

La minimización de esta función no es diferente de la descrita para el método de MCO, y lleva a la siguiente expresión:

$$\left(\mathbf{X'V^{-1}X}\right)\hat{\mathbf{b}} = \mathbf{X'V^{-1}y}$$

Esta expresión se resuelve, paralelamente al método anterior, dependiendo de la singularidad (rango completo o no) de la matriz de coeficientes mediante una de las siguientes expresiones:

$$\hat{\mathbf{b}} = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}y} \qquad \hat{\mathbf{b}}^{\mathbf{o}} = \left(\mathbf{X'V^{-1}X}\right)^{-}\mathbf{X'V^{-1}y}$$

No se va a desarrollar en este texto un ejemplo numérico de este método pero conviene resaltar algunos detalles:

- En el modelo fijo las matrices **V** y **R** son la misma matriz por lo que estas expresiones podrían escribirse igualmente como $\hat{\mathbf{b}} = \left(\mathbf{X'R^{-1}X}\right)^{-1}\mathbf{X'R^{-1}y}$ y $\hat{\mathbf{b}}^{\mathbf{o}} = \left(\mathbf{X'R^{-1}X}\right)^{-}\mathbf{X'R^{-1}y}$. Esta ecuación así expresada será más reconocible posteriormente cuando se vean los métodos de predicción.

- Obsérvese que en la definición presentada originalmente para el modelo fijo, en la que $\mathbf{V} = \mathbf{I}\sigma_e^2, \mathbf{V}^{-1} = \mathbf{I}\frac{1}{\sigma_e^2}$, y por tanto

$\left(\mathbf{X'V^{-1}X}\right)\hat{\mathbf{b}} = \mathbf{X'V^{-1}y}$ se reduce a $\left(\mathbf{X'X}\right)\hat{\mathbf{b}} = \mathbf{X'y}$. Es decir, el método MCG se transforma automáticamente en MCO por lo que ambos métodos son equivalentes con esa definición.

- La medida del error de los estimadores se lleva a cabo igualmente a partir de su varianza. El desarrollo algebraico es muy sencillo y similar al anterior y se propone como ejercicio al lector. El resultado lleva a $Var\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X'V^{-1}X}\right)^{-1}$. Obsérvese la equivalencia con $Var\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X'X}\right)^{-1}\sigma_e^2$ cuando $\mathbf{V} = \mathbf{I}\sigma_e^2$.

IV-3.3.2. Estimación por el método de Máxima Verosimilitud (MV)

Al contrario que los métodos minimocuadráticos vistos hasta aquí, la metodología de estimación por máxima verosimilitud precisa asumir una distribución de los datos basada en los parámetros que se desean estimar.

Bajo esta lógica los datos que se disponen se asumen generados a partir de esa distribución. En nuestro caso se asumirá una distribución normal para los datos, lo que se representa como $y_i : N(\mu, \sigma_e^2)$, y en álgebra matricial como $\mathbf{y} : N(\mathbf{Xb}, \mathbf{V})$.

La función de verosimilitud de los parámetros a estimar, dados los datos de que se disponen, tiene la misma expresión que su función de densidad, la función que describe los datos en función de los parámetros. El estimador por máxima verosimilitud consiste en localizar su máximo.

DERIVACIÓN DEL ESTIMADOR MAXIMOVEROSÍMIL DEL MODELO FIJO

La función de densidad de una variable que se distribuye $y_i : N(\mu, \sigma_e^2)$ es la siguiente:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

Si la variable se expresa en álgebra matricial, $\mathbf{y} : N(\mathbf{Xb}, \mathbf{V})$, esta expresión puede ponerse igualmente en álgebra matricial:

$$f(\mathbf{y}) = \frac{1}{\sqrt{2\pi}} \mathbf{V}^{-½} e^{-\frac{1}{2}(\mathbf{y-Xb})'\mathbf{V^{-1}}(\mathbf{y-Xb})}$$

La función de verosimilitud de los parámetros (**b**), dados los datos (**y**), es entonces:

$$L(\mathbf{b} \mid \mathbf{y}) = \frac{1}{\sqrt{2\pi}} \mathbf{V}^{-½} e^{-\frac{1}{2}(\mathbf{y-Xb})'\mathbf{V^{-1}}(\mathbf{y-Xb})}$$

El estimador de b por máxima verosimilitud será el que maximice esta expresión. En ella, el coeficiente $\frac{1}{\sqrt{2\pi}} \mathbf{V}^{-½}$ no depende de b por lo que es constante en relación a la incógnita, y el resto de la expresión se maximiza también si se encuentra el máximo de su exponente:

$$máx\left(\frac{1}{\sqrt{2\pi}} \mathbf{V}^{-½} e^{-\frac{1}{2}(\mathbf{y-Xb})'\mathbf{V^{-1}}(\mathbf{y-Xb})} \right) = máx\left(e^{-\frac{1}{2}(\mathbf{y-Xb})'\mathbf{V^{-1}}(\mathbf{y-Xb})} \right) =$$

$$= máx\left(-\frac{1}{2}(\mathbf{y-Xb})'\mathbf{V^{-1}}(\mathbf{y-Xb}) \right)$$

Obsérvese que maximizar esta expresión equivale a minimizar
(**y** - **Xb**)' **V-1** (**y** – **Xb**), lo que coincide con el criterio del método anterior por lo que llevará a la misma expresión:

$$\left(\mathbf{X'V^{-1}X}\right)\hat{\mathbf{b}} = \mathbf{X'V^{-1}y}$$

Los estimadores maximoverosímiles de los efectos fijos en un modelo fijo se obtienen mediante $\hat{\mathbf{b}} = \left(\mathbf{X'V^{-1}X}\right)^{-1} \mathbf{X'V^{-1}y}$, y por

tanto coincide con el método de MCG. Si además la estructura de la matriz de varianzas y covarianzas de los datos es $\mathbf{V} = \mathbf{I}\sigma_e^2$, entonces también coincide con el método de MCO. La diferencia radica en que el estimador MV precisa definir la variable como distribuida normalmente, cosa que los métodos minimocuadráticos no. Nótese que MV y MCG coinciden en la expresión para estimar efectos fijos en un modelo fijo, pero no tiene por qué ser así para obtener soluciones de parámetros que no sean efectos fijos, ni incluso para estimar efectos fijos con otros modelos.

IV-3.3.3. Estimación por el método BLUE

Las siglas del método BLUE se corresponden con:

- **B** (*Best*). El mejor estimador estadísticamente es aquél que posee la menor varianza.

- **L** (*Linear*). El estimador BLUE ha de ser una combinación lineal de los datos.

- **U** (*Unbiased*). El estimador BLUE ha de ser insesgado, es decir, su valor esperado debe coincidir con el verdadero valor del parámetro.

- **E** (*Estimator*). Estimador.

La deducción algebraica de las ecuaciones del BLUE será pospuesta para ser tratada conjuntamente con la predicción de los efectos aleatorios mediante el BLUP. Se basa en minimizar una expresión que combina las propiedades de insesgado y mínima varianza de forma conjunta. El resultado lleva nuevamente a:

$$\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\hat{\mathbf{b}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

IV-3.3.4. Relación entre los estimadores

Por tanto, el estimador BLUE de los efectos fijos en un modelo fijo coincide con el estimador por MV y con el estimador por MCG. Si además la estructura de la matriz de varianzas y covarianzas de los datos es sencilla ($\mathbf{V} = \mathbf{I}\sigma_e^2$), entonces también coincide con el método MCO.

Se ha mostrado que el estimador por MCO proporciona como estimaciones simplemente las medias de los datos por niveles de efectos fijos, y que los otros métodos tienen en cuenta además la distinta varianza de cada elemento y las relaciones entre ellos por lo que pueden considerarse medias ponderadas por el inverso de la varianza que además aprovechan la información que proporcionan las relaciones entre ellos.

Aunque para este modelo resulta excesivo el empleo de los modelos lineales, se ha presentado así una herramienta que será empleada posteriormente en modelos con mayor grado de complejidad.

CONCEPTOS CLAVE

- ¿Cuál es el fundamento del procedimiento ajuste por mínimos cuadrados como método de estimación?
- ¿Qué dimensiones tiene la matriz de coeficientes de las ecuaciones de resolución del modelo fijo? ¿Más o menos que la matriz de incidencia de los efectos fijos?
- ¿Es preciso realizar operaciones de multiplicación de matrices para construir las ecuaciones de resolución del modelo fijo?
- ¿Un modelo sin efectos fijos continuos presenta problemas de estimabilidad?
- ¿Para qué se utiliza una inversa generalizada? ¿Son únicas las soluciones obtenidas cuando es necesario hacer uso de una?
- ¿Qué son funciones estimables?
- ¿La diferencia entre niveles de un mismo efecto fijo son funciones estimables?
- ¿La media poblacional es una función estimable?
- ¿Dónde podemos encontrar los errores de estimación de los efectos fijos?
- ¿En qué se parecen los procedimientos de estimación de efectos fijos por mínimos cuadrados generalizados y por máxima verosimilitud cuando se analiza un modelo fijo?
- ¿Qué propiedades estadísticas tiene el BLUE?
- ¿Cuándo son equivalentes el procedimiento ajuste por mínimos cuadrados y por máxima verosimilitud?

QUINTA PARTE

PREDICCIÓN DEL MÉRITO GENÉTICO

RESUMEN

En este corto capítulo se trata desde un punto de vista general teórico la metodología de la predicción, es decir, la resolución de modelos que incluyen efectos aleatorios, como introducción a los dos capítulos siguientes. Se establece de forma general en qué consiste la predicción recalcando que incluye también estimación de efectos fijos. Se definen las propiedades estadísticas de los tres métodos teóricos de predicción de efectos aleatorios y se comenta que sólo dos de ellos, el BLP o índices de selección, y el BLUP, permiten obtener soluciones analíticas del modelo.

V-1. Predicción del mérito genético

Así como la estimación de los efectos fijos asume la existencia de un único valor verdadero que se desea conocer, aquélla no puede aplicarse para los efectos aleatorios, ya que no poseen un único valor desconocido sino que se asume la existencia de una distribución para los mismos. En esta situación no tendría sentido la estimación ya que se precisaría obtener valores para los infinitos datos de esa distribución teórica. Sin embargo, sí tiene sentido conocer el comportamiento del valor genético de un animal cuando vaya a ser utilizado como reproductor, lo que en la teoría supone predecir el valor de un elemento extraído de la distribución de valores genéticos aditivos antes de medirlo. Es por ello que para efectos aleatorios se habla de predicción en lugar de estimación como para los efectos fijos.

V-2. Métodos de predicción del mérito genético

Dado que el objetivo de la predicción es precisamente conocer de antemano el comportamiento de un individuo por su valor genético, y éste ha de ser medido en su rendimiento, hay que tener en cuenta que el valor genético se expresará en determinadas condiciones no genéticas que modifican su valor, como la ganadería, el sexo, o cualquier otro efecto de los contemplados dentro de la parte fija del modelo. En otras palabras, la predicción no puede entenderse de forma independiente a la estimación.

Así pues nuestro objetivo será conocer el valor esperado de una combinación lineal de efectos fijos y aleatorios. Definiremos dos matrices **K'** y **M'** que expresen esa combinación lineal de manera que el objetivo será predecir el valor esperado de dicha combinación lineal:

$$E\,(\mathbf{K'b} + \mathbf{M'u})$$

Obviamente en esta expresión **K'b** representa una combinación lineal que ha de ser función estimable.

Plantearemos a continuación secuencialmente los posibles predictores que desde un punto de vista teórico podemos utilizar incorporando una a una las distintas propiedades en los mismos.

V-2.1. El mejor predictor (Best Predictor o BP)

La única propiedad presente en el BP es la de mínimo error cuadrático medio y es por tanto equivalente al método de mínimos cuadrados para los efectos fijos. Sin embargo, los efectos aleatorios poseen varianza de entrada por definición por lo que ha de minimizarse la varianza del error de predicción. Así pues, la varianza de la diferencia entre el predictor y el verdadero valor se convierte en la expresión a minimizar.

El BP requiere que la distribución de la variable aleatoria, junto con su media y varianza, sean conocidas. Entonces, el hecho de utilizar el vector de datos **y** que depende de los efectos fijos y aleatorios bastaría para cumplir la propiedad de mínima varianza del error de predicción. Así el BP sería la media condicional del predictor dado el vector de datos. En otras palabras, será el valor esperado de esta combinación lineal dados unos datos, es decir, utilizando como fuente de información los datos de que se dispone:

$$E\,(\mathbf{K'b} + \mathbf{M'u} \mid \mathbf{y})$$

Sin embargo esta propiedad no es suficiente para encontrar una expresión analítica que nos permita resolver el modelo. Así pues el BP por sí solo no conduce a ningún resultado.

V-2.2. El mejor predictor lineal (Best Linear Predictor o BLP)

Una propiedad adicional de un predictor que permite llegar a una expresión analítica es la linealidad en las observaciones, es decir, que forzosamente sea necesario obtenerlo como combinación lineal de los datos. De hecho cuando se exige esta propiedad al predictor, ni siquiera la forma de la distribución es necesaria sino que basta conocer sus momentos de primer y segundo orden, es decir, su esperanza y su varianza. De acuerdo con la definición de nuestro modelo, estos son $E(\mathbf{y}) = \mathbf{Xb}$ y $Var(\mathbf{y}) = \mathbf{V}$.

ECUACIONES DE LA METODOLOGÍA BLP

Nos encontramos ante la necesidad de predecir una variable (**K'b** + **M'u**) a partir de otra (**y**), y las propiedades que ha de tener nuestro predictor son la de menor error cuadrático medio y su linealidad. La metodología estadística lineal con propiedades minimocuadráticas que se emplea para predecir una variable en función de otra se conoce como regresión lineal. Una recta de regresión que permite predecir la variable *y* a partir de la variable *x* es $\hat{y}_i = \beta_0 + \beta_1 x_i$. En esta ecuación β_0 es la ordenada en el origen y β_1 es el coeficiente de regresión. Se puede además prescindir de la ordenada en el origen si tanto la variable dependiente, la *y*, como la variable regresora, la *x*, se introducen en la recta como valor desviado con respecto a sus medias:

$$(\hat{y}_i - \overline{y}) = \beta_1 (x_i - \overline{x})$$

Por otro lado, el coeficiente de regresión de *y* sobre *x* se define como $\beta_1 = \frac{\sigma_{xy}}{\sigma^2_x}$, de manera que esta ecuación se puede escribir así:

$$(\hat{y}_i - \overline{y}) = \frac{\sigma_{yx}}{\sigma^2_x}(x_i - \overline{x})$$

Podemos trasladar esta ecuación en álgebra de escalares a nuestra definición del modelo en álgebra de matrices. Primeramente asumiremos que se conocen los verdaderos valores de nuestro vector de efectos fijos **b**. En esta situación, *E* (**K'b**) es exactamente **K'b**. La combinación lineal de los efectos aleatorios en esta situación no tiene tampoco sentido, sino que es el propio vector de efectos aleatorios **u** el que se desea predecir. Por otro lado, esta variable dependiente que se desea predecir, **u**, se define con media cero, por lo que, en nuestro caso $(\hat{y}_i - \overline{y})$ debe ser escrito sencillamente como **û**. La covarianza entre el vector de la variable que se desea predecir (**K'b** + **M'u**) y nuestro vector de datos (**y**)

será una matriz que podemos representar como **C'**, la varianza del vector de datos ya fue definida en el modelo como **V** (y aparecerá invertida en la expresión), y el vector de datos **y** aparecerá desviado con respecto a las medias de sus grupos supuestamente conocidas (**Xb** según la definición del propio modelo). Sustituyendo en la ecuación de la regresión anterior se obtiene:

$$E(\mathbf{K'b} + \mathbf{M'u}) = \mathbf{K'b} + \mathbf{C'V^{-1}(y - Xb)}$$

Debe quedar patente que un índice de selección no es otra cosa que una recta de regresión de **u** sobre **y**.

El BLP es entonces:

$$E(\mathbf{K'b} + \mathbf{M'u}) = \mathbf{K'b} + \mathbf{C'V^{-1}(y - Xb)}$$

En esta expresión C' es en realidad la matriz que relaciona la combinación lineal que se desea obtener con los datos y se define en forma de matriz de covarianzas entre ambos:

$$\mathbf{C'} = Cov\,(\mathbf{K'b} + \mathbf{M'u}, \mathbf{y'})$$

MATRIZ DE COVARIANZAS ENTRE VECTORES

- Una matriz de covarianzas entre dos vectores **a** y **b** se representa por **C'** = *Cov* (**a**, **b**). En este texto, con el fin de representar el vector que define las columnas de la matriz de covarianzas, el segundo vector aparecerá en forma transpuesta de manera que lo representaremos como **C'** = *Cov* (**a**, **b'**) y será:

$$\mathbf{C'} = \begin{bmatrix} Cov(a_1,b_1) & Cov(a_1,b_2) & Cov(a_1,b_3) & ..... & Cov(a_1,b_m) \\ Cov(a_2,b_1) & Cov(a_2,b_2) & Cov(a_2,b_3) & ..... & Cov(a_2,b_m) \\ ... & ... & ... & ..... & ... \\ ... & ... & ... & ..... & ... \\ Cov(a_n,b_1) & Cov(a_n,b_2) & Cov(a_n,b_3) & ..... & Cov(a_n,b_m) \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_{a_1b_1} & \sigma_{a_1b_2} & \sigma_{a_1b_3} & .... & \sigma_{a_1b_m} \\ \sigma_{a_1b_1} & \sigma_{a_2b_2} & \sigma_{a_2b_3} & .... & \sigma_{a_2b_m} \\ ... & ... & ... & .... & ... \\ ... & ... & ... & .... & ... \\ \sigma_{a_nb_1} & \sigma_{a_nb_2} & \sigma_{a_1b_3} & .... & \sigma_{a_1b_m} \end{bmatrix}$$

Obsérvese que en el lado derecho de la solución aparece **b** en lugar de $\hat{\mathbf{b}}$. Por tanto, en el BLP se está dando por conocido el verdadero valor de los efectos fijos y, por tanto, tampoco este modelo tiene solución salvo que se provoque un sesgo asumiendo que $\mathbf{b} = \hat{\mathbf{b}}$.

V-2.3. El mejor predictor lineal insesgado (Best Linear Unbiased Predictor o BLUP)

EL sesgo existente en el BLP es corregido en el BLUP. Para ello simplemente se utilizan los efectos fijos que son estimados mediante BLUE. Así el BLUP se obtiene a partir de:

$$E(\mathbf{K'b} + \mathbf{M'u}) = \mathbf{K'}\hat{\mathbf{b}} + \mathbf{C'V^{-1}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

En esta expresión **C'** y **V** son las mismas matrices que en el BLP pero los efectos fijos deben ser estimados por BLUE:

$$\hat{\mathbf{b}} = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}y}$$

Queda para más adelante mostrar cómo incluir el estimador BLUE en la predicción BLUP en un modelo que incluye tanto efectos fijos como aleatorios.

CONCEPTOS CLAVE

- ¿Por qué es preciso contar con la estimación de efectos fijos en la predicción?
- ¿Qué propiedades estadísticas tienen el BP, el BLP y el BLUP?
- ¿Cuál es la expresión analítica del BP?
- ¿Qué relación tiene la predicción con la regresión?
- ¿Por qué se provoca un sesgo al utilizar el BLP?
- ¿Qué incorpora la metodología BLUP para corregir el sesgo del BLP?

SEXTA PARTE

LOS ÍNDICES DE SELECCIÓN

RESUMEN

Se detalla la metodología de valoración genética conocida como Índices de Selección que presenta propiedades BLP. Una primera parte del capítulo desarrolla simplificaciones de la expresión teórica propuesta en el capítulo precedente y establece las expresiones que permiten obtener las precisiones de las valoraciones genéticas en función de la información considerada en cada caso. El cuerpo más grande del capítulo propone distintos índices con complejidad creciente, mostrando cómo se debe proceder en cada caso y obteniendo información útil sobre los parámetros de los que depende la precisión en cada índice. Se completa el capítulo con la utilización de índices combinados a partir de varios caracteres que se miden y se utilizan como criterio de selección para mejorar de forma óptima una combinación de caracteres objetivo de selección que se ponderan por su valor económico para formar un agregado genético-económico.

VI-1. Valoración genética mediante BLP: Los índices de selección

El método BLP descrito más arriba desde un punto de vista teórico equivale a lo que en la práctica se conoce como Índice de Selección. La expresión teórica desarrollada más arriba para este método es la siguiente:

$$E(\mathbf{K'b} + \mathbf{M'u}) = \mathbf{K'b} + \mathbf{C'V^{-1}(y - Xb)}$$

En esta expresión los efectos fijos se asumen conocidos. Dado que necesitamos un valor para el vector **b** de efectos fijos, estos se obtienen previamente con el empleo de un modelo fijo. Así, en la ecuación del modelo mixto original se define un nuevo residuo **e*** = **Zu** + **e**. El nuevo modelo ignora las relaciones entre individuos y proporciona estimaciones a partir del estimador minimocuadrático:

$$\hat{\mathbf{b}} = \left(\mathbf{X'R^{-1}X}\right)^{-1}\mathbf{X'R^{-1}y}$$

Y a continuación se asume que estos estimadores son el verdadero valor:

$$\boxed{\mathbf{b} = \hat{\mathbf{b}}}$$

Se ha querido encuadrar este paso porque tiene varias implicaciones.

En primer lugar, el hecho de ignorar los efectos aleatorios para estimar los efectos fijos provoca un sesgo en la estimación de los efectos fijos. Así, si la distribución de los niveles de los efectos aleatorios no es homogénea, se asignará parte de los efectos aleatorios al resultado de los efectos fijos. Pensemos por ejemplo en un efecto fijo "ganadero". Imaginemos dos ganaderos con un nivel de manejo similar. Uno de ellos es un hombre preocupado

por tener animales de mayor producción por lo que tendrá producciones mayores que el otro ganadero si éste es un hombre despreocupado. La solución proporcionada por el modelo fijo dará valores superiores para el primero, y estas diferencias en el valor de las ganaderías serán asignadas a diferencias en el manejo cuando en realidad son diferencias en el valor genético medio.

Dado que conocemos los efectos fijos, la combinación lineal de efectos fijos y aleatorios deja de tener sentido. $E(\mathbf{K'b}) = \mathbf{K'b}$, y entonces el interés ya no se centra en ninguna combinación lineal de efectos aleatorios sino que sólo nos interesa el valor genético de cada uno de ellos:

$$E(\mathbf{u}) = \hat{\mathbf{u}} = \mathbf{C'V^{-1}(y - Xb)}$$

Recordemos que los estimadores minimocuadráticos de los efectos fijos no son otra cosa que la media de los datos dentro de niveles de los efectos fijos. Así, en la práctica, se ajustan los datos para los efectos fijos restándoles las medias de sus grupos y se les ajusta un modelo aleatorio: $\mathbf{y^* = y - Xb = Zu + e}$, cuya solución es la expresada aquí:

$$\hat{\mathbf{u}} = \mathbf{C'V^{-1}y^*}$$

En este ajuste se producen las otras implicaciones de asumir los estimadores de **b** como los verdaderos valores de **b**. En primer lugar, al ser estimaciones sesgadas se origina también un sesgo en la predicción de los efectos aleatorios. Y, finalmente, dado que el valor de los efectos fijos se asume conocido sin error, cuando en realidad lo tiene, se introduce un error en **y*** que no existía en **y**. La magnitud de este error va a depender del error del estimador de la media para la que se ha ajustado, ya que, recuérdese, la varianza de la media es igual a la varianza de la variable partido por el número de datos ($\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$). Así pues, cuando se calcule la precisión de los valores genéticos se ignorará este error pudiendo dar por muy preciso un valor genético que en realidad no lo es tanto. Se originan así errores al medir la precisión que serán tan

grandes como los errores de las medias de los grupos a los que pertenecen.

En los índices de selección el vector **u**, vector de valores genéticos aditivos es aquello que se desea mejorar y se llama objetivo de selección. La información que se utiliza para lograr mejorar el objetivo es la fuente de información, lo que en el contexto de los índices de selección se llama criterio de selección. No existe una norma sobre el número de objetivos y criterios de selección a tener en cuenta, y éstos dependerán del índice concreto que estemos considerando. Dispondremos en el índice dos matrices cuya composición conviene recordar. Así, si se tienen en cuenta *n* objetivos y *m* criterios, los elementos **C'** y **V** son:

$$\mathbf{C'} = Cov(\mathbf{u}, \mathbf{y}^{*}\mathbf{'}) = \begin{bmatrix} \sigma_{u_1 y_1} & \sigma_{u_1 y_2} & \sigma_{u_1 y_3} & .... & \sigma_{u_1 y_m} \\ \sigma_{u_2 y_1} & \sigma_{u_2 y_2} & \sigma_{u_2 y_3} & .... & \sigma_{u_2 y_m} \\ ... & ... & ... & .... & ... \\ ... & ... & ... & .... & ... \\ \sigma_{u_n y_1} & \sigma_{u_n y_2} & \sigma_{u_n y_3} & .... & \sigma_{u_n y_m} \end{bmatrix}$$

$$\mathbf{V} = Var(\mathbf{y}^{*}) = \begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & \sigma_{y_1^* y_3^*} & .... & \sigma_{y_1^* y_m^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & \sigma_{y_2^* y_3^*} & .... & \sigma_{y_2^* y_m^*} \\ ... & ... & ... & .... & ... \\ ... & ... & ... & .... & ... \\ \sigma_{y_m^* y_1^*} & \sigma_{y_m^* y_2^*} & \sigma_{y_m^* y_3^*} & .... & \sigma^2_{y_m^*} \end{bmatrix}$$

Obsérvese que al no aportar los efectos fijos varianza a los datos, se cumple:

$$Var\,(\mathbf{y^*}) = Var\,(\mathbf{y} - \mathbf{Xb}) = Var(\mathbf{y}) = \mathbf{V}$$

A pesar de ser sesgados, los índices de selección han sido ampliamente utilizados cuando no existían herramientas informáticas ni se había derivado una metodología que permitiera resolver conjuntamente efectos fijos y aleatorios. Hoy en día los

índices de selección forman parte del pasado y es la metodología BLUP la que ha sido extendida en la valoración genética de todos los esquemas de selección a nivel internacional. Sin embargo su estudio es muy interesante en el aprendizaje de la valoración genética porque permite comparar la precisión de las valoraciones genéticas cuando se utilizan distintas fuentes de información, lo que permite organizar la recogida de datos de manera que su uso posterior sea mejor aprovechado. Así pues, la medida de la precisión es un concepto de elevado interés en el contexto de los índices de selección.

VI-2. La medida del error y de la precisión

Mientras que la varianza de los estimadores proporciona una buena medida de su error, no ocurre así con la varianza de los predictores. Según fue comentado anteriormente, que a diferencia de los efectos fijos, los efectos aleatorios presentan conceptualmente una varianza en su definición. Optaremos entonces por un concepto cercano pero diferente. La medida del error se hará entonces a partir de la varianza del error de predicción ya que fue este concepto el que se empleo como mínimo en la definición de "mejor" predictor.

Así, $\mathbf{u}$ es el vector que contiene los verdaderos valores de los efectos aleatorios, $\hat{\mathbf{u}}$ el vector con sus predictores, $\hat{\mathbf{u}} - \mathbf{u}$ es vector que contiene los errores de predicción, y su varianza (Varianza del Error de Predicción o VEP), se desarrolla a continuación:

$$VEP = Var\left(\hat{\mathbf{u}} - \mathbf{u}\right) = Var\left(\hat{\mathbf{u}}\right) + Var\left(\mathbf{u}\right) - Cov\left(\hat{\mathbf{u}}, \mathbf{u}'\right) - Cov\left(\mathbf{u}, \hat{\mathbf{u}}'\right)$$

Desarrollando por partes los distintos elementos de esta expresión:

$Var\left(\mathbf{u}\right) = \mathbf{G}$ por definición del modelo

$$Var\left(\hat{\mathbf{u}}\right) = Var\left(\mathbf{C}'\mathbf{V}^{-1}\mathbf{y}^{*}\right) = \mathbf{C}'\mathbf{V}^{-1}Var(\mathbf{y}^{*})\mathbf{V}^{-1}\mathbf{C} =$$
$$= \mathbf{C}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{C} = \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$$

$$Cov(\hat{\mathbf{u}},\mathbf{u}') = Cov\left(\mathbf{C}'\mathbf{V}^{-1}\mathbf{y}^{*},\mathbf{u}'\right) = \mathbf{C}'\mathbf{V}^{-1}Cov(\mathbf{y}^{*},\mathbf{u}') = \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$$

$$Cov(\hat{\mathbf{u}},\mathbf{u}') = Cov(\mathbf{u},\hat{\mathbf{u}}')' = \left(\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}\right)' = \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$$

Y sustituyendo:

$$VEP = \mathbf{G} - \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$$

Si interesante es la medida del error, igualmente interesante es la medida de la precisión. Pero en el caso de la precisión el valor se puede dar de forma adimensional al proporcionar la correlación entre el predictor y el verdadero valor ($\rho_{\hat{u}u}$). Como muchas de las correlaciones ésta se puede denominar repetibilidad del mérito genético. Por motivos de cálculo la precisión se calcula como su potencia al cuadrado, valor que se denomina fiabilidad y que es preferido en muchos catálogos como forma de informar del grado de certidumbre del valor genético de un individuo. Obtendremos entonces la fiabilidad a partir del cuadrado de $\rho_{\hat{u}u}$ para facilitar su desarrollo, aunque una vez calculada la precisión de cada individuo, debería obtenerse de ésta la raíz cuadrada. Como todas las correlaciones, se obtendrá la covarianza entre los vectores dividido por el producto de las desviaciones típicas. Dado que trabajamos el cuadrado del parámetro se dividiría por el producto de las matrices de varianzas y covarianzas, pero por tratarse de matrices en las que la operación división no se encuentra definida, se multiplicará por las inversas:

$$\boldsymbol{\rho}^{2}_{\hat{\mathbf{u}}\mathbf{u}} = \left[Cov(\hat{\mathbf{u}},\mathbf{u}')\right]^{2}\left[Var(\hat{\mathbf{u}})Var(\mathbf{u})\right]^{-1}$$

Y como $Cov(\hat{\mathbf{u}},\mathbf{u}') = Var(\hat{\mathbf{u}}) = \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$, el cuadrado del numerador se cancela con *Var(*$\hat{\mathbf{u}}$*)* del denominador:

$$\boldsymbol{\rho}^{2}_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}\right)\mathbf{G}^{-1}$$

Obsérvese que $\boldsymbol{\rho}^{2}_{\hat{\mathbf{u}}\mathbf{u}}$ es una matriz de correlaciones entre cada valor genético aditivo y cada predictor. En la diagonal se encontrará el

cuadrado de la precisión de cada uno de los predictores por lo que será necesario después extraer la raíz cuadrada.

RELACIÓN ENTRE LA VARIANZA DEL ERROR DE PREDICCIÓN Y LA FIABILIDAD

- Obtención de la fiabilidad a partir de la varianza del error de predicción. Partimos de VEP:

$$VEP = \mathbf{G} - \mathbf{C'V^{-1}C}$$

Multiplicamos ambos lados de la igualdad por la inversa de **G**:

$$VEP\mathbf{G^{-1}} = \mathbf{GG^{-1}} - \left(\mathbf{C'V^{-1}C}\right)\mathbf{G^{-1}} = \mathbf{I} - \left(\mathbf{C'V^{-1}C}\right)\mathbf{G^{-1}}$$

Despejamos $\left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1}$ para obtener la precisión:

$$\mathbf{I} - VEP\mathbf{G}^{-1} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1}$$

$$\boxed{\boldsymbol{\rho}^2_{\hat{u}u} = \mathbf{I} - VEP\mathbf{G}^{-1}}$$

- Obtención de la varianza del error de predicción a partir de la fiabilidad. Se deduce de la expresión anterior:

$$\boxed{VEP = \left(\mathbf{I} - \boldsymbol{\rho}^2_{\hat{u}u}\right)\mathbf{G}}$$

VI-3. Utilidad de los índices de selección en mejora animal

Los índices de selección fueron ampliamente utilizados en el siglo pasado cuando aún no se había desarrollado la informática ni había sido desarrollada la forma de resolver el modelo lineal mixto. Teóricamente para cada individuo se debía estudiar la información de la que se disponía, es decir, si teníamos su propio dato, el de un hermano, el del padre, etc., es decir, los criterios de selección a considerar. En función del tipo de información se desarrollaría un índice que proporcionase los coeficientes que establecerían la combinación lineal de esos criterios de selección que llevarían al predictor de su valor genético aditivo. Para ello habría que desarrollar los coeficientes que formarían parte de las matrices **C'** y **V** en cada caso.

El desarrollo de un índice de selección específico par cada individuo es una tarea inabordable ya que las combinaciones de criterios de selección disponibles posibles son infinitas por lo que esta forma de proceder no era operativa.

Así que la forma de proceder era bien distinta. Se comparaba la precisión que se obtenía al usar determinados tipos de información para cada caso y se diseñaba la recogida de datos de acuerdo con esa decisión. Así nacieron por ejemplo los centros de inseminación artificial, es decir, porque proporcionaban una precisión elevada para evaluar machos que pudieran ser concentrados en un centro de inseminación artificial.

El estudio de la precisión de diferentes tipos de índices de selección resulta enormemente didáctico por lo que a continuación se desarrollan algunos ejemplos.

VI-4. Desarrollo de índices de selección concretos

Procederemos a continuación a desarrollar diversos índices de selección. Para cada caso el proceso será el siguiente:

1. En primer lugar desarrollaremos los valores concretos de $\mathbf{C'V^{-1}}$ para establecer los coeficientes que multiplican a los criterios de regresión contemplados en el índice. Recuérdese que se trata de una regresión, de manera que **y*** será la variable regresora y $\hat{\mathbf{u}}$ la dependiente en: $\hat{\mathbf{u}} = \mathbf{C'V^{-1}y}^{*}$.

2. Se obtendrá **C** (la matriz transpuesta de **C'** ya desarrollada) y **G**, para, ayudados de los coeficientes del índice ya obtenidos, calcular la precisión mediante $\boldsymbol{\rho}^{2}_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1}$, teniendo en cuenta que posteriormente habrá que obtener la raíz cuadrada de la precisión.

Recordemos que en los dos pasos anteriores las tres matrices a detallar son:

$$\mathbf{C'} = Cov\,(\mathbf{u}, \mathbf{y^{*'}}) \qquad \mathbf{V} = Var(\mathbf{y^{*}}) \qquad \mathbf{G} = Var(\mathbf{u})$$

En algunos casos será preciso hacer uso de la ecuación principal de la herencia aditiva según la cual el valor genético aditivo de un individuo *i* es igual a la media de los valores genéticos aditivos de sus padres *j* y *k* más un muestreo mendeliano φ_i, es decir, aquella parte del valor aditivo de un individuo que lo hace diferente del de sus hermanos y que se debe a que durante la meiosis un gameto es portador únicamente de uno de los dos alelos del reproductor que porta de forma aleatoria:

$$u_i = \tfrac{1}{2}u_j + \tfrac{1}{2}u_k + \varphi_i$$

Dado que los alelos que quedan en los gametos son asignados aleatoriamente, las covarianzas entre un muestreo mendeliano y cualquier otro, o entre el mismo y el valor genético aditivo de otro individuo, son nulas al ser independientes.

Asimismo, cuando se trate de calcular la covarianza entre un dato y_i y cualquier otra cosa que no sea el propio dato, éste se descompondrá en el valor genético aditivo que ha originado ese dato y un residuo:

$$y_i^* = u_i + e_i$$

Finalmente, toda relación de parentesco no contemplada en el índice se asumirá que no existe. Como consecuencia, toda covarianza genética aditiva entre individuos cuyo parentesco se desconozca, valdrá cero.

Para evitar repeticiones innecesarias, en los sucesivos índices a desarrollar en este texto se obviarán algunos de los razonamientos ya detallados en índices previos.

VI-4.1. Índice de selección individual

Se usa el propio dato y_i^* como fuente de información (como variable regresora o también como criterio de selección) para evaluar al animal (para obtener $\hat{u}_i$ como predictor de u_i).

- Desarrollo del índice

C' es en principio una matriz de covarianzas entre el vector de objetivos y el vector de criterios. Pero en este caso, el vector de objetivos no es tal vector, sino un solo elemento u_i, ni el vector de criterios es tal vector sino un único dato y_i^*. Por tanto, en este caso:

$$\mathbf{C'} = Cov(u_i, y_i^*) = Cov(u_i, u_i + e_i) = Cov(u_i, u_i) + Cov(u_i, e_i) = \\ = Var(u_i) = \sigma_u^2$$

En esta expresión se ha hecho uso de algunas de las definiciones del modelo. Según ellas, toda covarianza entre un valor genético aditivo y un residuo es nula. Finalmente la varianza del valor genético aditivo de un individuo, por ser una realización de la distribución a la que pertenece, coincidirá con la varianza de dicha distribución. En este caso, al no conocer los padres del individuo *i*, se trata de un individuo fundador cuyo valor genético aditivo pertenece a la distribución de los valores genéticos aditivos de los fundadores, y la varianza de u_i sería la varianza genética aditiva del carácter. En realidad se puede demostrar que éste será el valor de la varianza genética aditiva de todo individuo no consanguíneo, como se verá más adelante.

V es en principio una matriz de varianzas y covarianzas del vector de datos, pero nuevamente sólo tenemos un dato, por lo que **V** no es tal matriz sino simplemente un escalar. Razonando de igual modo que en el caso anterior, y_i^* pertenece a la distribución de datos por lo que su varianza será la varianza fenotípica:

$$\mathbf{V} = Var(y_i^*) = \sigma_p^2$$

Y así, el índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y^*} = \sigma_u^2 \frac{1}{\sigma_p^2} y_i^* = \frac{\sigma_u^2}{\sigma_p^2} y_i^* = h^2 y_i^*$$

Así pues, integrando la información aportada hasta este punto, los datos serán desviados de las medias de sus grupos y el resultado multiplicado por la heredabilidad (h^2) nos proporciona el predictor del valor genético del individuo, de manera que la selección se hará sencillamente escogiendo al individuo que posea un mejor valor.

- Desarrollo de la precisión del índice

En el índice ya se tiene detallado el valor de **C'V^{-1}**. Para la precisión sólo nos falta **C** que es la matriz transpuesta de **C'**, y **G**. Como una transpuesta consiste en poner las filas como columnas y viceversa, y **C'** es un escalar, en este caso **C** = **C'** = σ_u^2. **G** es también en teoría una matriz y en este caso un escalar cuyo valor es la varianza de u_i que nuevamente coincide con σ_u^2. Así pues, la precisión del índice de selección individual es:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = h^2\sigma_u^2\left(\sigma_u^2\right)^{-1} = h^2 \qquad \rho_{\hat{u}u} = \sqrt{h^2} = h$$

Por tanto, la precisión del índice de selección coincide con la heredabilidad del carácter o, más concretamente, con su raíz cuadrada. Por tanto, si la heredabilidad del carácter es alta, este tipo de índice proporciona una buena valoración, mientras que si es baja habrá que buscar una alternativa.

VI-4.2. Índice de selección a partir de la información de uno de los padres

Se usa el dato del padre del individuo y_p^* como fuente de información para evaluar al animal.

- Desarrollo del índice

C' es nuevamente en principio una matriz de covarianzas entre el vector de objetivos y el vector de criterios. En todos los casos como el anterior en el que sólo hay un objetivo y un criterio C' es un único elemento y por tanto un escalar:

$$\mathbf{C'} = Cov(u_i, y_p^*) = Cov(u_i, u_p + e_p) = Cov(u_i, u_p) + Cov(u_i, e_p) =$$
$$= Cov(u_i, u_p) = Cov(\tfrac{1}{2}u_p + \tfrac{1}{2}u_m + \varphi_i, u_p) =$$
$$= Cov(\tfrac{1}{2}u_p, u_p) + Cov(\tfrac{1}{2}u_m, u_p) + Cov(\varphi_i, u_p) =$$
$$= \tfrac{1}{2}\, Var(u_p) + \tfrac{1}{2}\, Cov(u_m, u_p) = \tfrac{1}{2}\, \sigma_u^2$$

Las novedades en este desarrollo en relación a la misma matriz del ejemplo precedente son dos. Cuando se ha precisado calcular la covarianza entre los valores genéticos aditivos de dos individuos que pertenecen a generaciones diferentes, se ha descompuesto el más joven en función de los valores genéticos aditivos de sus padres. La otra novedad ha sido igualar a cero la covarianza entre los valores genéticos aditivos de los padres del individuo, dado que depende de su parentesco y éste, al no existir información en el índice, se asume que no existe.

V es nuevamente en principio una matriz de varianzas y covarianzas del vector de datos, pero nuevamente sólo tenemos un dato, por lo que **V** no es tal matriz sino simplemente un escalar. Razonando de igual modo que en el caso anterior, y_p^* pertenece a la distrbución de datos por lo que su varianza será la varianza fenotípica:

$$\mathbf{V} = Var(y_p^*) = \sigma_p^2$$

Y así, el índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \tfrac{1}{2}\sigma_u^2 \frac{1}{\sigma_p^2} y_i^* = \tfrac{1}{2}\frac{\sigma_u^2}{\sigma_p^2} y_i^* = \tfrac{1}{2}h^2 y_i^*$$

- Precisión del índice

Para la precisión nuevamente tenemos **C'V^{-1}** y sólo nos falta **C** como transpuesta de **C'**, y **G**. Nuevamente es un escalar por lo que **C** = **C'** = $\tfrac{1}{2}\sigma_u^2$. **G** sigue siendo la varianza de u_i que nuevamente coincide con σ_u^2. Así pues, la precisión del índice de selección es:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \tfrac{1}{2}h^2\,\tfrac{1}{2}\sigma_u^2\left(\sigma_u^2\right)^{-1} = \tfrac{1}{4}h^2$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{\tfrac{1}{4}h^2} = \tfrac{1}{2}h}$$

Por tanto, la precisión del índice de selección en este caso se reduce a la mitad. En el caso de tener que elegir entre el dato del individuo y el de su padre, siempre será mejor el primero.

VI-4.3. Índice de selección a partir del dato de un hijo

Se usa el dato del hijo del individuo y_h^* como regresora para evaluar al animal.

- Desarrollo del índice

C'. La única consideración a hacer aquí es que ahora el individuo más joven (*h*) es el que poporciona el dato por lo que será éste el que haya que descomponer en función de los valores genéticos aditivos de sus padres. El padre es en este caso *i* y se ha mantenido *m* para la madre:

$$\mathbf{C'} = Cov(u_i, y_h^*) = Cov(u_i, u_h + e_h) = Cov(u_i, u_h) + Cov(u_i, e_h) =$$
$$= Cov(u_i, u_h) = Cov(u_i, \tfrac{1}{2}u_i + \tfrac{1}{2}u_m + \varphi_h) =$$
$$= Cov(u_i, \tfrac{1}{2}u_i) + Cov(u_i, \tfrac{1}{2}u_m) + Cov(u_i, \varphi_h) =$$
$$= \tfrac{1}{2}\, Var(u_i) = \tfrac{1}{2}\, \sigma_u^2$$

V:

$$\mathbf{V} = Var(y_h^*) = \sigma_p^2$$

Y así, el coeficiente del índice de selección sería el mismo que en el caso precedente:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \tfrac{1}{2}\sigma_u^2 \frac{1}{\sigma_p^2} y_h^* = \tfrac{1}{2}\frac{\sigma_u^2}{\sigma_p^2} y_h^* = \tfrac{1}{2} h^2 y_h^*$$

- Precisión del índice

Para la precisión otra vez $\mathbf{C} = \mathbf{C'} = \tfrac{1}{2}\sigma_u^2$ y $\mathbf{G} = \sigma_u^2$:

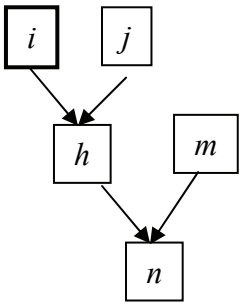
$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \tfrac{1}{2} h^2\, \tfrac{1}{2}\sigma_u^2 \left(\sigma_u^2\right)^{-1} = \tfrac{1}{4} h^2$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{\tfrac{1}{4} h^2} = \tfrac{1}{2} h}$$

Por tanto, la precisión del índice de selección es exactamente la misma cuando el dato es de un padre y cuando el dato es de un hijo. Es razonable ya que la relación de parentesco es la misma entre ambos individuos. En el caso de tener que elegir entre el dato del hijo y el de su padre, ambos proporcionan la misma precisión. Sin embargo si se trata de utilizar más de un padre o más de un hijo ha de tenerse en cuenta que en el caso de utilizar el dato del padre el intervalo generacional se acorta al tener al individuo evaluado incluso antes de que nazca. Sin embargo como máximo dispondremos de datos de dos padres, mientras que si se utilizan datos de hijos podremos disponer de muchos más, especialmente si se diseña apropiadamente la recogida de datos como sucede con la utilización de los centros de inseminación artificial.

VI-4.4. Índice de selección a partir del dato de un nieto

Se presenta ahora el caso de un índice con el dato de un pariente más alejado. En este caso será el nieto *n*, hijo de *h* que a su vez es hijo de *i*:



- Desarrollo del índice

C':

$$\mathbf{C'} = Cov\,(u_i,\ y_n^*) = Cov\,(u_i,\ u_n + e_n) = Cov\,(u_i,\ u_n) =$$
$$= Cov\,(u_i,\ \tfrac{1}{2}u_h + \tfrac{1}{2}u_m + \varphi_n) = \tfrac{1}{2}\,Cov\,(u_i,\ u_h) =$$
$$= \tfrac{1}{2}\,Cov\,(u_i,\ \tfrac{1}{2}u_i + \tfrac{1}{2}u_j + \varphi_h) = \tfrac{1}{2}\,\tfrac{1}{2}\,Cov\,(u_i,\ u_i) =$$
$$= \tfrac{1}{4}\,Var\,(u_i) = \tfrac{1}{4}\,\sigma_u^2$$

V:

$$\mathbf{V} = Var\,(y_n^*) = \sigma_p^2$$

Y así, el coeficiente del índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y^*} = \tfrac{1}{4}\sigma_u^2 \frac{1}{\sigma_p^2} y_n^* = \tfrac{1}{4}\frac{\sigma_u^2}{\sigma_p^2} y_n^* = \tfrac{1}{4} h^2 y_n^*$$

- Precisión del índice

Para la precisión $\mathbf{C} = \mathbf{C'} = ¼\ \sigma_u^2$ y $\mathbf{G} = \sigma_u^2$:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = ¼\, h^2\ ¼\,\sigma_u^2\left(\sigma_u^2\right)^{-1} = 1/16\, h^2$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{1/16\, h^2} = ¼\, h}$$

Así que a medida que el individuo que proporciona la fuente de información se aleja en parentesco del individuo a evaluar la precisión se reduce. Por tanto, si sólo se puede escoger un dato para evaluar a un animal, mejor si puede ser el del propio animal, y si no, mejor cuanto más cercano sea el pariente que lo proporciona.

VI-4.5. Índice de selección a partir de la media de los *n* datos del propio individuo

Este índice de selección sigue teniendo la misma dimensión dado que sólo se evalúa un individuo a partir de un solo dato. Sin embargo, este dato, la media, está compuesto de otros, lo que dificulta ligeramente su desarrollo.

- Desarrollo del índice

C':

$$\mathbf{C'} = Cov\,(u_i,\ \bar{y}^*) = Cov\left(u_i, \frac{y_1^* + y_2^* + ... + y_n^*}{n}\right) =$$

$$= \frac{1}{n} Cov\left(u_i, y_1^* + y_2^* + ... + y_n^*\right) =$$

$$= \frac{1}{n}\left[Cov\left(u_i, y_1^*\right) + Cov\left(u_i, y_2^*\right) + ... + Cov\left(u_i, y_n^*\right)\right] =$$

$$= \frac{1}{n} n\left[Cov\left(u_i, y_i^*\right)\right] = \sigma_u^2$$

En esta expresión los subíndices de los datos corresponden al número de observación del individuo en lugar de hacer referencia a la propiedad del individuo, ya que en este caso son todos del individuo *i*. Obsérvese que todas las covarianzas dentro del corchete se corresponden a la covarianza entre el valor genético de un individuo y su dato, por lo que la suma es igual a *n* veces una de ellas.

V: Esta matriz corresponde a la varianza de una media, la cuál, si todos los elementos que la compusieran fueran independientes, simplemente sería la varianza de la variable partido por el número de datos. Sin embargo, todos estos datos pertenecen al mismo individuo por lo que no son independientes. Ello obliga a considerar todas las covarianzas entre cada dos datos. Por ello habrá que extender esta media para el desarrollo de **V**, aunque al final ha de conducir a la varianza de un único elemento y por ello tiene que terminar dando un escalar. Para el cálculo de esta matriz es preciso recordar el concepto de repetibilidad (R) de un dato. Se trata del cociente ente la varianza genética aditiva más la varianza ambiental permanente partido por la varianza fenotípica pero aquí nos interesa desde el punto de vista de su interpretación. Es igual a la correlación entre dos medidas de un carácter del mismo individuo, y por lo tanto es igual al cociente entre la covarianza entre dos datos de un individuo y el producto de sus desviaciones típicas, que por pertenecer ambas a la misma distribución, es igual a la varianza fenotípica. Por tanto, Si llamamos y_i^* y y_j^* a dos datos cualquiera de un individuo se puede deducir su covarianza que será necesaria después:

$$R = \frac{\sigma_{y_i^* y_j^*}}{\sigma_{y_i^*}\sigma_{y_j^*}} = \frac{\sigma_{y_i^* y_j^*}}{\sigma_p \sigma_p} = \frac{\sigma_{y_i^* y_j^*}}{\sigma_p^2} \Rightarrow \sigma_{y_i^* y_j^*} = R\sigma_p^2$$

También es necesario observar que la suma de los elementos de un vector se obtiene premultiplicándolo por un vector fila de unos, y que esta suma proporciona la media al ser dividida por *n*. Asimismo, un vector fila de unos premultiplicando una matriz da

lugar a un vector fila con las sumas de los elementos de cada columna. Del mismo modo, un vector fila postmultiplicado por un vector columna de unos conduce a la suma de los elementos del vector. De acuerdo con todo esto se puede desarrollar **V**:

$$\mathbf{V} = Var(\overline{y}_n^*) = Var\left(\frac{1}{n}\mathbf{1'y}\right) = Var\left(\frac{1}{n}\begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}\begin{bmatrix}y_1^* \\ y_2^* \\ \vdots \\ y_n^*\end{bmatrix}\right) =$$

$$= \frac{1}{n^2}\begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix} Var(\mathbf{y}) \begin{bmatrix}1 \\ 1 \\ \vdots \\ 1\end{bmatrix} =$$

$$= \frac{1}{n^2}\mathbf{1'}Var(\mathbf{y})\mathbf{1} = \frac{1}{n^2}\mathbf{1'}\begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & \sigma_{y_1^* y_3^*} & \dots. & \sigma_{y_1^* y_n^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & \sigma_{y_2^* y_3^*} & \dots. & \sigma_{y_2^* y_n^*} \\ \dots & \dots & \dots & \dots. & \dots \\ \dots & \dots & \dots & \dots. & \dots \\ \sigma_{y_n^* y_1^*} & \sigma_{y_n^* y_2^*} & \sigma_{y_n^* y_3^*} & \dots. & \sigma^2_{y_n^*} \end{bmatrix}\mathbf{1} =$$

$$= \frac{1}{n^2}\mathbf{1'}\begin{bmatrix} \sigma_p^2 & R\sigma_p^2 & R\sigma_p^2 & \dots. & R\sigma_p^2 \\ R\sigma_p^2 & \sigma_p^2 & R\sigma_p^2 & \dots. & R\sigma_p^2 \\ \dots & \dots & \dots & \dots. & \dots \\ \dots & \dots & \dots & \dots. & \dots \\ R\sigma_p^2 & R\sigma_p^2 & R\sigma_p^2 & \dots. & \sigma_p^2 \end{bmatrix}\mathbf{1} =$$

$$= \frac{1}{n^2}\begin{bmatrix}\sigma_p^2 + (n-1)R\sigma_p^2 & \sigma_p^2 + (n-1)R\sigma_p^2 & \dots & \sigma_p^2 + (n-1)R\sigma_p^2\end{bmatrix}\begin{bmatrix}1 \\ 1 \\ \vdots \\ 1\end{bmatrix} =$$

$$= \frac{1}{n^2} n\left[\sigma_p^2 + (n-1)R\sigma_p^2\right] = \frac{\sigma_p^2 + (n-1)R\sigma_p^2}{n}$$

Y así, el coeficiente del índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \sigma_u^2 \frac{n}{\sigma_p^2 + (n-1)R\sigma_p^2} \bar{y}^* =$$

$$= \frac{n\sigma_u^2 \big/ \sigma_p^2}{\left[\sigma_p^2 + (n-1)R\sigma_p^2\right] \big/ \sigma_p^2} \bar{y}^* = \frac{nh^2}{1+(n-1)R} \bar{y}^*$$

- Precisión del índice

Para la precisión $\mathbf{C} = \mathbf{C'} = \sigma_u^2$ y $\mathbf{G} = \sigma_u^2$:

$$\boldsymbol{\rho}_{\hat{\mathbf{u}}\mathbf{u}}^{\mathbf{2}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \frac{nh^2}{1+(n-1)R} \sigma_u^2 \left(\sigma_u^2\right)^{-1} = \frac{nh^2}{1+(n-1)R}$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{\frac{nh^2}{1+(n-1)R}}}$$

Por tanto, cuando se usa la media de *n* medidas del individuo como criterio de selección, la precisión ya no sólo depende de la heredabilidad sino también del número de medidas que componen esa media y de la repetibilidad del carácter.

Se puede dar así respuesta a una interesante pregunta que puede plantearse previamente al diseño del registro de información: ¿cuántos datos hacen falta para evaluar un individuo con una determinada precisión para ciertos valores de heredabilidad y repetibilidad? Para responder a esta pregunta basta despejar n en la expresión anterior y sustituir los valores de las condiciones deseadas:

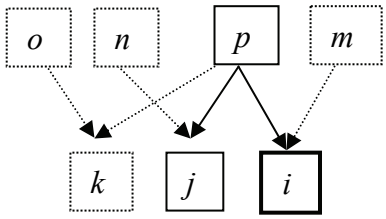
$$\boxed{n = \frac{\rho_{\hat{u}u}^2 - R\rho_{\hat{u}u}^2}{h^2 - R\rho_{\hat{u}u}^2}}$$

VI-4.6. Índice de selección a partir de la media de los datos de *n* hermanos de padre

Desarrollaremos dos casos diferentes, el primero cuando la media no incluye el dato del propio individuo a evaluar y la segunda cuando sí la incluye.

VI-4.6.1. La media de los datos no incluye el dato del individuo

La representación gráfica de la relación de parentesco entre el individuo a evaluar *i* y sus medios hermanos (*j*, *k*, …) es la siguiente:



En esta figura *p* es el padre de *i*, *j* y *k*, mientras que las madres *o*, *n* y *m*, son todas distintas. Seguimos con índices que utilizan un único criterio.

- Desarrollo del índice

C':

$$\mathbf{C'} = Cov\,(u_i,\ \bar{y}^*) = Cov\left(u_i, \frac{y_1^* + y_2^* + \ldots + y_n^*}{n}\right) =$$

$$= \frac{1}{n} Cov\left(u_i, y_1^* + y_2^* + \ldots + y_n^*\right) =$$

$$= \frac{1}{n}\left[Cov\left(u_i, y_1^*\right) + Cov\left(u_i, y_2^*\right) + ... + Cov\left(u_i, y_n^*\right)\right] =$$

$$= \frac{1}{n} n\left[Cov\left(u_i, y_j^*\right)\right] = Cov\left(u_i, u_j\right) =$$

$$= Cov\left(u_i, u_j\right) = Cov\left(½u_p + ½u_m + \varphi_i, ½u_p + ½u_n + \varphi_j\right) =$$

$$= Var\left(½u_p\right) = ¼\sigma_u^2$$

En esta expresión los subíndices de los datos distinguen a cada uno de los hermanos del individuo evaluado. Nuevamente todas las covarianzas dentro del corchete se corresponden a la covarianza entre el valor genético de un individuo y el dato de su medio hermano, por lo que la suma es igual a *n* veces una de ellas. Finalmente, por estar los dos individuos en la misma generación, el valor genético aditivo de los dos individuos se descompone en función de sus respectivos padres.

V: Nuevamente, al tratarse de la varianza de una media de datos correlacionados, aunque al final el resultado será un escalar, hay que desarrollar la matriz de covarianzas entre datos de medios hermanos. Por tanto, necesitaremos conocer el valor de una de estas covarianzas entre los datos de dos medios hermanos de i, por ejemplo, j y k:

$$\sigma_{y_j^* y_k^*} = Cov\left(y_j^*, y_k^*\right) = Cov\left(u_j, u_k\right) = Cov\left(½u_p, ½u_p\right) =$$

$$= Var\left(½u_p\right) = ¼\sigma_u^2$$

Nótese que la covarianza entre los datos de dos medios hermanos será la misma independientemente de que uno de ellos vaya a ser evaluado o no. De acuerdo con esto se puede desarrollar **V** para este caso:

$$\mathbf{V} = Var(\overline{y}_n^*) = Var\left(\frac{1}{n}\mathbf{1'y}\right) = Var\left(\frac{1}{n}\begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}\begin{bmatrix}y_1^* \\ y_2^* \\ \vdots \\ y_n^*\end{bmatrix}\right) =$$

$$= \frac{1}{n^2}\begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix} Var(\mathbf{y}) \begin{bmatrix}1 \\ 1 \\ \vdots \\ 1\end{bmatrix} = \frac{1}{n^2}\mathbf{1'}Var(\mathbf{y})\mathbf{1} =$$

$$= \frac{1}{n^2}\mathbf{1'}\begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & \sigma_{y_1^* y_3^*} & \dots. & \sigma_{y_1^* y_n^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & \sigma_{y_2^* y_3^*} & \dots. & \sigma_{y_2^* y_n^*} \\ \dots & \dots & \dots & \dots. & \dots \\ \dots & \dots & \dots & \dots. & \dots \\ \sigma_{y_n^* y_1^*} & \sigma_{y_n^* y_2^*} & \sigma_{y_n^* y_3^*} & \dots. & \sigma^2_{y_n^*} \end{bmatrix}\mathbf{1} =$$

$$= \frac{1}{n^2}\mathbf{1'}\begin{bmatrix} \sigma^2_p & \frac{1}{4}\sigma^2_u & \frac{1}{4}\sigma^2_u & \dots. & \frac{1}{4}\sigma^2_u \\ \frac{1}{4}\sigma^2_u & \sigma^2_p & \frac{1}{4}\sigma^2_u & \dots. & \frac{1}{4}\sigma^2_u \\ \dots & \dots & \dots & \dots. & \dots \\ \dots & \dots & \dots & \dots. & \dots \\ \frac{1}{4}\sigma^2_u & \frac{1}{4}\sigma^2_u & \frac{1}{4}\sigma^2_u & \dots. & \sigma^2_p \end{bmatrix}\mathbf{1} =$$

$$= \frac{1}{n^2}\begin{bmatrix}\sigma^2_p + (n-1)\frac{1}{4}\sigma^2_u & \sigma^2_p + (n-1)\frac{1}{4}\sigma^2_u & \dots & \sigma^2_p + (n-1)\frac{1}{4}\sigma^2_u\end{bmatrix}\begin{bmatrix}1 \\ 1 \\ \vdots \\ 1\end{bmatrix} =$$

$$= \frac{1}{n^2} n\left[\sigma^2_p + (n-1)\frac{1}{4}\sigma^2_u\right] = \frac{\sigma^2_p + (n-1)\frac{1}{4}\sigma^2_u}{n}$$

Y así, el coeficiente del índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \tfrac{1}{4}\sigma_u^2 \frac{n}{\sigma_p^2 + (n-1)\tfrac{1}{4}\sigma_u^2}\,\bar{y}^* =$$

$$= \frac{\tfrac{1}{4}n\sigma_u^2 \big/ \sigma_p^2}{\left[\sigma_p^2 + (n-1)\tfrac{1}{4}\sigma_u^2\right] \big/ \sigma_p^2}\,\bar{y}^* = \frac{\tfrac{1}{4}nh^2}{1+\tfrac{1}{4}(n-1)h^2}\,\bar{y}^*$$

- Precisión del índice

Para la precisión $\mathbf{C} = \mathbf{C'} = \tfrac{1}{4}\sigma_u^2$ y $\mathbf{G} = \sigma_u^2$:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \frac{\tfrac{1}{4}nh^2}{1+\tfrac{1}{4}(n-1)h^2}\,\tfrac{1}{4}\sigma_u^2\left(\sigma_u^2\right)^{-1} =$$

$$= \tfrac{1}{16}\frac{nh^2}{1+\tfrac{1}{4}(n-1)h^2}$$

$$\boxed{\rho_{\hat{u}u} = \tfrac{1}{4}\sqrt{\frac{nh^2}{1+\tfrac{1}{4}(n-1)h^2}}}$$

Por tanto, cuando se usa la media de los datos de *n* medios hermanos del individuo como criterio de selección, la precisión depende de la heredabilidad y del número de hermanos medidos.

Así, si se desea conocer el número de datos necesarios de medios hermanos de un individuo para obtener una precisión determinada se podría despejar *n* en esta expresión para dar:

$$\boxed{n = \frac{4\rho^2_{\hat{u}u}h^2 - 16\rho^2_{\hat{u}u}}{4\rho^2_{\hat{u}u}h^2 - h^2}}$$

VI-4.6.2. La media de los datos sí incluye el dato del individuo

- Desarrollo del índice

C'. Uno de los datos que participan de la media es precisamente y_i^*, por lo que no todas las covarianzas dentro del corchete son idénticas. Esto va a dar lugar a cambios en **C'** ya que sólo *n-1* covarianzas genéticas aditivas serán entre medios hermanos ($¼\sigma_u^2$), mientras que una de ellas será la varianza genética aditiva de un individuo (σ_u^2):

$$\mathbf{C'} = Cov\left(u_i, \bar{y}^*\right) = Cov\left(u_i, \frac{y_1^* + y_2^* + \ldots + y_i^* + \ldots + y_n^*}{n}\right) =$$

$$= \frac{1}{n} Cov\left(u_i, y_1^* + y_2^* + \ldots + y_i^* + \ldots + y_n^*\right) =$$

$$= \frac{1}{n}\left[Cov\left(u_i, y_1^*\right) + Cov\left(u_i, y_2^*\right) + \ldots + Cov\left(u_i, y_i^*\right) + \ldots + Cov\left(u_i, y_n^*\right)\right] =$$

$$= \frac{1}{n}\left[(n-1)Cov\left(u_i, y_j^*\right) + Cov\left(u_i, y_i^*\right)\right] = \frac{1}{n}\left[(n-1)¼\sigma_u^2 + \sigma_u^2\right] =$$

$$= \frac{\sigma_u^2 + (n-1)¼\sigma_u^2}{n}$$

La matriz **V** sin embargo no es distinta del caso anterior ya que, como fue comentado más arriba, la varianza de la media de los datos de *n* hermanos es la misma independientemente de que uno de ellos vaya a ser valorado genéticamente a partir de ella o no.

El índice de selección queda de esta manera:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \frac{\sigma_u^2 + (n-1)\,¼\,\sigma_u^2}{\sigma_p^2 + (n-1)\,¼\,\sigma_u^2}\,\overline{y}^* =$$

$$= \frac{\left[\sigma_u^2 + (n-1)\,¼\,\sigma_u^2\right] \Big/ \sigma_p^2}{\left[\sigma_p^2 + (n-1)\,¼\,\sigma_u^2\right] \Big/ \sigma_p^2}\,\overline{y}^* = \frac{h^2 + (n-1)\,¼\,h^2}{1 + ¼\,(n-1)h^2}\,\overline{y}^*$$

- Precisión del índice

Para la precisión $\mathbf{C} = \mathbf{C'} = \dfrac{\sigma_u^2 + (n-1)\,¼\,\sigma_u^2}{n}$ y $\mathbf{G} = \sigma_u^2$:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \frac{h^2 + (n-1)\,¼\,h^2}{1 + ¼\,(n-1)h^2}\,\frac{\sigma_u^2 + (n-1)\,¼\,\sigma_u^2}{n}\left(\sigma_u^2\right)^{-1} =$$

$$= \frac{h^2\left[1 + ¼\,(n-1)\right]^2}{n\left[1 + ¼\,(n-1)h^2\right]}$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{\frac{h^2\left[1 + ¼\,(n-1)\right]^2}{n\left[1 + ¼\,(n-1)h^2\right]}} = \frac{h\left[1 + ¼\,(n-1)\right]}{\sqrt{n\left[1 + ¼\,(n-1)h^2\right]}}}$$

La precisión sigue dependiendo de los mismos parámetros, la heredabilidad y el número de datos, aunque la expresión difiere algo al ser uno de los datos de esa media propiedad del individuo evaluado.

VI-4.7. Índice de selección a partir de la media de los datos de *n* hijas

Dado que la inseminación artificial se ha implantado de forma masiva en los esquemas de mejora de rumiantes, este índice es particularmente interesante porque permite estudiar el número de

hijas necesarias para obtener la precisión deseada al valorar un macho en un carácter con una heredabilidad determinada.

- Desarrollo del índice

C': Se desea evaluar al individuo *i* a partir de los datos de sus hijas *j*, *k*, etc.

$$\mathbf{C'} = Cov\,(u_i,\ \bar{y}^*) = Cov\left(u_i, \frac{y_1^* + y_2^* + \ldots + y_n^*}{n}\right) =$$

$$= \frac{1}{n} Cov\left(u_i, y_1^* + y_2^* + \ldots + y_n^*\right) =$$

$$= \frac{1}{n} n\left[Cov\left(u_i, y_j^*\right)\right] = Cov\left(u_i, u_j\right) =$$

$$= Cov\left(u_i, \tfrac{1}{2}u_i + \tfrac{1}{2}u_n + \varphi_j\right) = Var\left(\tfrac{1}{2}u_i\right) = \tfrac{1}{2}\sigma_u^2$$

En esta expresión los subíndices de los datos distinguen a cada una de los hijas del individuo evaluado. Como en casos anteriores todas las covarianzas dentro del corchete se corresponden al mismo valor, en este caso a la covarianza entre el valor genético de un individuo y el dato de hija, por lo que la suma es igual a *n* veces una de ellas.

V: Como en índices previos, hay que desarrollar la matriz de covarianzas entre los datos que componen la media originados por distintos individuos. Al ser todas hijas del mismo macho se trata de medias hermanas, varianza que ya fue desarrollada en un índice previo:

$$\mathbf{V} = \frac{\sigma_p^2 + (n-1)\,\tfrac{1}{4}\sigma_u^2}{n}$$

Y así, el coeficiente del índice de selección sería:

$$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y}^* = \tfrac{1}{2}\sigma_u^2 \frac{n}{\sigma_p^2 + (n-1)\tfrac{1}{4}\sigma_u^2}\bar{y}^* =$$

$$= \frac{\tfrac{1}{2}n\sigma_u^2 \Big/ \sigma_p^2}{\left[\sigma_p^2 + (n-1)\tfrac{1}{4}\sigma_u^2\right] \Big/ \sigma_p^2}\bar{y}^* = \frac{\tfrac{1}{2}nh^2}{1+\tfrac{1}{4}(n-1)h^2}\bar{y}^*$$

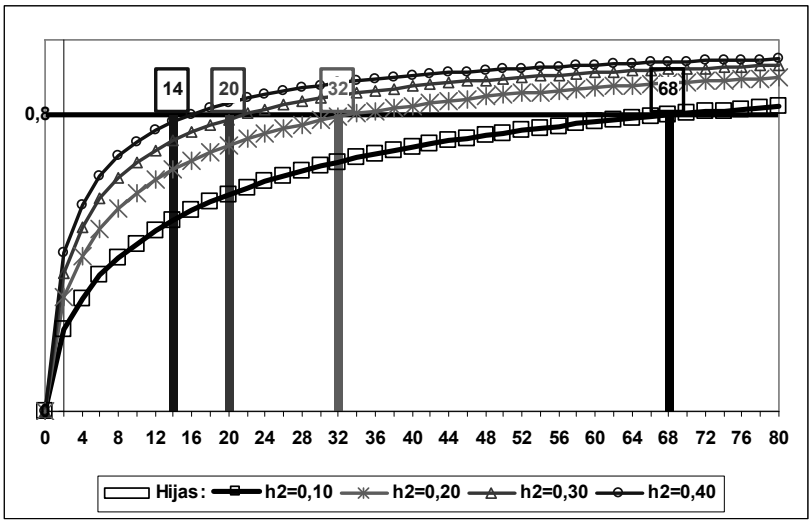
- Precisión del índice

Para la precisión $\mathbf{C} = \mathbf{C'} = \tfrac{1}{2}\sigma_u^2$ y $\mathbf{G} = \sigma_u^2$:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \frac{\tfrac{1}{2}nh^2}{1+\tfrac{1}{4}(n-1)h^2}\tfrac{1}{2}\sigma_u^2\left(\sigma_u^2\right)^{-1} = \tfrac{1}{4}\frac{nh^2}{1+\tfrac{1}{4}(n-1)h^2}$$

$$\boxed{\rho_{\hat{u}u} = \tfrac{1}{2}\sqrt{\frac{nh^2}{1+\tfrac{1}{4}(n-1)h^2}}}$$

Esta expresión es útil por ejemplo en el diseño de un circuito de inseminación artificial ya que se puede aproximar la precisión del valor genético obtenido para un animal en función de la heredabilidad y del número de hijas medidas. En la siguiente figura se ve esta evolución de la precisión en función del número de hijas para cuatro valores comunes de heredabilidad:

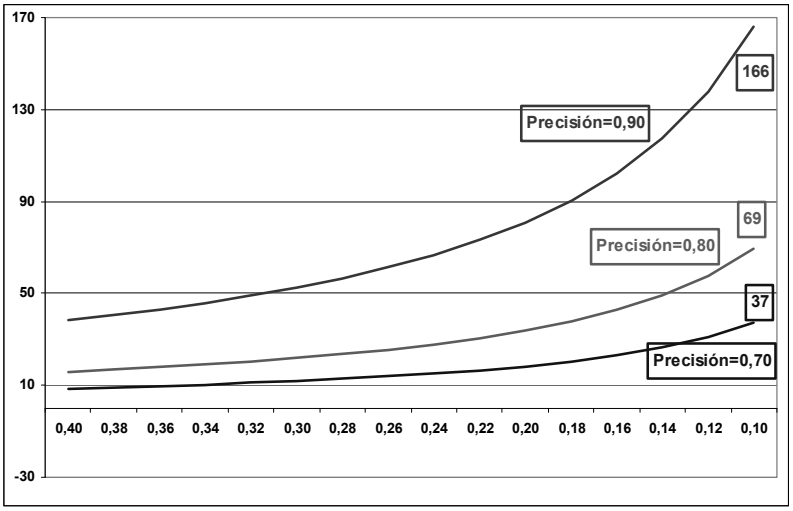


Así por ejemplo, serán necesarias 68 hijas medidas para obtener un macho valorado con una precisión del 80% si la heredabilidad es de 0,10. En cambio, esa misma precisión se logra con tan solo 14 hijas si la heredabilidad es de 0,40.

A partir de la última expresión se puede deducir el número de datos que sería necesario para tener evaluado un determinado macho con una determinada precisión para un cierto valor de heredabilidad:

$$n = \frac{4\rho_{\hat{u}u}^2 - \rho_{\hat{u}u}^2 h^2}{h^2 - \rho_{\hat{u}u}^2 h^2}$$

Podemos también representar el número de hijas necesarias en función de la heredabilidad teniendo en cuenta distintos valores de precisión antes de decidir cuál será el valor que vamos a utilizar para considerar a un animal probado. Así por ejemplo, de acuerdo con esta expresión, para una heredabilidad de 0,10 se necesitarán 166 hijas para tener valorado a un individuo si la precisión exigida es del 90%, pero sólo 37 si nos basta con una precisión del 70%:

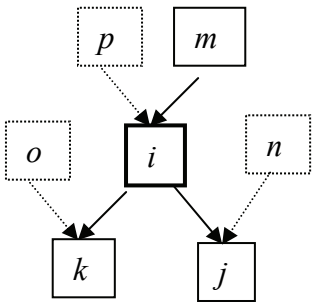


VI-4.8. Índice de selección cuando se utilizan varias fuentes de información

Obviamente la combinación de diferentes criterios de selección en un solo índice abre incontables posibilidades para desarrollar. Como ejemplo se desarrollará a continuación un índice de selección razonablemente complejo con tres criterios:

- El dato del individuo
- El dato de su madre
- La media de *n* hijas

La representación gráfica de las relaciones de parentesco entre los distintos individuos es la siguiente:



En este índice el individuo a evaluar sigue siendo uno sólo pero el vector de fuentes de información ha dejado de ser un único elemento.

- Desarrollo del índice

C'. Ahora **C'** presenta la covarianza entre el valor genético aditivo del individuo y las tres fuentes de información por lo que ahora es un vector fila de 3 elementos:

$$\mathbf{C'} = Cov(u_i, \mathbf{y}^{*\prime}) = Cov\left(u_i, \begin{bmatrix} y_i^* \\ y_m^* \\ \bar{y}^* \end{bmatrix}'\right) =$$

$$= Cov\left(u_i, \begin{bmatrix} y_i^* & y_m^* & \bar{y}^* \end{bmatrix}\right) = \begin{bmatrix} \sigma_{u_i y_i^*} & \sigma_{u_i y_m^*} & \sigma_{u_i \bar{y}^*} \end{bmatrix}$$

Por tanto son tres los elementos que hay que calcular, alguno ya calculado anteriormente:

- $\sigma_{u_i y_i^*} = Cov(u_i, y_i^*) = \sigma_u^2$
- $\sigma_{u_i y_m^*} = Cov(u_i, y_m^*) = Cov(u_i, u_m) = Cov(\frac{1}{2} u_m, u_m) = \frac{1}{2}\sigma_u^2$
- $\sigma_{u_i \bar{y}^*} = Cov\left(u_i, \frac{y_1^* + \ldots + y_n^*}{n}\right) = \frac{1}{n} Cov\left(u_i, y_1^* + \ldots + y_n^*\right) =$

$$= Cov\left(u_i, y_j^*\right) = Cov\left(u_i, \tfrac{1}{2} u_i\right) = \tfrac{1}{2}\sigma_u^2$$

Por tanto, el vector **C'** queda:

$$\mathbf{C'} = Cov(u_i, \mathbf{y}^{*\prime}) = \begin{bmatrix} \sigma_{u_i y_i^*} & \sigma_{u_i y_m^*} & \sigma_{u_i \bar{y}^*} \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_u^2 & \frac{1}{2}\sigma_u^2 & \frac{1}{2}\sigma_u^2 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}\sigma_u^2$$

V'. Ahora se trata de la varianza de un vector, luego será una matriz de varianzas y covarianzas de tamaño 3:

$$\mathbf{V'} = Var(\mathbf{y}^{*\prime}) = Var\begin{bmatrix} y_i^* \\ y_m^* \\ \bar{y}^* \end{bmatrix} = \begin{bmatrix} \sigma^2_{y_i^*} & \sigma_{y_i^* y_m^*} & \sigma_{y_i^* \bar{y}^*} \\ \sigma_{y_m^* y_i^*} & \sigma^2_{y_m^*} & \sigma_{y_m^* \bar{y}^*} \\ \sigma_{\bar{y}^* y_i^*} & \sigma_{\bar{y}^* y_m^*} & \sigma^2_{\bar{y}^*} \end{bmatrix}$$

Deben entonces describirse 9 parámetros, pero dado que la matriz es simétrica, sólo serán necesarios los de la diagonal y 3 elementos de fuera de la diagonal. Para calcular la varianza de la media de los datos de las hijas del individuo, obsérvese que son medio hermanas entre sí, y esta varianza ya la hemos calculado en un índice previo:

- $\sigma^2_{y_i^*} = Var(y_i^*) = \sigma^2_{y_m^*} = Var(y_m^*) = \sigma_p^2$

- $\sigma^2_{\bar{y}^*} = \frac{1}{n^2}\mathbf{1'}\begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & \dots & \sigma_{y_1^* y_n^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & \dots & \sigma_{y_2^* y_n^*} \\ \dots & \dots & \dots & \dots \\ \sigma_{y_n^* y_1^*} & \sigma_{y_n^* y_2^*} & \dots & \sigma^2_{y_n^*} \end{bmatrix}\mathbf{1} = \frac{\sigma_p^2 + (n-1)\,¼\,\sigma_u^2}{n}$

- $\sigma_{y_i^* y_m^*} = \sigma_{y_m^* y_i^*} = Cov(y_i^*, y_m^*) = Cov(u_i, u_m) = ½\sigma_u^2$

- $\sigma_{y_i^* \bar{y}^*} = \sigma_{\bar{y}^* y_i^*} = Cov\left[y_i^*, \frac{1}{n}\left(y_1^* + y_2^* + \dots + y_n^*\right)\right] = Cov(u_i, u_j) = ½\sigma_u^2$

- $\sigma_{y_m^* \bar{y}^*} = \sigma_{\bar{y}^* y_m^*} = \frac{1}{n} Cov(y_m^*, y_1^* + y_2^* + \dots + y_n^*) = Cov(u_m, u_j) =$

$= Cov(u_m, ½u_i) = ½\,Cov(u_m, u_i) = ½\,Cov(u_m, ½u_m) = ¼\sigma_u^2$

Así que la matriz **V** queda finalmente de la siguiente manera:

$$\mathbf{V} = Var\begin{bmatrix} y_i^* \\ y_m^* \\ \overline{y}^* \end{bmatrix} = \begin{bmatrix} \sigma_p^2 & \tfrac{1}{2}\sigma_u^2 & \tfrac{1}{2}\sigma_u^2 \\ \tfrac{1}{2}\sigma_u^2 & \sigma_p^2 & \tfrac{1}{4}\sigma_u^2 \\ \tfrac{1}{2}\sigma_u^2 & \tfrac{1}{4}\sigma_u^2 & \dfrac{\sigma_p^2 + (n-1)\tfrac{1}{4}\sigma_u^2}{n} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & \tfrac{1}{2}h^2 & \tfrac{1}{2}h^2 \\ \tfrac{1}{2}h^2 & 1 & \tfrac{1}{4}h^2 \\ \tfrac{1}{2}h^2 & \tfrac{1}{4}h^2 & \dfrac{1+(n-1)\tfrac{1}{4}h^2}{n} \end{bmatrix} \sigma_p^2$$

Para construir el índice se precisa la inversa de la matriz V por lo que en este texto se dejará indicado. El índice finalmente queda:

$\hat{\mathbf{u}} = \hat{u}_i = \mathbf{C'V^{-1}y^*} =$

$$= \begin{bmatrix} 1 & \tfrac{1}{2} & \tfrac{1}{2} \end{bmatrix} \sigma_u^2 \begin{bmatrix} 1 & \tfrac{1}{2}h^2 & \tfrac{1}{2}h^2 \\ \tfrac{1}{2}h^2 & 1 & \tfrac{1}{4}h^2 \\ \tfrac{1}{2}h^2 & \tfrac{1}{4}h^2 & \dfrac{1+(n-1)\tfrac{1}{4}h^2}{n} \end{bmatrix}^{-1} \frac{1}{\sigma_p^2} \begin{bmatrix} y_i^* \\ y_m^* \\ \overline{y}^* \end{bmatrix} =$$

$$= h^2 \begin{bmatrix} 1 & \tfrac{1}{2} & \tfrac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & \tfrac{1}{2}h^2 & \tfrac{1}{2}h^2 \\ \tfrac{1}{2}h^2 & 1 & \tfrac{1}{4}h^2 \\ \tfrac{1}{2}h^2 & \tfrac{1}{4}h^2 & \dfrac{1+(n-1)\tfrac{1}{4}h^2}{n} \end{bmatrix}^{-1} \begin{bmatrix} y_i^* \\ y_m^* \\ \overline{y}^* \end{bmatrix} =$$

$$= \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} y_i^* \\ y_m^* \\ \overline{y}^* \end{bmatrix} = b_1 y_i^* + b_2 y_m^* + b_3 \overline{y}^*$$

Compruébese a través de las dimensiones de los elementos del índice que el producto de todos los elementos que preceden al

vector de observaciones conduce a un vector fila de tantos elementos como criterios y que ha sido representado en la expresión del índice por **b'**, de forma que $\hat{u}_i$ = **b'y***. Así, aunque los criterios de selección sean más de uno, el predictor del valor genético del individuo es único y se obtiene como una combinación lineal de las fuentes de información que actúan así como regresoras, siendo los elementos del vector **b'** los coeficientes del índice de selección.

- Precisión del índice

Para la precisión **C** = **C'** = $\begin{bmatrix} 1 \\ ½ \\ ½ \end{bmatrix}\sigma_u^2$ y **G** = σ_u^2:

$$\boldsymbol{\rho}^2_{\hat{u}u} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} =$$

$$= h^2 \begin{bmatrix} 1 & ½ & ½ \end{bmatrix} \begin{bmatrix} 1 & ½h^2 & ½h^2 \\ ½h^2 & 1 & ¼h^2 \\ ½h^2 & ¼h^2 & \dfrac{1+(n-1)¼h^2}{n} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ ½ \\ ½ \end{bmatrix} \sigma_u^2 \left(\sigma_u^2\right)^{-1};$$

$$\rho_{\hat{u}u} = \sqrt{h^2 \begin{bmatrix} 1 & ½ & ½ \end{bmatrix} \begin{bmatrix} 1 & ½h^2 & ½h^2 \\ ½h^2 & 1 & ¼h^2 \\ ½h^2 & ¼h^2 & \dfrac{1+(n-1)¼h^2}{n} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ ½ \\ ½ \end{bmatrix}}$$

Nuevamente se ha dejado indicado el resultado por la necesidad de invertir la matriz **V** pero es sencillo razonar por las dimensiones de las matrices que el resultado será un único número, es decir, un escalar. La precisión sigue dependiendo de los mismos parámetros, la heredabilidad y el número de datos. La repetibilidad sólo entrará

en juego cuando entre las fuentes de información aparezcan medidas repetidas sobre un mismo individuo.

VI-4.9. Índice de selección cuando la fuente de información es un carácter diferente

En ocasiones el objetivo de selección es un carácter difícil de medir o presenta una heredabilidad mucho más baja que otro con el que tiene una elevada correlación genética. Se puede hacer entonces selección usando como criterio el segundo de ellos para obtener una respuesta correlacionada en el primero.

Para el desarrollo de este apartado denotaremos al carácter objetivo como *x* y al carácter criterio como *y*.

En esta situación los índices de selección difieren muy ligeramente de los desarrollados hasta este punto. Simplemente, en la matriz **C'** que relaciona los objetivos con los criterios, la varianza genética aditiva del carácter objetivo (aquí denotada como $\sigma^2_{u_x}$) deberá ser reemplazada por la covarianza genética aditiva entre los dos caracteres ($\sigma_{u_x u_y}$). Nótese que ha de conocerse la correlación genética entre los dos caracteres ($r_{u_x u_y}$) para poder obtener dicha covarianza:

$$r_{u_x u_y} = \frac{\sigma_{u_x u_y}}{\sigma_{u_x}\sigma_{u_y}}$$

Veamos dos ejemplos representativos de este apartado.

VI-4.9.1. Índice de selección cuando la fuente de información es el dato del propio individuo en un carácter diferente

Siguiendo la notación comentada para este tipo de índices desarrollaremos sin mucho detalle el índice de selección individual en un carácter correlacionado.

- Índice de selección

$$\hat{\mathbf{u}} = \hat{u}_{i_x} = \mathbf{C'V^{-1}y^*} = \sigma_{u_x u_y} \frac{1}{\sigma^2_{p_y}} y^*_{i_y} = \frac{\sigma_{u_x u_y}}{\sigma^2_{p_y}} y^*_{i_y}$$

- Precisión del índice de selección

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} = \frac{\sigma_{u_x u_y}}{\sigma^2_{p_y}} \sigma_{u_x u_y} \left(\sigma^2_{u_x}\right)^{-1} =$$

$$= \frac{\sigma_{u_x u_y}}{\sigma^2_{p_y}} \sigma_{u_x u_y} \frac{1}{\sigma^2_{u_x}} \frac{\sigma^2_{u_y}}{\sigma^2_{u_y}} = \frac{\left(\sigma_{u_x u_y}\right)^2}{\sigma^2_{u_x}\sigma^2_{u_y}} \frac{\sigma^2_{u_y}}{\sigma^2_{p_y}} = r^2_{u_x u_y} h^2_y$$

$$\boxed{\rho_{\hat{u}u} = \sqrt{r^2_{u_x u_y} h^2_y} = r_{u_x u_y} h_y}$$

En el desarrollo de la precisión se ha añadido en numerador y denominador la varianza genética aditiva del carácter criterio de selección para dejar el resultado expresado en función de parámetros genéticos. La precisión se reduce con respecto a la heredabilidad del carácter criterio de selección en una proporción igual al valor absoluto de la correlación genética entre los dos caracteres. Así, para valores elevados de esta correlación genética e importantes diferencias en la heredabilidad a favor del criterio, un índice de selección a partir de un carácter correlacionado podría ser preferible.

VI-4.9.2 Índice de selección cuando se utilizan varias fuentes de información en un carácter diferente

Utilizaremos el mismo ejemplo desarrollado más arriba pero en un carácter diferente. Las fuentes de información que se utilizan son:

- El dato del individuo
- El dato de su madre

- La media de *n* hijas

Todas ellas medidas en un carácter diferente.

- Desarrollo del índice

Este desarrollo coincide casi enteramente con el mismo caso en el que se utilizan las mismas fuentes de información en el propio carácter, pero con algunas modificaciones. Durante el desarrollo del índice las varianzas genéticas aditivas aparecidas en el desarrollo de **C'** deben sustituirse por covarianzas genéticas entre los dos caracteres, mientras que tanto las varianzas genéticas aditivas como las fenotípicas que aparecen en el desarrollo de **V** son del carácter criterio. El índice queda como sigue:

$$\hat{\mathbf{u}} = \hat{u}_{i_x} = \mathbf{C'V^{-1}y^*} =$$

$$= \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \sigma_{u_x u_y} \begin{bmatrix} 1 & \frac{1}{2}h_y^2 & \frac{1}{2}h_y^2 \\ \frac{1}{2}h_y^2 & 1 & \frac{1}{4}h_y^2 \\ \frac{1}{2}h_y^2 & \frac{1}{4}h_y^2 & \frac{1+(n-1)\frac{1}{4}h_y^2}{n} \end{bmatrix}^{-1} \frac{1}{\sigma_{p_y}^2} \begin{bmatrix} y_{i_y}^* \\ y_{m_y}^* \\ \overline{y}_y^* \end{bmatrix}$$

- Precisión del índice

En la precisión debe tenerse en cuenta además que las varianzas genéticas aditivas que aparecen en la matriz **G** corresponden al carácter objetivo:

$$\boldsymbol{\rho}_{\hat{\mathbf{u}}\mathbf{u}}^{\mathbf{2}} = \left(\mathbf{C'V^{-1}C}\right)\mathbf{G}^{-1} =$$

$$= \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \sigma_{u_x u_y} \begin{bmatrix} 1 & \frac{1}{2}h_y^2 & \frac{1}{2}h_y^2 \\ \frac{1}{2}h_y^2 & 1 & \frac{1}{4}h_y^2 \\ \frac{1}{2}h_y^2 & \frac{1}{4}h_y^2 & \frac{1+(n-1)\frac{1}{4}h_y^2}{n} \end{bmatrix}^{-1} \frac{1}{\sigma_{p_y}^2} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \sigma_{u_x u_y} \left(\sigma_{u_x}^2\right)^{-1}$$

$$\rho_{\hat{u}u} = r_{u_x u_y} h_y \sqrt{\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2}h_y^2 & \frac{1}{2}h_y^2 \\ \frac{1}{2}h_y^2 & 1 & \frac{1}{4}h_y^2 \\ \frac{1}{2}h_y^2 & \frac{1}{4}h_y^2 & \frac{1+(n-1)\frac{1}{4}h_y^2}{n} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}}$$

VI-5. Índices de selección para varios caracteres

Aunque normalmente el interés del mejorador se centra en un carácter concreto, es frecuente que otros caracteres contribuyan al beneficio económico de la explotación. En este contexto es deseable encontrar una combinación lineal de los *n* caracteres objetivo de selección, que pondere adecuadamente la ganancia que se obtiene por el valor genético que cada individuo tiene para cada uno de ellos. Dicha combinación lineal se llama agregado genético económico o genotipo agregado, y se representa por *H*:

$$H = \mathbf{v'u} = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = v_1 u_1 + v_2 u_2 + \dots + v_n u_n$$

En esta expresión los subíndices de los valores genéticos no sirven para identificar individuos sino para distinguir entre caracteres en el objetivo de selección. Los coeficientes de esta combinación lineal presentes en el vector **v'** se llaman pesos económicos y se corresponden con la ganancia en unidad monetaria que se obtiene por el incremento en una unidad de cada carácter del objetivo. Obsérvese que estos coeficientes tienen en cuenta simultáneamente la importancia económica del carácter, su escala y su variabilidad genética. Por ejemplo, los beneficios de una explotación de ovino lechero se obtendrán principalmente por la cantidad de leche vendida, pero el ganadero ingresará también más dinero, por ejemplo, por el porcentaje de grasa en leche y por la

venta de corderos. Obviamente el ganadero gana más por la venta de un cordero que por la venta de un litro de leche, de manera que el peso económico del primero será mucho más elevado que el segundo. Sin embargo, en el agregado genético económico la escala de cada carácter será muy diferente siendo enormemente más bajos los correspondientes al número de corderos, lo que compensa el que sea ponderado por un peso económico mucho mayor. Obsérvese también que los valores genéticos tendrán menor valor absoluto si la variabilidad genética es menor, por lo que estos pesos económicos también tienen en cuenta la variabilidad genética del carácter.

Es frecuente que no todos los caracteres del objetivo de selección puedan medirse con facilidad. En su lugar se registran otros con los que tiene correlación genética. Los caracteres que se recogen y que se introducen en el índice de selección para predecir el valor del genotipo agregado de cada individuo se llaman criterios de selección y pueden ser los propios objetivos, algunos de ellos y otros no, o todos diferentes. Para el desarrollo de este apartado asumiremos que existen *n* caracteres en el objetivo y *m* en el criterio.

Otra cuestión a tener en cuenta es la correlación genética entre los distintos caracteres del objetivo ya que la mejora en uno de ellos podría no compensar por el empeoramiento en otro de mayor valor económico.

Aprovecharemos la teoría ya desgranada de los índices de selección para predecir los valores genéticos de cada carácter. Predeciremos el genotipo agregado $H = \mathbf{v'u}$ a partir de $\hat{H} = \mathbf{v'\hat{u}}$ empleando para ello la expresión ya conocida para resolver los índices de selección $\mathbf{\hat{u} = C'V^{-1}y^{*}}$:

$$\hat{H} = \mathbf{v'\hat{u}} = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \mathbf{v'C'V^{-1}y^*} =$$

$$= \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} \sigma_{u_1 y_1} & \sigma_{u_1 y_2} & .. & \sigma_{u_1 y_m} \\ \sigma_{u_2 y_1} & \sigma_{u_2 y_2} & .. & \sigma_{u_2 y_m} \\ \dots & \dots & .. & \dots \\ \sigma_{u_n y_1} & \sigma_{u_n y_2} & .. & \sigma_{u_n y_m} \end{bmatrix} \begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & .. & \sigma_{y_1^* y_m^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & .. & \sigma_{y_2^* y_m^*} \\ .. & .. & .. & .. \\ \sigma_{y_m^* y_1^*} & \sigma_{y_m^* y_2^*} & .. & \sigma^2_{y_m^*} \end{bmatrix}^{-1} \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_m^* \end{bmatrix}$$

Denotaremos **b'** al vector fila de *m* elementos que contiene el producto de todas las matrices que preceden al vector de criterios de selección:

$$\mathbf{b' = v'C'V^{-1}} =$$

$$= \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} \sigma_{u_1 y_1} & \sigma_{u_1 y_2} & .. & \sigma_{u_1 y_m} \\ \sigma_{u_2 y_1} & \sigma_{u_2 y_2} & .. & \sigma_{u_2 y_m} \\ \dots & \dots & .. & \dots \\ \sigma_{u_n y_1} & \sigma_{u_n y_2} & .. & \sigma_{u_n y_m} \end{bmatrix} \begin{bmatrix} \sigma^2_{y_1^*} & \sigma_{y_1^* y_2^*} & .. & \sigma_{y_1^* y_m^*} \\ \sigma_{y_2^* y_1^*} & \sigma^2_{y_2^*} & .. & \sigma_{y_2^* y_m^*} \\ .. & .. & .. & .. \\ \sigma_{y_m^* y_1^*} & \sigma_{y_m^* y_2^*} & .. & \sigma^2_{y_m^*} \end{bmatrix}^{-1}$$

El índice de selección *I* se obtendrá entonces como el producto de este vector y el de criterios de selección:

$$I = \mathbf{b'y^*} = b_1 y_1^* + b_2 y_2^* + \dots + b_m y_m^*$$

Así, el predictor del genotipo agregado, expresado como una combinación lineal de los predictores de los valores genéticos aditivos para cada uno de los caracteres, se puede expresar también como una combinación lineal de criterios de selección. Si en la primera, los coeficientes que ponderan los valores genéticos se llaman pesos económicos, en la segunda son los coeficientes del índice de selección:

$$\hat{H} = \mathbf{v'\hat{u}} = v_1\hat{u}_1 + v_2\hat{u}_2 + ... + v_n\hat{u}_n = I = \mathbf{b'y}^* = b_1y_1^* + b_2y_2^* + ... + b_my_m^*$$

Se obtiene así un único valor que se puede emplear para seleccionar a los individuos por el valor económico de sus genes. El genotipo agregado tiene así en cuenta la importancia económica de los caracteres presentes en el objetivo de selección, la escala de cada uno de ellos, la variabilidad genética de los caracteres del objetivo, y la correlación que tienen tanto los caracteres del objetivo con los criterios como la que existe entre los distintos criterios de selección.

CONCEPTOS CLAVE

- ¿Cómo se ajustan los efectos fijos al utilizar los índices de selección?
- ¿Con qué método de predicción se identifica la metodología de los índices de selección?
- ¿Por qué un índice de selección es un método sesgado?
- ¿Qué diferencia hay entre objetivos y criterios de selección?
- ¿En qué consiste índice de selección individual? ¿Qué precisión tiene? ¿alta o baja?
- En caso de tener que escoger un dato para evaluar a un individuo que no tiene dato, ¿de qué pariente lo obtendrías?
- ¿Qué parámetros intervienen en la precisión de un índice de selección obtenido a partir de la media de los datos del propio individuo?
- ¿En qué tipo de índices la precisión depende de la repetibilidad del carácter?
- ¿Qué parámetros intervienen en la precisión de un índice de selección obtenido a partir de la media de los hijos de un individuo?
- ¿Cómo se lograría alcanzar la precisión deseada al valorar a un individuo cuando la heredabilidad del carácter es muy baja?
- ¿Qué parámetro genético es preciso conocer cuando la fuente de información del índice es de un carácter diferente al objetivo?
- ¿Qué son los pesos económicos?
- ¿Qué parámetros contribuyen al cálculo de los pesos de un índice de selección?
- ¿Pueden los caracteres objetivo de selección ser diferentes de los caracteres criterio de selección? ¿Pueden ser los mismos?

SÉPTIMA PARTE

VALORACIÓN GENÉTICA ANIMAL MEDIANTE BLUP

RESUMEN

Se detalla la metodología de valoración genética actual más estándar conocida BLUP por sus propiedades estadísticas. Una buena parte del capítulo está destinada al manejo de la información genealógica a través de la matriz numerador de relaciones aditivas conocida más coloquialmente como matriz de parentescos, mostrando cómo se puede evitar el cálculo de su inversa, elemento necesario en la resolución. Se desarrolla un ejemplo numérico con todo detalle interpretando la información reunida en las ecuaciones de resolución del modelo mixto. Se comentan las soluciones obtenidas, tanto para los valores genéticos como para la parte fija del modelo y se muestra también la forma en que pueden ser utilizadas en la práctica. Se concluye el capítulo con la medida de la precisión de los valores genéticos así como su aplicación en el pequeño ejemplo práctico utilizado.

VII-1. Valoración genética mediante BLUP

BLUP son las siglas de (B)est (L)inear (U)nbiased (P)rediction. Es el método empleado en la actualidad para valorar genéticamente prácticamente todas las poblaciones animales a nivel mundial. Las diferencias con respecto a la metodología de los índices de selección son dos:

1. Posee la propiedad de insesgado. Ésta es en realidad la verdadera diferencia teórica con el BLP. La propiedad de insesgado se logra utilizando los estimadores de los efectos fijos como lo que son, estimadores de los efectos fijos, lo que obliga al método a resolver conjuntamente los efectos fijos y los aleatorios de manera que unos sean tenidos en cuenta al resolver los otros. Así por ejemplo, las medias de los datos de las ganaderías son ajustadas para los valores genéticos de los animales, teniendo en cuenta así su valor genético, a diferencia del BLP en el que se asume que los animales de todas las ganaderías poseen el mismo valor genético medio.

2. Utiliza toda la información de parentesco disponible para cada animal. Mientras que en los índices de selección se restringe la información a incluir para valorar a cada animal, en el BLUP se utiliza toda de forma simultánea. El modelo que junto con este hecho, además proporciona un valor genético para todos y cada uno de los animales presentes en el pedigrí, tengan o no tengan dato, se llama Modelo Animal, y se ha impuesto frente a otros modelos equivalentes de menor coste a medida que se ha incrementado la capacidad de los ordenadores.

VII-2. Derivación de las ecuaciones del BLUP

La forma de obtener las ecuaciones de resolución del BLUP consiste en aplicar las propiedades que por definición debe cumplir este predictor:

- Ha de ser lineal en las observaciones. Es decir, debe poder obtenerse como combinación lineal de los datos. Si llamamos **L'** a la matriz que define esa combinación lineal, el predictor será de la forma **L'y**.

- Ha de ser insesgado. Es decir, el valor esperado del predictor (**L'y**) debe coincidir con el valor esperado de lo que se desea de predecir (**K'b** + **M'u**):

$$E(\mathbf{L'y}) = E(\mathbf{K'b} + \mathbf{M'u})$$

Haciendo de uso de la definición de las esperanzas del modelo tenemos:

$$\mathbf{L'Xb} = \mathbf{K'b}$$
$$\mathbf{L'Xb} - \mathbf{K'b} = 0$$
$$(\mathbf{L'X} - \mathbf{K'})\mathbf{b} = 0$$

De manera que $(\mathbf{L'X} - \mathbf{K'})$ representa el sesgo.

- Ha de ser el mejor, es decir, ha de tener mínima varianza del error de predicción.

(**K'b** + **M'u**) es lo que se desea predecir y **L'y** es el predictor, así que la diferencia entre ambos es el error de predicción:

$$EP = (\mathbf{K'b} + \mathbf{M'u}) - \mathbf{L'y}$$

Y la varianza del error de predicción:

$$V(EP) = Var(\mathbf{K'b} + \mathbf{M'u} - \mathbf{L'y}) = Var(\mathbf{M'u} - \mathbf{L'y}) =$$
$$= \mathbf{M'}V(\mathbf{u})\mathbf{M} + \mathbf{L'}V(\mathbf{y})\mathbf{L} - \mathbf{M'}Cov(\mathbf{u},\mathbf{y'})\mathbf{L} - \mathbf{L'}Cov(\mathbf{y},\mathbf{u'})\mathbf{M} =$$
$$= \mathbf{M'GM} + \mathbf{L'VL} - \mathbf{M'GZ'L} - \mathbf{L'ZGM}$$

Obsérvese que la matriz $Cov(\mathbf{u},\mathbf{y'})$ identificada como **C'** en el contexto de los índices de selección, es aquí equivalente a

GZ', tal y como quedó definida esta covarianza en la definición del modelo.

Finalmente, la combinación de sesgo y mínima varianza del error de predicción deben combinarse en una sola función a minimizar con la ayuda de un nuevo parámetro que se convierte en incógnita que permita poner sesgo y varianza en la misma escala. Esta nueva incógnita se llama multiplicador de LaGrange que denotaremos por **Φ**, y llamando *F* a la función que combina sesgo y varianza del error de predicción, obtenemos la siguiente función que habrá de ser minimizada:

$$F = \mathrm{V(EP)} + (\mathbf{L'X} - \mathbf{K'})\mathbf{\Phi}$$

DERIVACIÓN DE LAS ECUACIONES DEL MODELO MIXTO DE HENDERSON

La forma de obtener el mínimo de la función *F* es derivar la función respecto a los parámetros desconocidos **L** y **Φ** e igualar las derivadas a matrices nulas:

$$\delta F / \delta \mathbf{L} = 2\mathbf{VL} - 2\,\mathbf{ZGM} + \mathbf{X\Phi} = \mathbf{0}$$
$$\delta F / \delta \mathbf{\Phi} = \mathbf{X'L} - \mathbf{K} = \mathbf{0}$$

Se define $\boldsymbol{\theta} = \frac{1}{2}\,\mathbf{\Phi}$, y se simplifica:

$$\mathbf{VL} + \mathbf{X\theta} = \mathbf{ZGM}$$
$$\mathbf{X'L} = \mathbf{K}$$

Estas ecuaciones pueden ser escritas en forma matricial de la siguiente manera:

$$\begin{bmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X'} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{ZGM} \\ \mathbf{K} \end{bmatrix}$$

Se sustituyen **V** por su valor en la definición del modelo (**V** = **ZGZ'** + **R**):

$$\begin{bmatrix} \mathbf{ZGZ'} + \mathbf{R} & \mathbf{X} \\ \mathbf{X'} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{L} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{ZGM} \\ \mathbf{K} \end{bmatrix}$$

Estas ecuaciones se pueden reescribir. Empezaremos desarrollando el producto de los dos bloques superiores por el vector de incógnitas e igualado al lado derecho correspondiente. Los pasos que siguen a continuación son algebraicamente sencillos:

$$(\mathbf{ZGZ'} + \mathbf{R})\mathbf{L} + \mathbf{X\theta} = \mathbf{ZGM}$$

$$\mathbf{ZGZ'L} + \mathbf{RL} + \mathbf{X\theta} = \mathbf{ZGM}$$

$$\mathbf{ZGZ'L} - \mathbf{ZGM} + \mathbf{RL} + \mathbf{X\theta} = \mathbf{0}$$

$$\mathbf{Z}(\mathbf{GZ'L} - \mathbf{GM}) + \mathbf{RL} + \mathbf{X\theta} = \mathbf{0}$$

$$\mathbf{Z}\left[\mathbf{G}(\mathbf{Z'L} - \mathbf{M})\right] + \mathbf{RL} + \mathbf{X\theta} = \mathbf{0}$$

Definimos la matriz **S** como **S** = **G**(**Z'L** - **M**), de donde se despeja **M** como **M** = **Z'L** - **G**$^{-1}$**S**. Las ecuaciones anteriores quedan ahora. La ecuación previa, más las que quedaban del bloque anterior, e incorporando esta definición de **M**, se pueden escribir conjuntamente como:

$$\begin{bmatrix} \mathbf{R} & \mathbf{X} & \mathbf{Z} \\ \mathbf{X'} & \mathbf{0} & \mathbf{0} \\ \mathbf{Z'} & \mathbf{0} & -\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{L} \\ \mathbf{\theta} \\ \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{K} \\ \mathbf{M} \end{bmatrix}$$

Del primer grupo de ecuaciones se obtiene $\mathbf{ZS} + \mathbf{RL} + \mathbf{X\theta} = \mathbf{0}$. Y despejando **L** se obtiene:

$$\mathbf{L} = -\mathbf{R}^{-1}\mathbf{X\theta} - \mathbf{R}^{-1}\mathbf{ZS}$$

Los otros dos bloques de ecuaciones quedan:

$$\mathbf{X'L} = \mathbf{K}$$

$$\mathbf{Z'L} - \mathbf{G}^{-1}\mathbf{S} = \mathbf{M}$$

Y sustituyendo el valor de **L** en estas dos ecuaciones:

$$\mathbf{X'}\left(-\mathbf{R}^{-1}\mathbf{X\theta} - \mathbf{R}^{-1}\mathbf{ZS}\right) = \mathbf{K}$$

$$\mathbf{Z'}\left(-\mathbf{R}^{-1}\mathbf{X\theta} - \mathbf{R}^{-1}\mathbf{ZS}\right) - \mathbf{G}^{-1}\mathbf{S} = \mathbf{M}$$

Y reordenando las ecuaciones adecuadamente nos queda:

$$\mathbf{X'R}^{-1}\mathbf{X\theta} + \mathbf{X'R}^{-1}\mathbf{ZS} = -\mathbf{K}$$

$$\mathbf{Z'R}^{-1}\mathbf{X\theta} + \mathbf{Z'R}^{-1}\mathbf{ZS} + \mathbf{G}^{-1}\mathbf{S} = -\mathbf{M}$$

Reagrupando las ecuaciones en un sistema nos quedan:

$$\begin{bmatrix} \mathbf{X'R}^{-1}\mathbf{X} & \mathbf{X'R}^{-1}\mathbf{Z} \\ \mathbf{Z'R}^{-1}\mathbf{X} & \mathbf{Z'R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{\theta} \\ \mathbf{S} \end{bmatrix} = \begin{bmatrix} -\mathbf{K} \\ -\mathbf{M} \end{bmatrix}$$

Definamos la inversa generalizada de la matriz de coeficientes de la expresión anterior como una matriz **C** cuyos bloques correspondientes a los bloques originales son definidos por los subíndices:

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

Entonces se puede despejar :

$$\begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{S} \end{bmatrix} = -\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{K} \\ \mathbf{M} \end{bmatrix}$$

Podemos ahora volver sobre el valor de **L** presentado más arriba como **L** = - $\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\theta}$ - $\mathbf{R}^{-1}\mathbf{Z}\mathbf{S}$:

$$\mathbf{L} = -\begin{bmatrix} \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1}\mathbf{Z} \end{bmatrix}\begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{S} \end{bmatrix}$$

$$\mathbf{L} = -\begin{bmatrix} \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1}\mathbf{Z} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{K} \\ \mathbf{M} \end{bmatrix}$$

Una vez conocido el valor de **L** recordemos que el BLUP de **K'b** + **M'u** es **L'y**, y entonces:

$$\mathbf{L}' = \left\{-\begin{bmatrix} \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1}\mathbf{Z} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{K} \\ \mathbf{M} \end{bmatrix}\right\}' = \begin{bmatrix} \mathbf{K}' & \mathbf{M}' \end{bmatrix}\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix}$$

$$\mathbf{L}'\mathbf{y} = \begin{bmatrix} \mathbf{K}' & \mathbf{M}' \end{bmatrix}\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Si $\mathbf{L}'\mathbf{y} = \mathbf{K}'\hat{\mathbf{b}} + \mathbf{M}'\hat{\mathbf{u}}$ entonces: $\mathbf{L}'\mathbf{y} = \begin{bmatrix} \mathbf{K}' & \mathbf{M}' \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix}$, por lo que:

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Y recordando que **C** es la inversa de la matriz definida algo más arriba, se pueden estimar los efectos fijos **b** y predecir los aleatorios **u** simultáneamente resolviendo el sistema siguiente:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Esta expresión se conoce con el nombre de ecuaciones del modelo mixto de Henderson siendo en la actualidad esta nomenclatura utilizada como sinónimo de BLUP.

Los predictores BLUP de los efectos aleatorios y los estimadores BLUE de los fijos, son entonces obtenidos resolviendo este sistema de ecuaciones:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

Estas ecuaciones, enormemente utilizadas se deben a Henderson y se conocen por ello como ecuaciones del modelo mixto de Henderson o MME (*Mixed Model Equations*).

Debe entonces completarse la matriz de coeficientes del lado izquierdo, el vector del lado derecho o términos independientes, y resolver invirtiendo la matriz de coeficientes y multiplicando por el vector del lado derecho. Obsérvese que todos los elementos de este sistema de ecuaciones son conocidos. **X** y **Z** son las matrices de incidencia o matrices diseño de los efectos fijos y aleatorios, matrices de ceros y unos (excepto para variables regresoras o covariables), **y** es el vector de datos que contiene simplemente las observaciones y **R** y **G** son las matrices de varianzas y covarianzas que corresponden respectivamente a los efectos residuales y los efectos genéticos aditivos. Ambas aparecen en forma invertida, pero **R** es sencilla de eliminar de acuerdo con la definición más habitual del modelo, mientras que la inversa de **G** requiere un estudio algo más detallado.

VII-3. Ecuaciones simplificadas del BLUP

Durante la definición del modelo lineal mixto normalmente se asume homogeneidad de varianza residual e independencia entre residuos. En esta situación $\mathbf{R} = \mathbf{I}\sigma_e^2$, y por tanto su inversa es:

$$\mathbf{R}^{-1} = \mathbf{I}\frac{1}{\sigma_e^2}$$

De este modo, multiplicando la matriz de coeficientes y el vector del lado derecho por σ_e^2, se mantendrá la igualdad y desaparecerá la inversa de **R,** a excepción de donde encontramos $\mathbf{G^{-1}}$. Según la definición del modelo $\mathbf{G} = \mathbf{A}\sigma_u^2$ siendo **A** la matriz numerador de

relaciones aditivas que desarrollaremos después y σ_u^2 la varianza genética aditiva del carácter. En esta posición quedará finalmente:

$$\mathbf{G}^{-1}\sigma_e^2 = \mathbf{A}^{-1}\frac{1}{\sigma_u^2}\sigma_e^2 = \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} = \mathbf{A}^{-1}\alpha$$

Obsérvese que en esta expresión se ha igualado alfa al cociente de varianzas: $\alpha = \frac{\sigma_e^2}{\sigma_u^2}$.

Finalmente las ecuaciones simplificadas del BLUP son las siguientes:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

VII-4. Ejemplo para aplicación de la metodología BLUP

A continuación se va a desarrollar la metodología BLUP con ayuda de un sencillo ejemplo. En el mismo se desea valorar genéticamente a seis individuos de los que sólo cinco poseen dato propio. Sus registros se encuentran distribuidos en dos niveles distintos de un mismo efecto fijo, el efecto ganadería. Los datos son los siguientes:

Animal	Padre	Madre	Ganadería	Peso
1	-	-	-	-
2	-	-	1	415
3	-	2	1	430
4	1	2	2	420
5	3	2	2	400
6	3	4	2	405

Prácticamente todos los elementos que aparecen en las ecuaciones del modelo mixto de Henderson pueden ser deducidas a partir de la ecuación del modelo correspondiente a este ejemplo:

$$\begin{bmatrix} 415 \\ 430 \\ 420 \\ 400 \\ 405 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\ \mathbf{b} + \mathbf{Z}\ \mathbf{u} + \mathbf{e}$$

En la ecuación del modelo aparecen definidas las matrices **X** y **Z**, así como el vector de datos **y**. Únicamente desconocemos el contenido de la inversa de la matriz **A**, matriz denominada como matriz de relaciones aditivas en el momento de definir el modelo. Estrictamente debe llamarse matriz numerador de relaciones aditivas o NRM (*Numerator Relationship Matrix*). Su estructura contiene la aportación de los parentescos entre individuos a las valoraciones genéticas por lo que su papel es clave y merece un estudio más detallado.

VII-5. La matriz numerador de relaciones aditivas (A)

Recordemos que **A** se define como la matriz que define la estructura de la matriz de varianzas y covarianzas del vector de efectos genéticos aditivos **G** de manera que, siendo σ_u^2 la varianza genética aditiva del carácter:

$$\mathbf{G} = Var(\mathbf{u}) = \mathbf{A}\sigma_u^2 \qquad \text{y} \qquad \mathbf{A} = \frac{\mathbf{G}}{\sigma_u^2} = \frac{Var(\mathbf{u})}{\sigma_u^2}$$

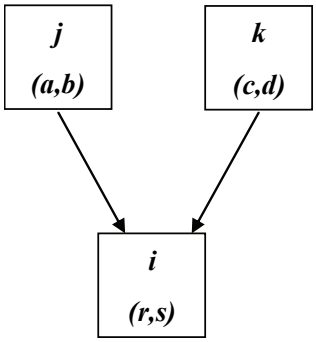
De esta manera, la relación entre los elementos a_{ii} de la diagonal y a_{jk} de fuera de la diagonal de **A** y los correspondientes elementos de **G** es la siguiente:

$$a_{ii} = \frac{g_{ii}}{\sigma_u^2} = \frac{\sigma_{u_i}^2}{\sigma_u^2} \qquad \text{y} \qquad a_{jk} = \frac{g_{jk}}{\sigma_u^2} = \frac{\sigma_{u_j u_k}}{\sigma_u^2}$$

Comenzaremos por hacer unas consideraciones previas necesarias sobre conceptos que no han sido desarrollados previamente.

VII-5.1. El coeficiente de consanguinidad de un individuo

Un individuo *j* que posee dos alelos *a* y *b* para un hipotético gen se cruza con otro *k* de sexo contrario que posee dos alelos *c* y *d* para ese mismo gen. Un descendiente suyo *i* posee para ese mismo gen dos alelos, *r* heredado de *j* y *s* heredado de *k*:



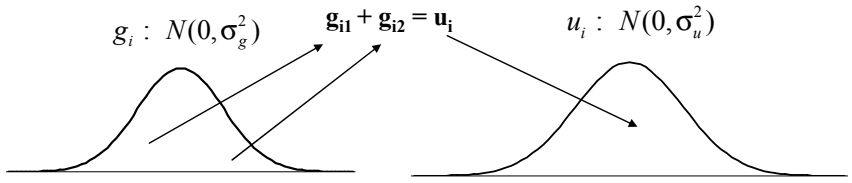
Se define el coeficiente de consanguinidad F_i del individuo *i* como la probabilidad de que los dos alelos del individuo i sean idénticos por descendencia ($F_i = P(r = s)$). En concreto *r* ha sido heredado de *j* por lo que tendrá una probabilidad de ½ de ser el alelo *a* y otro ½ de ser el alelo *b*. Igualmente *s* es con igual probabilidad de ½ el alelo *c* o el *d*. Por tanto tenemos cuatro posibilidades de alelos en i, todos con probabilidad ¼, que los alelos de *i* sean *a* y *c*, *a* y *d*, *b* y *c*, o *b* y *d*. La consanguinidad de *i* queda:

$$F_i = P(r = s) = ¼ \, [P(a = c) + P(a = d) + P(b = c) + P(b = d)]$$

Así para individuos con padres conocidos la consanguinidad se originará si los padres tienen algún tipo de parentesco entre sí, es decir, si tienen algún antepasado en común. En este caso existe una determinada probabilidad de que hayan heredado el mismo alelo ambos padres y lo hayan transmitido al hijo. Para los individuos fundadores se asumirá que poseen dos alelos diferentes entre sí y también diferentes de cualquier otro alelo de cualquier individuo de la población fundadora.

VII-5.2. La distribución de valores genéticos aditivos y la distribución de alelos

Hasta aquí se ha considerado el valor genético de un individuo como un elemento indivisible, pero cuando se trata de estudiar el flujo de alelos, es preciso recordar que la parte aditiva del genotipo de un individuo es la suma de los dos alelos que lo componen. Definiremos entonces una distribución de alelos con media cero y varianza la varianza alélica (σ_g^2) de la que se extraen aleatoriamente dos alelos para formar individuos fundadores. Estos individuos fundadores pasan a formar parte de otra distribución, también de media cero y de varianza la varianza genética aditiva (σ_u^2):



Obsérvese que los individuos con al menos un padre conocido no pertenecen a esta distribución. Así, si la población está sometida a selección, ni la media ni la varianza se mantendrán constantes, por lo que un individuo no fundador no tendrá por qué pertenecer a esta distribución.

VII-5.3. Los elementos de A

Tomaremos como referencia un individuo *i* cuyo valor genético aditivo u_i es la suma de sus dos alelos *r* y *s*. Ya que $\mathbf{A} = \frac{\mathbf{G}}{\sigma_u^2} = \frac{Var(\mathbf{u})}{\sigma_u^2}$, un elemento a_{ii} de la diagonal de **A** será de la forma:

$$a_{ii} = \frac{\sigma_{u_i}^2}{\sigma_u^2} = \frac{Var(u_i)}{\sigma_u^2} = \frac{Var(r+s)}{\sigma_u^2} = \frac{Var(r)+Var(s)+2Cov(r,s)}{\sigma_u^2}$$

Dado que r y s son dos alelos extraídos aleatoriamente de la distribución de alelos, su varianza es la varianza de la distribución, la varianza alélica (σ_g^2). En cuanto a la covarianza entre ambos, si se trata del mismo alelo, entonces será la varianza de uno de ellos y por tanto también la varianza alélica, pero si son distintos, entonces serán extracciones independientes de la distribución y por tanto la covarianza será nula. Ante el desconocimiento de si se trata o no del mismo alelo, esta covarianza se obtendrá mediante el producto de ambos elementos, la varianza alélica y la probabilidad de que sean el mismo ($P(r = s)\,\sigma_g^2$). Así si se trata del mismo alelo la probabilidad vale 1 y la covarianza es la varianza alélica, mientras que si se trata de alelos distintos esa probabilidad vale cero y la covarianza también. Si además tenemos en cuenta que la probabilidad de que estos dos alelos sean idénticos ha sido ya definida como el coeficiente de consanguinidad F_i del individuo *i*, tenemos:

$$a_{ii} = \frac{\sigma_g^2 + \sigma_g^2 + 2P(r=s)\sigma_g^2}{\sigma_u^2} = \frac{2\sigma_g^2 + 2\sigma_g^2 F_i}{\sigma_u^2} = \frac{2\sigma_g^2(1+F_i)}{\sigma_u^2}$$

Necesitamos ahora conocer la relación entre la varianza alélica y la varianza genética aditiva. Recurriremos para ello al valor genético aditivo de un individuo fundador ya que de él sabemos que pertenece a la distribución de valores genéticos aditivos y por tanto, su varianza es la de su distribución. Sabemos también de él

que tiene dos alelos completamente independientes y por tanto, su consanguinidad es cero:

Si *i* es fundador: $Var(u_i) = \sigma_u^2 = 2\sigma_g^2 + 2\sigma_g^2 F = 2\sigma_g^2(1+F_i) = 2\sigma_g^2$.
Puesto que la varianza del valor genético de este individuo es la varianza genética aditiva y también el doble de la varianza alélica, ambos valores son iguales:

$$\boxed{\sigma_u^2 = 2\sigma_g^2}$$

VII-5.3.1. Los elementos de la diagonal de A

Podemos ahora recurrir a la expresión anterior para encontrar una expresión general para los elementos de la diagonal de **A**:

$$a_{ii} = \frac{2\sigma_g^2(1+F_i)}{\sigma_u^2} = \frac{\sigma_u^2(1+F_i)}{\sigma_u^2} = 1 + F_i$$

Por tanto, los elementos de la diagonal de A son igual a uno más la consanguinidad del individuo. Aún necesitamos conocer la consanguinidad del individuo para completar el elemento de la diagonal. Para ello desarrollaremos la covarianza entre los valores genéticos aditivos de sus padres *j* y *k*, teniendo en cuenta que se deben a la suma de los valores de sus alelos:

$$\sigma_{u_j u_k} = Cov(u_j, u_k) = Cov(a+b, c+d) =$$
$$= Cov(a,c) + Cov(a,d) + Cov(b,c) + Cov(b,d) =$$

Se ha descompuesto la covarianza entre dos sumas en la suma de las cuatro correspondientes covarianzas entre alelos. Procedemos ahora a descomponer la covarianza alélica como previamente en el producto de la varianza alélica y la probabilidad de que ambos alelos sean el mismo:

$$\sigma_{u_j u_k} = P(a=c)\sigma_g^2 + P(a=d)\sigma_g^2 + P(b=c)\sigma_g^2 + P(b=d)\sigma_g^2 =$$
$$= \left[P(a=c) + P(a=d) + P(b=c) + P(b=d)\right]\sigma_g^2$$

Entre corchetes se encuentra la misma suma de probabilidades que aparece en la definición del coeficiente de consanguinidad dividida por cuatro. Así pues, este corchete es cuatro veces la consanguinidad. Además sabemos que la varianza alélica (σ_g^2) es la mitad de la varianza genética aditiva (σ_u^2). Sustituyendo se obtiene:

$$\sigma_{u_j u_k} = 4F_i \frac{\sigma_u^2}{2} = 2F_i\sigma_u^2 \Rightarrow F_i = \tfrac{1}{2}\frac{\sigma_{u_j u_k}}{\sigma_u^2} = \tfrac{1}{2}a_{jk}$$

En esta expresión se ha tenido en cuenta que la covarianza entre los valores genéticos de dos individuos *j* y *k* dividida por la varianza genética aditiva del carácter es precisamente un elemento de la matriz **A**, el que corresponde al cruce de los padres de *i*. Por tanto, para obtener un elemento de la diagonal de **A** se calculará:

$$\boxed{a_{ii} = 1 + F_i = 1 + \tfrac{1}{2}a_{jk}}$$

Este valor será igual a 1 más la consanguinidad, y ésta, si no se conocen ambos padres del individuo vale cero, y si se conocen es igual a la mitad del elemento de la propia matriz **A** donde cruzan sus padres. Aún nos queda sin embargo conocer cómo se obtienen los elementos de fuera de la diagonal de la matriz **A**.

VII-5.3.2. Los elementos de fuera de la diagonal de A

Para afrontar el cálculo de los elementos de fuera de la diagonal de **A** introducimos ahora un nuevo individuo, el individuo *h*, no siendo *h* más joven que *i*. Calcularemos la covarianza de los valores genéticos aditivos de *h* e *i* teniendo en cuenta la descomposición del valor genético aditivo de *i* en función de la media de los valores genéticos aditivos de sus padres y un muestreo mendeliano:

$$\sigma_{u_h u_i} = Cov(u_h, u_i) = Cov\left(u_h, \tfrac{1}{2}u_j + \tfrac{1}{2}u_k + \varphi_i\right) =$$
$$= Cov\left(u_h, \tfrac{1}{2}u_j\right) + Cov\left(u_h, \tfrac{1}{2}u_k\right) + Cov(u_h, \varphi_i) =$$
$$= \tfrac{1}{2}Cov(u_h, u_j) + \tfrac{1}{2}Cov(u_h, u_k) = \tfrac{1}{2}\sigma_{u_h u_j} + \tfrac{1}{2}\sigma_{u_h u_k}$$

Y dividiendo a ambos lados de la igualdad para encontrar elementos de **A**:

$$\frac{\sigma_{u_h u_i}}{\sigma_u^2} = \tfrac{1}{2}\frac{\sigma_{u_h u_j}}{\sigma_u^2} + \tfrac{1}{2}\frac{\sigma_{u_h u_k}}{\sigma_u^2}$$

$$\boxed{a_{hi} = \tfrac{1}{2}a_{hj} + \tfrac{1}{2}a_{hk}}$$

Además, la matriz **A** es una matriz simétrica por lo que $a_{ih} = a_{hi}$, es decir, el parentesco de *h* con *i* es igual al de *i* con *h*, o dicho de otra manera, el porcentaje de genes que comparte *i* con *h* es igual al que comparte *h* con *i*.

Por tanto, para obtener un elemento de fuera de la diagonal de la matriz de relaciones aditivas se calculará la media de otros valores de la misma fila que están en las columnas de los padres. Si alguno o los dos padres son desconocidos, el elemento de la columna del padre o padres desconocidos tendrá o tendrán valor nulo. Es importante hacer notar que la descomposición en función de los padres debe hacerse para el individuo más joven, ya que en este trazado retrospectivo de genes se pretende localizar los antepasados comunes de cada dos individuos, y éstos no se localizarían aunque existiesen si el desglose se hiciera en el individuo más viejo.

MATRIZ DE RELACIONES ADITIVAS Y MATRIZ NUMERADOR DE RELACIONES ADITIVAS

Los coeficientes de la matriz de relaciones aditivas representan el porcentaje de genes que comparten los dos individuos que se cruzan en esa posición, por lo que suponen una medida de parentesco entre ellos. Tal y como se han definido aquí los coeficientes de **A**, algunos elementos pueden presentar valores superiores al 100%. Este hecho se observa claramente en la diagonal de los individuos

consanguíneos cuyo valor excede a uno. Efectivamente, cualquier alelo presente en un individuo aparece representado en sí mismo, por lo que este coeficiente no puede ser inferior al 100%. Pero es que además, en un individuo consanguíneo existe una cierta probabilidad de que aparezca repetido en el otro alelo del individuo. Para corregir este efecto los valores obtenidos aquí deben ser ajustados para la consanguinidad de los dos individuos que se cruzan en cada posición. Por ello a la matriz A presentada aquí se la llama Matriz Numerador de Relaciones Aditivas (NRM). Los elementos de la Matriz de Relaciones Aditivas **A*** se obtienen a partir de los de la matriz **A** de la siguiente manera:

$$a^*_{jk} = \frac{a_{jk}}{\sqrt{1+F_j}\sqrt{1+F_k}}$$

En cualquier caso en este texto en el contexto de la valoración genética a la matriz **A** se le denominará Matriz de Relaciones Aditivas. Con frecuencia es denominada también Matriz de Parentescos, aunque existen otras medidas de parentesco de no menor importancia que los coeficientes de **A**.

VII-5.3.3. El método tabular de construcción de la matriz A.

La construcción de la matriz de relaciones aditivas puede realizarse mediante el método tabular. Haremos uso para ello de las expresiones vistas en el apartado anterior para construir los elementos de la diagonal y de fuera de la diagonal de la matriz **A**:

$$a_{ii} = 1 + F_i = 1 + \tfrac{1}{2} a_{jk}$$
$$a_{hi} = \tfrac{1}{2} a_{hj} + \tfrac{1}{2} a_{hk}$$

En el método tabular el orden de construcción es importante y debe seguirse escrupulosamente. Los pasos que hay que seguir son los siguientes:

1. Si en la genealogía existen *n* animales, crear una tabla con *n* x *n* celdas, en el ejemplo nuestro una tabla con 6 filas y 6 columnas.

2. Comenzar cada fila por el elemento de la diagonal con la primera expresión, la que sirve para elementos de la diagonal. En cada elemento de la diagonal, si no se conocen los dos padres del individuo la consanguinidad

vale cero y si se conocen los dos padres, entonces se obtendría multiplicando por ½ el elemento de la propia matriz **A** donde se cruzan los padres del individuo. Este elemento debe haber sido ya calculado si el método se sigue apropiadamente.

3. Completar la fila de la matriz utilizando ahora la expresión para elementos de fuera de la diagonal. Si observamos la expresión, se trata de hacer la media de otros dos coeficientes de la misma matriz **A** que tienen el mismo primer subíndice que el que se desea calcular, es decir, que están en la misma fila, y cuyo segundo subíndice coincide con los padres del individuo señalado por el segundo subíndice del que se desea calcular. En otras palabras, se anota en esta posición la media de los dos elementos que hay en la misma fila para las columnas de los padres. Si alguno de los padres no es conocido, ese elemento no existiría en la matriz y se bastaría con multiplicar el otro por ½. Una vez deducida la regla, es aconsejable señalar encima de cada individuo la información de sus padres para localizar las columnas rápidamente.

4. Terminada la fila debe copiarse como columna ya que se trata de una matriz simétrica.

5. Comenzar nuevamente en un elemento diagonal y proceder como en el paso 2.

Siguiendo la genealogía de nuestro ejemplo construiremos la matriz **A** calculando los elementos exactamente en este orden:

Animal	Padre	Madre
1	-	-
2	-	-
3	-	2
4	1	2
5	3	2
6	3	4

Elemento Diagonal de la fila 1:

$$a_{11} = 1 + F_1 = 1 + \tfrac{1}{2} a_{00} = 1$$

Resto de la fila 1:

$$a_{12} = \tfrac{1}{2} a_{10} + \tfrac{1}{2} a_{10} = 0 \qquad a_{13} = \tfrac{1}{2} a_{10} + \tfrac{1}{2} a_{12} = 0$$

$$a_{14} = \tfrac{1}{2} a_{11} + \tfrac{1}{2} a_{10} = \tfrac{1}{2} \qquad a_{15} = \tfrac{1}{2} a_{13} + \tfrac{1}{2} a_{12} = 0$$

$$a_{16} = \tfrac{1}{2} a_{13} + \tfrac{1}{2} a_{14} = \tfrac{1}{4}$$

Se copia la fila 1 como columna 1:

$$a_{21} = a_{12} = 0; \qquad a_{31} = a_{13} = 0; \qquad a_{41} = a_{14} = \tfrac{1}{2};$$

$$a_{51} = a_{15} = 0; \qquad a_{51} = a_{15} = 0;$$

Elemento Diagonal de la fila 2:

$$a_{22} = 1 + F_2 = 1 + \tfrac{1}{2} a_{00} = 1$$

Resto de la fila 2:

$$a_{23} = \tfrac{1}{2} a_{20} + \tfrac{1}{2} a_{22} = \tfrac{1}{2} \qquad a_{24} = \tfrac{1}{2} a_{21} + \tfrac{1}{2} a_{22} = \tfrac{1}{2}$$

$$a_{25} = \tfrac{1}{2} a_{23} + \tfrac{1}{2} a_{22} = \tfrac{3}{4} \qquad a_{26} = \tfrac{1}{2} a_{23} + \tfrac{1}{2} a_{24} = \tfrac{1}{2}$$

Se copia la fila 2 como columna 2:

$$a_{32} = a_{23} = \tfrac{1}{2}; \qquad a_{42} = a_{24} = \tfrac{1}{2}; \qquad a_{52} = a_{25} = \tfrac{3}{4};$$

$$a_{62} = a_{26} = \tfrac{1}{2}$$

Elemento Diagonal de la fila 3:

$$a_{33} = 1 + F_3 = 1 + \tfrac{1}{2} a_{02} = 1$$

Resto de la fila 3:

$$a_{34} = \tfrac{1}{2} a_{31} + \tfrac{1}{2} a_{32} = \tfrac{1}{4} \qquad a_{35} = \tfrac{1}{2} a_{33} + \tfrac{1}{2} a_{32} = \tfrac{3}{4}$$

$$a_{36} = \tfrac{1}{2} a_{33} + \tfrac{1}{2} a_{34} = \tfrac{5}{8}$$

Se copia la fila 3 como columna 3:

$$a_{43} = a_{34} = \tfrac{1}{4}; \qquad a_{53} = a_{35} = \tfrac{3}{4}; \qquad a_{63} = a_{36} = \tfrac{5}{8}$$

Elemento Diagonal de la fila 4:

$$a_{44} = 1 + F_4 = 1 + \tfrac{1}{2} a_{12} = 1$$

Resto de la fila 4:

$$a_{45} = \tfrac{1}{2} a_{43} + \tfrac{1}{2} a_{42} = \tfrac{3}{8} \qquad a_{46} = \tfrac{1}{2} a_{43} + \tfrac{1}{2} a_{44} = \tfrac{5}{8}$$

Se copia la fila 4 como columna 4:

$a_{54} = a_{45} = 3/8$; $\quad a_{64} = a_{46} = 5/8$

Elemento Diagonal de la fila 5:

$a_{55} = 1 + F_5 = 1 + 1/2\, a_{23} = 1 + 1/2\, 1/2 = 5/4$

Resto de la fila 5 y elemento correspondiente de la parte simétrica:

$a_{56} = 1/2\, a_{53} + 1/2\, a_{54} = 9/16$ $\quad a_{65} = a_{56} = 9/16$

Y finalmente elemento Diagonal de la fila 6:

$a_{66} = 1 + F_6 = 1 + 1/2\, a_{34} = 1 + 1/2\, 1/4 = 7/8$

La tabla queda de la siguiente manera:

			(2)	(1-2)	(2-3)	(3-4)
	1	**2**	**3**	**4**	**5**	**6**
1	1	0	0	½	0	¼
2	0	1	½	½	¾	½
3	0	½	1	¼	¾	⅝
4	½	½	¼	1	⅜	⅝
5	0	¾	¾	⅜	5/4	9/16
6	¼	½	⅝	⅝	9/16	⅞

Ha quedado así construida la matriz **A**. Esta matriz contiene toda la información sobre parentescos y consanguinidad de los individuos de la población. Así por ejemplo, en el cruce entre el *4* y el *1* hay un valor de 0,5 ya que comparten el 50% de los genes al ser *4* hijo de *1*, y por ejemplo entre *1* y *6* hay ¼ ya que comparten la cuarta parte de los genes al ser *6* nieto de *1* por ser hijo de *4* que a su vez es hijo de *1*. Obsérvese cómo la matriz **A** proporciona todos los valores de todos los cruces cuando las relaciones se complican. Asimismo, todos los coeficientes de consanguinidad destacan en lo que excede a *1* el valor de la diagonal de cada individuo. Los dos individuos consanguíneos en el ejemplo son el *5* y el *6*. Obsérvese por ejemplo cómo, en el caso del *5*, el *2* es simultáneamente su madre y su abuela paterna.

La matriz **A** queda así definida, lo mismo que la matriz **G** que proporcionaba todas las varianzas y covarianzas entre todos los niveles del efecto genético aditivo:

$$\mathbf{A}=\begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 & \frac{1}{4} & \frac{3}{4} & \frac{5}{8} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 & \frac{3}{8} & \frac{5}{8} \\ 0 & \frac{3}{4} & \frac{3}{4} & \frac{3}{8} & \frac{5}{4} & \frac{9}{16} \\ \frac{1}{4} & \frac{1}{2} & \frac{5}{8} & \frac{5}{8} & \frac{9}{16} & \frac{9}{8} \end{bmatrix} \Rightarrow \mathbf{G}=\begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 & \frac{1}{4} & \frac{3}{4} & \frac{5}{8} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 & \frac{3}{8} & \frac{5}{8} \\ 0 & \frac{3}{4} & \frac{3}{4} & \frac{3}{8} & \frac{5}{4} & \frac{9}{16} \\ \frac{1}{4} & \frac{1}{2} & \frac{5}{8} & \frac{5}{8} & \frac{9}{16} & \frac{9}{8} \end{bmatrix} \sigma_u^2$$

VII-5.3.4. Construcción de la inversa de la matriz de relaciones aditivas $\mathbf{A}^{-1}$

La matriz de relaciones aditivas **A** es muy interesante desde el punto de vista descriptivo ya que proporciona el parentesco entre todos los individuos en forma de porcentaje de genes compartidos. Sin embargo, la matriz que es utilizada en las ecuaciones del modelo mixto es su inversa. La inversión de matrices es un proceso computacionalmente costoso por lo que debe evitarse con poblaciones de tamaño grande.

Existe una forma de construir la inversa de la matriz de relaciones aditivas mediante unas reglas sencillas sin necesidad de construir primero la matriz original e invertirla después. Se deducen a continuación estas reglas que se deben a Henderson y que pueden ser aún más sencillas si se acepta una inversa de **A** aproximada.

Para ello se precisa expresar la matriz **A** en álgebra matricial y obtener la inversa de esa expresión. Por tanto nuestro objetivo es encontrar una expresión de $\mathbf{A}=\dfrac{Var(\mathbf{u})}{\sigma_u^2}$

Podemos expresar la ecuación de la herencia aditiva de todos los individuos de nuestro ejemplo obviando en la expresión el valor

genético aditivo de los padres desconocidos. Estas ecuaciones quedarían así:

$$
\begin{aligned}
u_1 &= \varphi_1 \\
u_2 &= \varphi_2 \\
u_3 &= \tfrac{1}{2} u_2 + \varphi_3 \\
u_4 &= \tfrac{1}{2} u_1 + \tfrac{1}{2} u_2 + \varphi_4 \\
u_5 &= \tfrac{1}{2} u_2 + \tfrac{1}{2} u_3 + \varphi_5 \\
u_6 &= \tfrac{1}{2} u_3 + \tfrac{1}{2} u_4 + \varphi_6
\end{aligned}
$$

Estas ecuaciones pueden expresarse en álgebra matricial si definimos la matriz de padres **P**. Esta matriz es de tamaño *n* x *n* y todos sus elementos son ceros a excepción de las posiciones en las que el individuo que define la columna es padre del que define la fila. Así por ejemplo, las filas primera y segunda tendrán todos sus elementos nulos al no conocerse los padres de los dos primeros individuos, mientras que en la fila 3 habrá un único 1 en la columna 2 por ser éste el padre de aquél, en la cuarta habrá unos en las columnas de sus padres que son la 1 y la 2, y así sucesivamente. La matriz de padres de nuestro ejemplo es:

$$
\mathbf{P} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0
\end{bmatrix}
$$

INTERPRETACIÓN DE LAS POTENCIAS DE P

Una propiedad interesante de **P** es que la matriz **P***P, $\mathbf{P}^2$, es la matriz de los padres de los padres, es decir, la matriz de abuelos. En nuestro caso los únicos individuos de los que se conocen abuelos son el 5 y el 6, de modo que las cuatro primeras filas de $\mathbf{P}^2$ son elementos nulos. En concreto el 2 es abuelo por partida doble del 6, ya que es padre de 3 y de 4 que son los padres de 6:

$$\mathbf{P^2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Obsérvese que en todo archivo de pedigrí existe una potencia de P a partir de la cual todas las superiores son matrices nulas.

Obsérvese cómo las ecuaciones anteriores se pueden expresar en álgebra matricial de la siguiente manera:

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{bmatrix}$$

$$\mathbf{u} = ½ \quad \mathbf{P} \quad \mathbf{u} + \boldsymbol{\varphi}$$

Despejamos el vector **u**:

$$\mathbf{u} = ½\,\mathbf{P}\mathbf{u} + \boldsymbol{\varphi}$$
$$\mathbf{u} - ½\,\mathbf{P}\mathbf{u} = \boldsymbol{\varphi}$$
$$\left(\mathbf{I} - ½\,\mathbf{P}\right)\mathbf{u} = \boldsymbol{\varphi}$$
$$\mathbf{u} = \left(\mathbf{I} - ½\,\mathbf{P}\right)^{-1}\boldsymbol{\varphi}$$

DESARROLLO EN SERIE DE TAYLOR DE $\left(\mathbf{I} - ½\,\mathbf{P}\right)^{-1}$

Si e parte de la expresión original y se va sustituyendo recursivanebte u por su valor se llega a:

$$\mathbf{u} = \tfrac{1}{2}\mathbf{P}\mathbf{u} + \boldsymbol{\varphi}$$

$$\mathbf{u} = \tfrac{1}{2}\mathbf{P}\left(\tfrac{1}{2}\mathbf{P}\mathbf{u} + \boldsymbol{\varphi}\right) + \boldsymbol{\varphi} = \tfrac{1}{4}\mathbf{P}^2\mathbf{u} + \tfrac{1}{2}\mathbf{P}\boldsymbol{\varphi} + \boldsymbol{\varphi} =$$

$$= \tfrac{1}{4}\mathbf{P}^2\left(\tfrac{1}{2}\mathbf{P}\mathbf{u} + \boldsymbol{\varphi}\right) + \tfrac{1}{2}\mathbf{P}\boldsymbol{\varphi} + \boldsymbol{\varphi} = \tfrac{1}{8}\mathbf{P}^3\boldsymbol{\varphi} + \tfrac{1}{4}\mathbf{P}^2\boldsymbol{\varphi} + \tfrac{1}{2}\mathbf{P}\boldsymbol{\varphi} + \boldsymbol{\varphi} =$$

.....

$$= (\mathbf{I} + \tfrac{1}{2}\mathbf{P} + \tfrac{1}{4}\mathbf{P}^2 + \tfrac{1}{8}\mathbf{P}^3 + ...)\boldsymbol{\varphi}$$

Como fue comentado anteriormente esta serie no es infinita ya que existe siempre una potencia de **P** a partir de la cual todas valen **0**. Por otro lado, anteriormente se ha mostrado que $\mathbf{u} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1}\boldsymbol{\varphi}$, es decir, que se pueden igualar ambos factores de φ en las dos expresiones:

$$\left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1} = (\mathbf{I} + \tfrac{1}{2}\mathbf{P} + \tfrac{1}{4}\mathbf{P}^2 + \tfrac{1}{8}\mathbf{P}^3 + ...)$$

Este resultado muestra que no es preciso invertir esta matriz sino que puede ser construida directamente a partir de la matriz **P**.

La matriz $\left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1}$, normalmente denotada por **T**, es una matriz triangular inferior (todos los elementos por encima de la diagonal principal son nulos) que representa el porcentaje de genes transmitidos por descendencia de cada individuo de esa columna con el individuo de esa fila. En el caso en el que el individuo de la columna sea fundador representa además el porcentaje de genes que comparten ya que no existe ninguna otra vía en el pedigrí para ello. En el ejemplo, marcando la parte de la matriz que corresponde a columnas de fundadores, es de la siguiente manera:

$$\mathbf{T} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1} = \left[\begin{array}{cc|cccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \tfrac{1}{2} & 1 & 0 & 0 & 0 \\ \tfrac{1}{2} & \tfrac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & \tfrac{3}{4} & \tfrac{1}{2} & 0 & 1 & 0 \\ \tfrac{1}{4} & \tfrac{1}{2} & \tfrac{1}{2} & \tfrac{1}{2} & 0 & 1 \end{array}\right]$$

Para llegar hasta la matriz **A** necesito la varianza del vector **u**. La obtenemos teniendo en cuenta que **u** es el producto de una parte constante y un vector φ variable:

$$Var(\mathbf{u}) = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1} Var(\boldsymbol{\varphi})\left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}'\right)^{-1}$$

Nótese que para obtener la transpuesta de $\left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1}$ ha bastado transponer la matriz **P** ya que **I** es la matriz identidad que es una

matriz simétrica. Para llegar a **A** sólo falta dividir por σ_u^2, lo que hacemos a ambos lados de la igualdad:

$$\mathbf{A} = \frac{Var(\mathbf{u})}{\sigma_u^2} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1} \frac{Var(\boldsymbol{\varphi})}{\sigma_u^2} \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}'\right)^{-1}$$

Como fue comentado anteriormente las covarianzas entre dos muestreos mendelianos distintos son nulas ya que no parece haber ninguna relación entre lo que distingue a un hermano de otro por azar, y con más razón si son dos individuos sin parentesco. Así pues la matriz $Var(\boldsymbol{\varphi})$ es una matriz diagonal y su cociente por σ_u^2 sigue siendo una matriz diagonal que llamaremos **D**:

$$\mathbf{D} = \frac{Var(\boldsymbol{\varphi})}{\sigma_u^2} = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & d_{nn} \end{bmatrix} = \begin{bmatrix} \dfrac{\sigma_{\varphi_1}^2}{\sigma_u^2} & 0 & \dots & 0 \\ 0 & \dfrac{\sigma_{\varphi_2}^2}{\sigma_u^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dfrac{\sigma_{\varphi_n}^2}{\sigma_u^2} \end{bmatrix}$$

Así la expresión para obtener **A** queda:

$$\mathbf{A} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1} \mathbf{D} \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}'\right)^{-1}$$

Llamando a $\mathbf{T} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1}$, **A** se puede escribir como **A** = **TDT'**. Finalmente, al calcular la inversa de **A** se invierte el orden de de los factores y cada uno de ellos:

$$\mathbf{A}^{-1} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}'\right) \mathbf{D}^{-1} \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)$$

Así pues se puede obtener la inversa de **A** mediante el producto de otras tres, de las cuales sólo una de ellas es inversa, pero al tratarse

de la inversa de una matriz diagonal es sencilla de obtener mediante la inversión de cada uno de los valores de la diagonal.

VII-5.3.5. Obtención de los elementos de la matriz $\mathbf{D}^{-1}$

En la expresión anterior los elementos de las matrices **I** y **P** son conocidos y son todos ceros y unos. .Sólo nos falta conocer los elementos de la matriz **D** y su inversa. Al tratarse de una matriz diagonal sólo los elementos de la diagonal son distintos de cero y se corresponden con $d_{ii} = \frac{\sigma^2_{\varphi_i}}{\sigma^2_u}$. Llegaremos hasta el valor de cada d_{ii} partiendo de la descomposición del valor genético aditivo del individuo i en función de los valores genéticos aditivos de sus padres y el muestreo mendeliano:

$$u_i = \tfrac{1}{2} u_j + \tfrac{1}{2} u_k + \varphi_i$$

Para llegar hasta d_{ii} se necesita la varianza de φ_i, así que se calcula la varianza a la derecha y a la izquierda de esta igualdad, teniendo en cuenta que la única covarianza no nula es la que existe entre u_j y u_k que entra con valor doble por lo que $2Cov(\tfrac{1}{2}u_j, \tfrac{1}{2}u_k)) = 2\,\tfrac{1}{2}\,\tfrac{1}{2}\sigma_{u_j u_k} = \tfrac{1}{2}\sigma_{u_j u_k}$:

$$\sigma^2_{u_i} = \tfrac{1}{4}\sigma^2_{u_j} + \tfrac{1}{4}\sigma^2_{u_k} + \tfrac{1}{2}\sigma_{u_j u_k} + \sigma^2_{\varphi_i}$$

Y finalmente sólo falta dividir por σ^2_u:

$$\frac{\sigma^2_{u_i}}{\sigma^2_u} = \tfrac{1}{4}\frac{\sigma^2_{u_j}}{\sigma^2_u} + \tfrac{1}{4}\frac{\sigma^2_{u_k}}{\sigma^2_u} + \tfrac{1}{2}\frac{\sigma_{u_j u_k}}{\sigma^2_u} + \frac{\sigma^2_{\varphi_i}}{\sigma^2_u}$$

Las varianzas y covarianzas entre valores genéticos aditivos divididos por σ^2_u se corresponden con elementos de **A** de la diagonal o de fuera de la diagonal, de manera que la expresión resulta:

$$a_{ii} = ¼\, a_{jj} + ¼\, a_{kk} + ½\, a_{jk} + d_{ii}$$

Recordemos ahora que $a_{ii} = 1 + F_i = 1 + ½\, a_{jk}$. En el lado izquierdo está a_{ii} que es igual a $1 + F_i$ y en el derecho está $½\, a_{jk}$ que es F_i. Se despeja d_{ii} y se obtiene:

$$d_{ii} = 1 - ¼\, a_{jj} - ¼\, a_{kk}$$

Dado que d_{ii} depende únicamente de elementos diagonales de **A**, los cuales son igual a 1 más la consanguinidad de los individuos, el problema de construir la inversa de la matriz de relaciones aditivas se ha reducido finalmente a calcular la consanguinidad de los individuos que actúan como padres.

VII-5.3.6. Regla de construcción de $\mathbf{A}^{-1}$

Se parte ahora de la expresión anterior y se aprovecha que la matriz $\mathbf{D}^{-1}$ es diagonal para posponer el producto de los elementos de esta matriz y multiplicar previamente los elementos de los paréntesis:

$$\mathbf{A}^{-1} = \left(\mathbf{I} - ½\, \mathbf{P}'\right) \mathbf{D}^{-1} \left(\mathbf{I} - ½\, \mathbf{P}\right) = \left\{\mathbf{I} - ½\, \mathbf{P} - ½\, \mathbf{P}' + ¼\, \mathbf{P}'\mathbf{P}\right\} d_{ii}^{-1}$$

Los coeficientes que se derivan de las expresiones en el interior de la llave son las siguientes:

- **I**: La matriz identidad está compuesta por unos en todas las posiciones de la diagonal, por lo que se anotará un 1 en la diagonal del individuo.
- $-½\, \mathbf{P}$. La matriz **P** posee unos en el cruce entre cada individuo y sus padres. Por tanto, se anotará –½ en el cruce entre el individuo y sus padres.
- $-½\, \mathbf{P}'$. Esta matriz es la transpuesta de la anterior por lo que se anotará –½ en el cruce entre los padres y su hijo.

- $\frac{1}{4}\mathbf{P}'\mathbf{P}$. Puede comprobarse que el producto **P'P** da lugar a unos en las diagonales de los padres y al cruce del padre con la madre y de la madre con el padre. Por tanto, se anotará ¼ en las diagonales de los padres y en el cruce entre los padres.

Estos coeficientes para cada individuo deben ser luego multiplicados por el inverso del elemento correspondiente de **D**. Todo ello puede resumirse en la siguiente tabla:

	i	**j**	**k**
i	1	–½	–½
j	–½	¼	¼
k	–½	¼	¼

x d_{ii}^{-1}; siendo $d_{ii} = 1 - \frac{1}{4}a_{jj} - \frac{1}{4}a_{kk}$

VII-5.3.7. Reglas de Henderson para la construcción de A^{-1} aproximada

El cálculo de la consanguinidad para poblaciones muy grandes puede suponer un elevado coste computacional, tanto en tiempo como en cantidad de memoria utilizada. En general la consanguinidad de las poblaciones sometidas a selección artificial no es elevada, por lo que podemos ignorar la consanguinidad de los padres en las reglas de construcción de la inversa de **A** proporcionando reglas sencillas para construirla.

Así, si en $d_{ii} = 1 - \frac{1}{4}a_{jj} - \frac{1}{4}a_{kk}$ se asume que la consanguinidad de *j* y *k* es nula en todos los casos, entonces los valores de d_{ii} y de su inverso d_{ii}^{-1} dependerán únicamente de si *j* y *k* son o no conocidos:

			d_{ii} =	d_{ii}^{-1} =
d_{ii} =	*j* y *k* desconocidos	$a_{jj} = a_{kk} = 0$	1	1
	j conocido y *k* desconocido o *j* desconocido y *k* conocido	$a_{jj} = 1; a_{kk} = 0$ *o* $a_{jj} = 0; a_{kk} = 1$	¾	4/3
	j y *k* conocidos	$a_{jj} = a_{kk} = 1$	½	2

Utilizando estos valores de d_{ii}^{-1} en la tabla anterior, y teniendo en cuenta que los coeficientes que afectan a individuos desconocidos no pueden anotarse, se deducen las reglas de Henderson para la construcción de la inversa de la matriz de relaciones aditivas.

- Si se desconocen ambos padres del individuo *i*:
  - Se anota un 1 en la diagonal del individuo, en la posición (*i*, *i*).
- Si sólo se conoce uno de los padres, por ejemplo *j*:
  - Se anota 4/3 en la diagonal del individuo (*i*, *i*).
  - −2/3 donde se cruza el individuo *i* con el padre conocido *j*, en los dos lugares donde se da el cruce: (*i*, *j*) y (*j*, *i*).
  - 1/3 en la diagonal del padre conocido: (*j*, *j*).
- Si se conocen ambos padres del individuo *i* se añade:
  - 2 en el elemento (*i*, *i*).
  - -1 en los cruces entre el individuo *i* y sus padres *j* y *k* en todos los lugares donde se dan: (*i*, *j*), (*j*, *i*), (*i*, *k*) y (*k*, *i*).
  - ½ en la diagonal de los padres y en los dos lugares donde se cruzan entre sí: (*j*, *j*), (*k*, *k*), (*j*, *k*) y (*k*, *j*).

En primer lugar se crea una tabla de *n* filas y *n* columnas teniendo en cuenta que en algunas de las celdas se anotarán varios coeficientes que luego habrán de sumarse. A continuación se recorre el registro de pedigrí desde el primer individuo hasta el último y se anotan los siguientes coeficientes dependiendo del número de padres conocidos de cada individuo:

Una vez anotados todos los coeficientes se suman los valores anotados dentro de cada casilla y se obtiene la inversa. Nótese que estas reglas proporcionan exactamente $\mathbf{A}^{-1}$ si ningún individuo consanguíneo tiene hijos.

Siguiendo la genealogía de nuestro ejemplo construiremos la matriz $\mathbf{A}^{-1}$ anotando los elementos exactamente en este orden:

Animal	Padre	Madre
1	-	-
2	-	-
3	-	2
4	1	2
5	3	2
6	3	4

1. Primer individuo. No tiene padres conocidos; se anota un 1 en la posición (1, 1)
2. Segundo individuo. Tampoco tiene padres conocidos, se anota un 1 en la posición (2, 2)
3. Tercer individuo. Sólo un padre conocido (2). Se anota:
    a. $\frac{4}{3}$ en la posición (3, 3)
    b. $-\frac{2}{3}$ en las posiciones (2, 3) y (3, 2)
    c. $\frac{1}{3}$ en la posición (2, 2)
4. Cuarto individuo. Ambos padres conocidos (1 y 2). Se anota:
    a. 2 en la posición (4, 4)
    b. -1 en las posiciones (1, 4) , (2, 4), (4, 1) y (4, 2)
    c. ½ en las posiciones (1, 1), (2, 2), (1, 2) y (2, 1)
5. Quinto individuo. Ambos padres conocidos (2 y 3). Se anota:
    a. 2 en la posición (5, 5)

   b. -1 en las posiciones (2, 5) , (3, 5), (5, 2) y (5, 3)
   c. ½ en las posiciones (2, 2), (3, 3), (2, 3) y (3, 2)
6. Sexto individuo. Ambos padres conocidos (3 y 4). Se anota:
   a. 2 en la posición (6, 6)
   b. -1 en las posiciones (3, 6) , (4, 6), (6, 3) y (6, 4)
   c. ½ en las posiciones (3, 3), (4, 4), (3, 4) y (4, 3)

Una vez anotados todos los coeficientes la tabla queda así:

	1	**2**	(2) **3**	(1-2) **4**	(2-3) **5**	(3-4) **6**
1	$1+\frac{1}{2}$	$\frac{1}{2}$		-1		
2	$\frac{1}{2}$	$1+\frac{1}{3}+\frac{1}{2}+\frac{1}{2}$	$-\frac{2}{3}+\frac{1}{2}$	-1	-1	
3		$-\frac{2}{3}+\frac{1}{2}$	$\frac{4}{3}+\frac{1}{2}+\frac{1}{2}$	$\frac{1}{2}$	-1	-1
4	-1	-1	$\frac{1}{2}$	$2+\frac{1}{2}$		-1
5		-1	-1		2	
6			-1	-1		2

Y después de sumar los elementos dentro de cada casilla se obtiene la inversa de **A** sin necesidad de calcularla:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 & \frac{1}{4} & \frac{3}{4} & \frac{5}{8} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 & \frac{3}{8} & \frac{5}{8} \\ 0 & \frac{3}{4} & \frac{3}{4} & \frac{3}{8} & \frac{5}{4} & \frac{9}{16} \\ \frac{1}{4} & \frac{1}{2} & \frac{5}{8} & \frac{5}{8} & \frac{9}{16} & \frac{6}{8} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} & 0 & -1 & 0 & 0 \\ \frac{1}{2} & \frac{7}{3} & -\frac{1}{6} & -1 & -1 & 0 \\ 0 & -\frac{1}{6} & \frac{7}{3} & \frac{1}{2} & -1 & -1 \\ -1 & -1 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix}$$

Obsérvese que los individuos de una población con elevada profundidad de pedigrí estarán muy relacionados y la matriz **A** será una matriz densa, es decir, con pocos elementos nulos. Sin embargo, $\mathbf{A}^{-1}$, su inversa, sólo presenta elementos no nulos en las diagonales, entre los individuos que se cruzan y entre padres e hijos, por lo que tendrá un elevado número de ceros. Por otro

lado, aunque la matriz **A** es muy interesante desde el punto de vista descriptivo del porcentaje de genes compartidos entre los individuos de la población, resulta mucho más complicada de construir y tiene menos elementos nulos que su inversa, cuando es ésta última la que se necesita en la resolución del modelo de valoración genética. Es por ello que siempre se obtiene $\mathbf{A}^{-1}$ directamente cuando se trata de hacer valoraciones genéticas.

VII-5.4. La ponderación de la importancia de la información de parentesco

Al contemplar las ecuaciones del modelo mixto de Henderson se observa que toda la información de parentesco se encuentra concentrada en $\mathbf{A}^{-1}$. También se observa que esta matriz se incorpora en las ecuaciones multiplicada por un coeficiente α que fue definido como $\alpha = \frac{\sigma_e^2}{\sigma_u^2}$. Por tanto, se le dará mayor ponderación (mayor importancia) a la información de parentesco cuando la variabilidad de origen genético sea más pequeña en relación a la variabilidad residual. Esta interpretación es más clara cuando se expresa α en función de la heredabilidad $\left(h^2 = \frac{\sigma_u^2}{\sigma_p^2} \right)$ en un modelo en el que el único efecto aleatorio además del residuo es el efecto genético aditivo. En este modelo la varianza fenotípica es la suma de las otras dos componentes:

$$\sigma_p^2 = \sigma_u^2 + \sigma_e^2 \Rightarrow \sigma_e^2 = \sigma_p^2 - \sigma_u^2$$

De manera que:

$$\alpha = \frac{\sigma_p^2 - \sigma_u^2}{\sigma_u^2} = \frac{\left(\sigma_p^2 - \sigma_u^2\right) \Big/ \sigma_p^2}{\sigma_u^2 \Big/ \sigma_p^2} = \frac{\sigma_p^2 \Big/ \sigma_p^2 - \sigma_u^2 \Big/ \sigma_p^2}{\sigma_u^2 \Big/ \sigma_p^2} = \frac{1 - h^2}{h^2}$$

Así dando valores a la heredabilidad dentro de su espacio paramétrico se observa que a medida que la heredabilidad se acerca a cero, α tiende a infinito mientras que para un valor de heredabilidad igual a 1, α vale cero. En otras palabras, la información de los datos de los parientes será tanto más importante cuanto menor sea la heredabilidad, mientras que para valores altos de la heredabilidad el dato del propio individuo es el verdaderamente importante.

En nuestro ejemplo utilizaremos un valor de heredabilidad de 0,6, de forma que $\alpha = \frac{2}{3}$. Con este valor obtenemos $\mathbf{A}^{-1}\alpha$, para incorporar este producto en la matriz de coeficientes:

$$\mathbf{A}^{-1}\alpha = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} & 0 & -1 & 0 & 0 \\ \frac{1}{2} & \frac{7}{3} & -\frac{1}{6} & -1 & -1 & 0 \\ 0 & -\frac{1}{6} & \frac{7}{3} & \frac{1}{2} & -1 & -1 \\ -1 & -1 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix} \frac{2}{3} =$$

$$= \begin{bmatrix} 1 & \frac{1}{3} & 0 & -\frac{2}{3} & 0 & 0 \\ \frac{1}{3} & \frac{14}{9} & -\frac{1}{9} & -\frac{2}{3} & -\frac{2}{3} & 0 \\ 0 & -\frac{1}{9} & \frac{14}{9} & \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & -\frac{2}{3} & \frac{1}{3} & \frac{5}{3} & 0 & -\frac{2}{3} \\ 0 & -\frac{2}{3} & -\frac{2}{3} & 0 & \frac{4}{3} & 0 \\ 0 & 0 & -\frac{2}{3} & -\frac{2}{3} & 0 & \frac{4}{3} \end{bmatrix}$$

Otra cuestión interesante de comentar es que, en este modelo, α depende de la heredabilidad y, por lo tanto, no es preciso conocer las distintas varianzas de los efectos sino que basta con conocer la heredabilidad del carácter.

VII-5.5. Las ecuaciones del modelo mixto

Se van a mostrar las ecuaciones en nuestro ejemplo. La información disponible era la siguiente:

Animal	Padre	Madre	Ganadería	Peso
1	-	-	-	-
2	-	-	1	415
3	-	2	1	430
4	1	2	2	420
5	3	2	2	400
6	3	4	2	405

Es preciso recordar que el ejemplo se desarrolla sobre las ecuaciones simplificadas del BLUP, las cuales se expresan de la siguiente manera:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Para completar las ecuaciones sólo falta hacer el producto de las matrices **X'X**, **X'Z**, **Z'X**, **Z'Z**, **X'y** y **Z'y**. Sin embargo no es necesario hacer los distintos productos ya que se pueden obtener de forma sencilla.

Los bloques **X'X**, **X'Z**, **Z'X** y **Z'Z** se pueden obtener simplemente rellenando una tabla con tantas filas y columnas como incógnitas. El valor a incluir en cada casilla será el resultado de contar el número de observaciones que se ven afectadas por cada combinación de incógnitas. Por ejemplo, en el cruce entre la incógnita "ganadería 1" (denotada en el modelo como b_1) y la incógnita "valor genético aditivo del individuo 2" (en el modelo u_2), habrá que poner el número de observaciones que el individuo 2 ha tenido en la ganadería 1. En este caso se trata de una observación, así que habrá que anotar un 1. Los vectores del lado derecho de las ecuaciones **X'y** y **Z'y** se corresponden con la suma de las observaciones que se ven afectadas por cada una de las incógnitas. Así, se puede completar el vector del lado derecho

simplemente completando una tabla de una sola columna en la que se anotan estas sumas. Por ejemplo, el valor correspondiente a la ganadería 1 será la suma de los datos que se han recogido en esa ganadería, es decir, 415 + 430 = 845. Las tablas quedan de la siguiente manera:

	μ	b_1	b_2	u_1	u_2	u_3	u_4	u_5	u_6
μ	5	2	3	0	1	1	1	1	1
b_1	2	2	0	0	1	1	0	0	0
b_2	3	0	3	0	0	0	1	1	1
u_1	0	0	0	0	0	0	0	0	0
u_2	1	1	0	0	1	0	0	0	0
u_3	1	1	0	0	0	1	0	0	0
u_4	1	0	1	0	0	0	1	0	0
u_5	1	0	1	0	0	0	0	1	0
u_6	1	0	1	0	0	0	0	0	1

μ	2070
b_1	845
b_2	1225
u_1	0
u_2	415
u_3	430
u_4	420
u_5	400
u_6	405

Los cuatro bloques que se han marcado en la tabla de la izquierda se corresponden con los definidos en las ecuaciones del modelo mixto como **X'X**, **X'Z**, **Z'X** y **Z'Z**. Los dos vectores separados en el lado izquierdo se corresponden asimismo con **X'y** y **Z'y**.

Agrupando toda la información detallada hasta este punto se pueden completar las ecuaciones del modelo mixto:

$$\begin{bmatrix} 5 & 2 & 3 & 0 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 3 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0{,}33 & 0 & -0{,}67 & 0 & 0 \\ 1 & 1 & 0 & 0{,}33 & 2{,}56 & -0{,}11 & -0{,}67 & -0{,}67 & 0 \\ 1 & 1 & 0 & 0 & -0{,}11 & 2{,}56 & 0{,}33 & -0{,}67 & -0{,}67 \\ 1 & 0 & 1 & -0{,}67 & -0{,}67 & 0{,}33 & 2{,}67 & 0 & -0{,}67 \\ 1 & 0 & 1 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33 & 0 \\ 1 & 0 & 1 & 0 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \end{bmatrix} = \begin{bmatrix} 2070 \\ 845 \\ 1225 \\ 0 \\ 415 \\ 430 \\ 420 \\ 400 \\ 405 \end{bmatrix}$$

VII-5.6. Modelos de rango no completo e inversas generalizadas.

En principio, para resolver las ecuaciones del modelo mixto, ha de invertirse la matriz de coeficientes y multiplicar por el vector del lado derecho para dar lugar a las soluciones. Sin embargo, la inversión de la matriz de coeficientes no es posible ya que existe una combinación lineal entre las filas o columnas de esta matriz. Como consecuencia de esta dependencia lineal el determinante de la matriz es nulo y se dice que el modelo es de rango no completo. Esta combinación se debe al ajuste de efectos fijos discontinuos además de la media general, en este caso la ganadería, lo que provoca que la suma de las filas correspondientes a este efecto dé lugar a la correspondiente a la media y a que el modelo posee infinitas soluciones. Así pues, es necesario obtener las soluciones de los efectos fijos para un valor arbitrario de una de las incógnitas involucradas en la combinación lineal, es decir, se hace necesario resolver con la ayuda de una inversa generalizada. El tema fue tratado con la profundidad necesaria en el modelo fijo y todo aquello no va a ser repetido aquí con igual detalle.

Resolveremos el modelo por ejemplo asignando cero al valor de la media dado que facilita el razonamiento de las soluciones pero debe quedar claro que cualquier solución que asigne un valor arbitrario a cualquiera de las dos ganaderías o a la media habría conducido a las mismas soluciones de las funciones estimables.

Las nuevas ecuaciones quedan de la siguiente manera:

$$\begin{bmatrix} 2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0{,}33 & 0 & -0{,}67 & 0 & 0 \\ 1 & 0 & 0{,}33 & 2{,}56 & -0{,}11 & -0{,}67 & -0{,}67 & 0 \\ 1 & 0 & 0 & -0{,}11 & 2{,}56 & 0{,}33 & -0{,}67 & -0{,}67 \\ 0 & 1 & -0{,}67 & -0{,}67 & 0{,}33 & 2{,}67 & 0 & -0{,}67 \\ 0 & 1 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33 & 0 \\ 0 & 1 & 0 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \end{bmatrix} = \begin{bmatrix} 845 \\ 1225 \\ 0 \\ 415 \\ 430 \\ 420 \\ 400 \\ 405 \end{bmatrix}$$

VII-5.6.1. Interpretación de las ecuaciones del modelo mixto

Cada una de las filas de la matriz de coeficientes multiplicada por el vector de incógnitas e igualado al término del lado derecho, se interpreta como una ecuación para la incógnita de esa posición. De este modo para cada una de las incógnitas se ajustan todas las otras que le afectan. Al resolver todas las ecuaciones al mismo tiempo todas quedan ajustadas a la vez. A diferencia del BLP las soluciones de los efectos fijos tienen en cuenta los niveles de los efectos aleatorios que le afectan y lo mismo ocurre con las soluciones de los aleatorios con respecto a los fijos. Esta forma de resolver el modelo es lo que proporciona al BLUP la propiedad de insesgado.

Así por ejemplo la primera ecuación, la solución para la primera ganadería se obtendría:

$$2\hat{b}_1 + 0\hat{b}_2 + 0\hat{u}_1 + 1\hat{u}_2 + 1\hat{u}_3 + 0\hat{u}_4 + 0\hat{u}_5 + 0\hat{u}_6 = 845$$

$$2\hat{b}_1 + \hat{u}_2 + \hat{u}_3 = 845$$

$$\hat{b}_1 = \frac{845 - \hat{u}_2 - \hat{u}_3}{2}$$

En esta solución 845 es la suma de los datos de la ganadería 1 y 2 es el número de datos de la ganadería 1. La valoración genética mediante BLP utiliza en primer lugar un modelo fijo. Con el modelo fijo la ganadería 1 habría sido estimada sencillamente a partir de su media, es decir, de la siguiente manera:

$$\hat{b}_1 = \frac{845}{2}$$

Si comparamos ambas soluciones observamos que en la primera antes de dividir la suma de los datos por el número de datos, se resta el valor genético de los animales que han dado esos datos. Las soluciones para el efecto ganadería deberían proporcionar un valor que describiese la superioridad o inferioridad de los rendimientos de la ganadería por cualquier causa que no fuera el

valor genético de los animales que dan lugar a esos rendimientos, es decir, deberían permitir que el modelo separase un efecto del otro. La solución del modelo fijo ignora el valor genético de los animales que han dado lugar esos rendimientos al proporcionar el valor de la ganadería simplemente a partir de la media de los registros, mientras que las soluciones BLUE de los efectos fijos al resolver el modelo mixto descuentan de los rendimientos el valor genético de los animales que los produjeron. Así, dado que la media de los valores genéticos se define como cero, si los individuos son de tipo medio no aportarán ni restarán nada al valor medio del rebaño y entonces ambas soluciones coinciden. Sin embargo, si los animales son superiores a la media, antes de dividir por el número de datos se restará de los rendimientos el exceso que han producido con respecto al valor proporcionado por el manejo de la ganadería. Por el contrario, si los animales fueran genéticamente inferiores a la media su valor sería negativo, por lo que al restar sus valores genéticos de sus rendimientos, provocarán un incremento en los rendimientos medidos de los mismos, para corregir el menor valor de sus registros por su menor valor genético. Este ajuste se da simultáneamente para todas las incógnitas del modelo por lo que todas las soluciones son insesgadas, tanto los BLUE de los efectos fijos como los BLUP de los efectos aleatorios. Esta propiedad sólo es posible obtenerla mediante la resolución conjunta de todas las incógnitas del modelo.

VII-5.7. Resolución de las ecuaciones del modelo mixto

Los métodos de resolución de las ecuaciones del modelo mixto pueden ser divididos en métodos directos y métodos indirectos. Los primeros son aquellos que obtienen las soluciones de forma precisa. El mejor ejemplo de método directo es la inversión de la matriz de coeficientes y posterior multiplicación por los términos independientes:

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z}\\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\alpha\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X'y}\\ \mathbf{Z'y}\end{bmatrix}$$

Para nuestro ejemplo las soluciones obtenidas son:

$$\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{u}_1\\ \hat{u}_2\\ \hat{u}_3\\ \hat{u}_4\\ \hat{u}_5\\ \hat{u}_6\end{bmatrix}=\begin{bmatrix}2 & 0 & 0 & 1 & 1 & 0 & 0 & 0\\ 0 & 3 & 0 & 0 & 0 & 1 & 1 & 1\\ 0 & 0 & 1 & 0{,}33 & 0 & -0{,}67 & 0 & 0\\ 1 & 0 & 0{,}33 & 2{,}56 & -0{,}11 & -0{,}67 & -0{,}67 & 0\\ 1 & 0 & 0 & -0{,}11 & 2{,}56 & 0{,}33 & -0{,}67 & -0{,}67\\ 0 & 1 & -0{,}67 & -0{,}67 & 0{,}33 & 2{,}67 & 0 & -0{,}67\\ 0 & 1 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33 & 0\\ 0 & 1 & 0 & 0 & -0{,}67 & -0{,}67 & 0 & 2{,}33\end{bmatrix}^{-1}\begin{bmatrix}845\\ 1225\\ 0\\ 415\\ 430\\ 420\\ 400\\ 405\end{bmatrix}=\begin{bmatrix}425{,}5308\\ 409{,}5310\\ 4{,}0411\\ -5{,}3091\\ -0{,}7525\\ 3{,}4071\\ -5{,}8166\\ -1{,}1834\end{bmatrix}$$

OTROS MÉTODOS DE RESOLUCIÓN DE LAS ECUACIONES DEL MODELO MIXTO

Normalmente el tamaño de las ecuaciones no es tan reducido como en nuestro ejemplo, por lo que la inversión de la matriz de coeficientes no es siempre posible. Cuando el tamaño de las ecuaciones lo permite pueden utilizarse otros métodos directos basados en la manipulación de la matriz de coeficientes. En Mejora Genética, la matriz de coeficientes suele ser simétrica y no definida negativa, lo que permite su descomposición en el producto de una matriz triangular por su transpuesta. Esto puede ser empleado después para obtener las soluciones por dos sistemas consecutivos de forma que en cada ecuación sólo haya una variable a determinar. Este sistema ha sido referenciado como "back-forward solving".

No obstante, normalmente la gran cantidad de ecuaciones que aparecen en un modelo animal no permite que pueda ser resuelto por métodos directos, por lo que debemos recurrir a métodos indirectos.

Los métodos indirectos se basan en procedimientos iterativos. Se parte de valores aproximados de las incógnitas. Normalmente se utilizan los obtenidos mediante el cociente de los términos independientes entre los correspondientes valores situados en la diagonal de la matriz de coeficientes. Los dos métodos indirectos más utilizados son los conocidos con el nombre de Gauss-Seidel y Jacobi de segundo orden. El primero de ellos utiliza, dentro de cada iteración, los valores obtenidos para las incógnitas en ecuaciones precedentes, mientras que el segundo sólo utiliza las nuevas soluciones a partir de la primera ecuación de la siguiente iteración, y, por tanto, precisa retener en memoria dos valores para cada una de las incógnitas. Sin embargo, este procedimiento converge mejor cuando las relaciones entre las

distintas incógnitas son numerosas (por ejemplo las relaciones de parentesco).

Hoy en día existen a disposición de los usuarios programas preparados para evaluar animales a partir de una lista de datos y otra de genealogías, con lo que la evaluación de los reproductores se realiza sin dificultades en la actualidad.

VII-5.8. Interpretación de las soluciones de las ecuaciones del modelo mixto

VII-5.8.1. Soluciones de los efectos fijos

Las soluciones de los efectos fijos se corresponden con los dos primeros elementos del vector de soluciones más la solución de la media que ha sido forzada a valer cero:

$$\begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 0,0000 \\ 425,5308 \\ 409,5310 \end{bmatrix}$$

Al haber sido necesaria la obtención de una inversa generalizada, las soluciones obtenidas para los efectos fijos no son únicas sino que son tan sólo una de las infinitas posibles soluciones. Sin embargo, existen ciertas combinaciones lineales de las soluciones de los efectos fijos que son únicas. Se trata de las funciones estimables, que son además las soluciones que realmente son útiles:

- Diferencias entre niveles de efectos fijos. Es decir, las diferencias entre ganaderías, $\hat{b}_1 - \hat{b}_2$, siempre valdrá 15,9998. Esta diferencia entre las ganaderías se asigna exclusivamente al manejo de la ganadería ya que todos los otros efectos del modelo han sido ajustados. Por tanto, por ejemplo, siempre sabremos que un individuo de la ganadería 1 que tuviese un rendimiento 10 unidades superior a otro de la ganadería 2, es genéticamente peor que él ya que a igual valor genético debería rendir 16 unidades más. Obsérvese que estas combinaciones lineales serían funciones estimables también para otros efectos fijos discontinuos del modelo. Nuestro

ejemplo sólo incluye el efecto ganadería con dos niveles, pero en caso de haberse ajustado los correspondientes efectos fijos, podría conocerse también cuánto se produce más en primavera que en verano, cómo ha mejorado el manejo de un año a otro, o cuánto pesan más los machos que las hembras al destete.

- La media más un nivel de cada efecto fijo. Estas combinaciones lineales son precisamente las definidas en el modelo por **Xb**, en nuestro caso $\hat{\mu} + \hat{b}_1$ y $\hat{\mu} + \hat{b}_2$. Podremos conocer entonces en nuestro ejemplo la media de los rendimientos de cada ganadería, y lo que es más importante, podremos predecir el rendimiento de cada animal *i* en cada ganadería *j* mediante $\hat{\mu} + \hat{b}_j + \hat{u}_i$. Es este tipo de combinación lineal de efectos fijos y aleatorios la que se busca en la teoría de la predicción cuando se habla de **K'b** + **M'u**, y es por esta razón que la teoría de la predicción involucra también a los efectos fijos.

Es importante destacar que la falta de estimabilidad de cada una de las incógnitas fijas no afecta a los efectos aleatorios cuyas soluciones serán únicas independientemente de la inversa generalizada que se utilice para resolver el modelo.

VII-5.8.2. Soluciones de los efectos aleatorios

Las soluciones de los efectos aleatorios corresponden con los 6 últimos elementos del vector de soluciones:

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \end{bmatrix} = \begin{bmatrix} 4,0411 \\ -5,3091 \\ -0,7525 \\ 3,4071 \\ -5,8166 \\ -1,1834 \end{bmatrix}$$

El valor genético obtenido para cada individuo representa, en unidades del carácter, la superioridad que presenta con respecto a la media de los individuos de la población base. Así, si se trata de kilogramos, el individuo 1, en las mismas condiciones ambientales que los individuos de la población base, pesaría 4,04 kilogramos más que la media de los fundadores.

Una cuestión a destacar en la metodología BLUP es que proporciona valoraciones genéticas para todos los individuos del pedigrí aunque no tengan dato, siempre que exista al menos un pariente por lejano que sea que sí posea dato. Así, en nuestro caso hemos obtenido un valor genético para el individuo 1, a pesar de no disponer de dato propio. Para su evaluación han sido utilizados todos los datos de parientes del individuo, todo ello a través de la inversa de la matriz de relaciones aditivas, sin que sea preciso prestar atención alguna a esto.

La media de los valores genéticos de los animales fundadores deberá ser cero, tal y como se asume por hipótesis. Es por ello que en las soluciones aparecen valores positivos y negativos significando esto únicamente que el valor genético del individuo se encuentra por encima o por debajo de la media de los fundadores, valor éste que sirve de referencia. Sin embargo en nuestro ejemplo es difícil de comprobar que la media de los fundadores, a causa de la existencia de un individuo del que sólo se conoce un padre, por lo que existe un fundador que no ha sido identificado en nuestro ejemplo, el padre desconocido del individuo 3. Estos animales se llaman fundadores fantasmas.

Por otro lado, de no existir selección, la media de los valores genéticos no evolucionaría, y así, si tuviéramos suficientes datos que corrigieran el efecto de muestreo, la media de todos los valores genéticos sería igualmente cero. La suma de los elementos de un vector se obtiene premultiplicándolo por un vector de unos, y ha de ser cero para que la media también lo sea, lo que se expresa: $\mathbf{1'u} = 0$. Sin embargo, esto no es así bajo selección o con pocos datos como es nuestro caso y debe tenerse en cuenta el parentesco entre todos ellos para trazar retrospectivamente el pedigrí. Para ello se utiliza la

inversa de la matriz de relaciones aditivas entre individuos. Así, lo que sí se cumple en todos los casos es la siguiente expresión:

$$\mathbf{1'A^{-1}u} = 0$$

$$\mathbf{1'A^{-1}\hat{u}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3/2 & 1/2 & 0 & -1 & 0 & 0 \\ 1/2 & 7/3 & -1/6 & -1 & -1 & 0 \\ 0 & -1/6 & 7/3 & 1/2 & -1 & -1 \\ -1 & -1 & 1/2 & 5/2 & 0 & -1 \\ 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 4,0411 \\ -5,3091 \\ -0,7525 \\ 3,4071 \\ -5,8166 \\ -1,1834 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 2/3 & 2/3 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4,0411 \\ -5,3091 \\ -0,7525 \\ 3,4071 \\ -5,8166 \\ -1,1834 \end{bmatrix} = 4,0411 - 2/3\left(5,3091 + 0,7525\right) = 0$$

VII-5.8.3. Presentación de los valores genéticos de los animales

Ya ha sido comentado que la valoración genética proporciona predictores de los valores genéticos de los individuos con la media de los fundadores igual a cero. En poblaciones sometidas a selección que posean un profundo historial de pedigrí, los valores genéticos de los individuos de la generación más actual son más elevados, por lo que conviene cambiar la escala en la presentación y considerar arbitrariamente como base los animales nacidos, por ejemplo, un par de generaciones anteriores a la actual.

Sin embargo, en poblaciones con un historial reciente, los valores genéticos de los individuos candidatos a la selección pueden presentar valores negativos. En concreto, si nunca se ha llevado a cabo selección en esta población, la mitad de ellos aproximadamente presentarán valores positivos y la otra mitad valores negativos. Así, para intensidades de selección inferiores al

50% entre los individuos seleccionados se encontrarán individuos con valor negativo.

Para evitar el efecto que puedan tener los valores negativos sobre el responsable de conducir la selección, se suele hacer un cambio de escala. La media y la varianza de la nueva variable transformada pueden ser elegidas arbitrariamente. Una transformación frecuente es a media 100 y varianza 400, lo que se corresponde con una desviación típica de 20 ($I_{100} \sim N(100,400)$). La nueva variable ofrece valores positivos para todos los individuos y una variabilidad suficiente para distinguir bien entre ellos. Obsérvese que, dado que en una distribución normal el intervalo que cubre la media más menos 1,96 veces la desviación típica se corresponde con un área de 0,95, aproximadamente el 95% de los individuos se encontraría aproximadamente entre 60 y 140.

La transformación de una variable distribuida normalmente con media cero y varianza 1 a otra variable con la media y varianza que se desee, supone una transformación inversa a la de la tipificación de una variable, es decir, que se obtiene multiplicando por la nueva desviación típica y sumando la nueva media. Dado que los valores genéticos se obtienen con media cero y varianza σ_u^2 ($u_i \sim N(0,\sigma_u^2)$) la tipificación se logra sencillamente dividiendo el valor genético por la desviación típica. Así pues, la transformación se realiza de la siguiente manera:

$$I_{100_i} = \left(\frac{\hat{u}_i}{\sigma_u} \right) 20 + 100 = 100 + \frac{20}{\sigma_u} \hat{u}_i$$

VII-5.9. Medida del error de los valores genéticos

La medida de la precisión se realiza como siempre a partir de la varianza, y en este caso, por tratarse de vectores, por medio de sus matrices de varianzas y covarianzas, en el caso de los efectos fijos, la matriz de varianzas y covarianzas de los estimadores, y en el caso de los aleatorios, a partir de la matriz de varianzas y covarianzas de los errores de predicción.

Partiremos de la expresión utilizada en la resolución de las ecuaciones del modelo mixto:

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z}\\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X'R^{-1}y}\\ \mathbf{Z'R^{-1}y}\end{bmatrix}$$

Llamando **C** a la inversa de la matriz de coeficientes y sacando y como factor fuera del vector del lado derecho, la resolución de las ecuaciones del modelo mixto se puede expresar de la siguiente manera:

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{C_{11}} & \mathbf{C_{12}}\\ \mathbf{C_{12}'} & \mathbf{C_{22}}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}}\\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{y}$$

En esta expresión se ha dividido la matriz **C** en los cuatro bloques llamados $\mathbf{C_{11}}$, $\mathbf{C_{12}}$, $\mathbf{C_{21}}$ y $\mathbf{C_{22}}$, que son los bloques de la inversa de la matriz de coeficientes que en la original correspondían a $\mathbf{X'R^{-1}X}$, $\mathbf{X'R^{-1}Z}$, $\mathbf{Z'R^{-1}X}$ y $\mathbf{Z'R^{-1}Z + G^{-1}}$. Se desarrolla la varianza de esta expresión en la que la expresión matricial que premultiplica al vector de datos **y** es constante, para llegar a la medida de su error.

DERIVACIÓN DE LOS ERRORES DE LAS SOLUCIONES DEL MODELO MIXTO

De la expresión utilizada para resolver el modelo en la que la inversa de la matriz de coeficientes se ha separado en bloques, se pueden aislar de forma independiente las soluciones de los efectos fijos y las de los aleatorios:

$$\hat{\mathbf{b}}=\begin{bmatrix}\mathbf{C_{11}} & \mathbf{C_{12}}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}}\\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{y}$$

$$\hat{\mathbf{u}}=\begin{bmatrix}\mathbf{C_{12}'} & \mathbf{C_{22}}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}}\\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{y}$$

Dado el valor de $Var(\mathbf{y})$ desarrollado en la definición del modelo como $Var(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R}$, se puede calcular la varianza de las soluciones.

$$Var(\hat{\mathbf{b}})=\begin{bmatrix}\mathbf{C_{11}} & \mathbf{C_{12}}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}}\\ \mathbf{Z'R^{-1}}\end{bmatrix}(\mathbf{ZGZ'+R})\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C_{11}}\\ \mathbf{C_{12}'}\end{bmatrix}$$

$$Var(\hat{\mathbf{u}}) = \begin{bmatrix} \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix} (\mathbf{ZGZ}' + \mathbf{R}) \begin{bmatrix} \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{12} \\ \mathbf{C}_{22} \end{bmatrix}$$

Se desarrolla aquí el caso en el que la matriz de coeficientes es de rango completo aunque no es necesario que se cumpla esta condición para mantener la validez de los resultados. Dado que **C** es la inversa de la matriz de coeficientes se cumple:

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Se desarrollan algunos pasos algebraicos previos para obtener algunas igualdades que necesitaremos después:

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \left\{ \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{C}_{12}\mathbf{G}^{-1} \\ \mathbf{0} & \mathbf{C}_{22}\mathbf{G}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{C}_{12}\mathbf{G}^{-1} \\ \mathbf{0} & \mathbf{C}_{22}\mathbf{G}^{-1} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{C}_{12}\mathbf{G}^{-1} \\ \mathbf{0} & \mathbf{I} - \mathbf{C}_{22}\mathbf{G}^{-1} \end{bmatrix}$$

Del producto de los distintos bloques se obtienen cuatro igualdades que serán de utilidad:

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix} \mathbf{X} = \mathbf{I}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix} \mathbf{Z} = -\mathbf{C}_{12}\mathbf{G}^{-1}$$

$$\begin{bmatrix} \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix} \mathbf{X} = \mathbf{0}$$

$$\begin{bmatrix} \mathbf{C}_{12}' & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1} \end{bmatrix} \mathbf{Z} = \mathbf{I} - \mathbf{C}_{22}\mathbf{G}^{-1}$$

El álgebra que se desarrolla a continuación no se va a explicar detalladamente pero puede razonarse sin excesivo esfuerzo.

- A continuación se desarrolla en primer lugar la varianza de los BLUE de los efectos fijos:

$$Var(\hat{\mathbf{b}}) = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}(\mathbf{ZGZ'+R})\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{12}'\end{bmatrix} =$$

$$= \left(\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{Z}\right)\mathbf{G}\left(\mathbf{Z'}\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{12}'\end{bmatrix}\right) +$$

$$+\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{R}\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{12}'\end{bmatrix} =$$

$$= -\mathbf{C}_{12}\mathbf{G^{-1}G}\left(-\mathbf{C}_{12}\mathbf{G^{-1}}\right)' +$$

$$+\left[\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{X} \quad \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{Z}\right]\begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{12}'\end{bmatrix} =$$

$$= \mathbf{C}_{12}\mathbf{G^{-1}C}_{12}' + \begin{bmatrix}\mathbf{I} & -\mathbf{C}_{12}\mathbf{G^{-1}}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{12}'\end{bmatrix} =$$

$$= \mathbf{C}_{12}\mathbf{G^{-1}C}_{12}'\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{G^{-1}C}_{12}' = \mathbf{C}_{11}$$

Por tanto, en primer lugar, en cuanto a los efectos fijos:

$$\boxed{Var(\hat{\mathbf{b}}) = \mathbf{C}_{11}}$$

- Y de un modo similar se desarrolla la varianza de los BLUP de los efectos aleatorios:

$$Var(\hat{\mathbf{u}}) = \begin{bmatrix}\mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}(\mathbf{ZGZ'+R})\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{12} \\ \mathbf{C}_{22}\end{bmatrix} =$$

$$= \left(\begin{bmatrix}\mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{Z}\right)\mathbf{G}\left(\mathbf{Z'}\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{12} \\ \mathbf{C}_{22}\end{bmatrix}\right) +$$

$$+\begin{bmatrix}\mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{R}\begin{bmatrix}\mathbf{R^{-1}X} & \mathbf{R^{-1}Z}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{12} \\ \mathbf{C}_{22}\end{bmatrix} =$$

$$= \left(\mathbf{I} - \mathbf{C}_{22}\mathbf{G^{-1}}\right)\mathbf{G}\left(\mathbf{I} - \mathbf{C}_{22}\mathbf{G^{-1}}\right)' +$$

$$+\left[\begin{bmatrix}\mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{X} \quad \begin{bmatrix}\mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}}\end{bmatrix}\mathbf{Z}\right]\begin{bmatrix}\mathbf{C}_{12} \\ \mathbf{C}_{22}\end{bmatrix} =$$

$$= \left(\mathbf{G} - \mathbf{C}_{22}\right)\left(\mathbf{I} - \mathbf{G^{-1}C}_{22}\right) + \begin{bmatrix}\mathbf{0} & \mathbf{I} - \mathbf{C}_{22}\mathbf{G^{-1}}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{12} \\ \mathbf{C}_{22}\end{bmatrix} =$$

$$= \mathbf{G} - 2\mathbf{C}_{22} + \mathbf{C}_{22}\mathbf{G^{-1}C}_{22} + \mathbf{C}_{22} - \mathbf{C}_{22}\mathbf{G^{-1}C}_{22} = \mathbf{G} - \mathbf{C}_{22}$$

Y entonces, en relación a los efectos aleatorios:

$$Var(\hat{\mathbf{u}}) = \mathbf{G} - \mathbf{C}_{22}$$

- Se deja como ejercicio para el lector el desarrollo de las covarianzas entre los estimadores de los efectos fijos y los predictores de los efectos aleatorios. Estas resultan en matrices nulas:

$$Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}}) = \mathbf{0}$$

Pero para los efectos aleatorios lo que interesa conocer es la varianza del error de predicción. Para ello se recuerda que durante el desarrollo del BLP se mostró que $Cov(\hat{\mathbf{u}},\mathbf{u}) = Cov(\mathbf{u},\hat{\mathbf{u}}) = Var(\hat{\mathbf{u}})$:

$$Var(\hat{\mathbf{u}} - \mathbf{u}) = Var(\hat{\mathbf{u}}) + Var(\mathbf{u}) - Cov(\hat{\mathbf{u}},\mathbf{u}) - Cov(\mathbf{u},\hat{\mathbf{u}})$$
$$Var(\hat{\mathbf{u}} - \mathbf{u}) = Var(\mathbf{u}) - Var(\hat{\mathbf{u}}) = \mathbf{G} - (\mathbf{G} - \mathbf{C}_{22}) = \mathbf{C}_{22}$$

Así que finalmente podemos escribir:

$$Var\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}-\mathbf{C}_{22}\end{bmatrix} \quad \Rightarrow \quad Var\begin{bmatrix}\hat{\mathbf{b}}-\mathbf{b}\\ \hat{\mathbf{u}}-\mathbf{u}\end{bmatrix} = Var\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}-\mathbf{u}\end{bmatrix} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}$$

Es importante destacar cómo a medida que se aumenta el número de observaciones la varianza de los predictores $Var(\hat{\mathbf{u}})$ aumenta hasta un máximo de **G**, mostrando así que no es el parámetro adecuado para medir la falta de calidad del predictor. Por el contrario, la varianza del error de predicción $Var(\hat{\mathbf{u}} - \mathbf{u})$ disminuye hasta un mínimo de **0**.

Así pues, los errores de las soluciones se encuentran directamente en la inversa de la matriz de coeficientes:

$$Var\begin{bmatrix}\hat{\mathbf{b}}-\mathbf{b}\\ \hat{\mathbf{u}}-\mathbf{u}\end{bmatrix} = Var\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}-\mathbf{u}\end{bmatrix} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{12}' & \mathbf{C}_{22}\end{bmatrix}$$

Dado que la inversa de la matriz de coeficientes es una operación obligatoria para resolver el modelo, siempre tenderemos disponibles los errores de cada solución en su diagonal, a menos que se utilicen métodos de resolución indirectos, en cuyo caso, se han desarrollado métodos que aproximan dicha diagonal.

VII-5.10. Transformación de la varianza del error de predicción en precisión

La varianza del error de predicción de las valoraciones es de difícil interpretación por lo que las valoraciones genéticas suelen proporcionarse acompañadas preferentemente de su precisión (correlación entre el verdadero valor y el predictor) o de fiabilidad (el cuadrado de dicha correlación).

La relación entre la varianza del error de predicción y la precisión ya fue expuesta en el capítulo correspondiente a la valoración genética mediante BLP:

$$\boldsymbol{\rho}^2_{\hat{\mathbf{u}}\mathbf{u}} = \mathbf{I} - VEP\mathbf{G}^{-1}$$

Esta expresión puede ser detallada para un individuo concreto de la población:

$$\rho^2_{\hat{u}_i u_i} = 1 - \frac{VEP_i}{\sigma^2_u}$$

En este momento conviene recordar que en las ecuaciones simplificadas del BLUP la matriz de coeficientes había sido multiplicada por la varianza residual (σ^2_e) para eliminar $\mathbf{R}^{-1}$, por lo que en su inversa los elementos de la diagonal deberían ser multiplicados nuevamente por σ^2_e para obtener la *VEP* de cada individuo en la diagonal. Si llamamos c^{ii} al elemento diagonal de la inversa de la matriz de coeficientes de las ecuaciones simplificadas del modelo mixto, entonces $VEP_i = c^{ii}\ \sigma^2_e$ y esta expresión quedaría:

$$\rho^2_{\hat{u}_i u_i} = 1 - \frac{c^{ii}\sigma^2_e}{\sigma^2_u}$$

Y dado que $\alpha = \frac{\sigma^2_e}{\sigma^2_u}$:

$$\rho^2_{\hat{u}_i u_i} = 1 - c^{ii}\alpha$$

Al igual que durante la construcción de las ecuaciones del modelo mixto, para proporcionar la precisión de los valores genéticos no

es preciso conocer la varianza de cada componente aleatorio sino que basta con conocer la heredabilidad.

VII-5.10.1. Precisión de los animales del ejemplo

La inversa de la matriz de coeficientes del ejemplo era la siguiente:

$$\left[\begin{array}{c|c} \mathbf{X'X} & \mathbf{X'Z} \\ \hline \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\alpha \end{array}\right]^{-1} =$$

$$= \left[\begin{array}{cc|cccccc} 1,4591 & 0,8282 & -0,1012 & -1,0654 & -1,0328 & -0,6845 & -0,9544 & -0,8456 \\ 0,8282 & 1,4069 & -0,3958 & -0,8544 & -0,8020 & -1,0208 & -1,0672 & -1,1238 \\ \hline -0,1012 & -0,3958 & 1,3651 & 0,0794 & 0,1230 & 0,5873 & 0,2274 & 0,3726 \\ -1,0654 & -0,8544 & 0,0794 & 1,2879 & 0,8429 & 0,7631 & 0,9750 & 0,8250 \\ -1,0328 & -0,8020 & 0,1230 & 0,8429 & 1,2226 & 0,6059 & 0,9339 & 0,8661 \\ -0,6845 & -1,0208 & 0,5873 & 0,7631 & 0,6059 & 1,2625 & 0,8287 & 0,9713 \\ -0,9544 & -1,0672 & 0,2274 & 0,9750 & 0,9339 & 0,8287 & 1,4352 & 0,9648 \\ -0,8456 & -1,1238 & 0,3726 & 0,8250 & 0,8661 & 0,9713 & 0,9648 & 1,4352 \end{array}\right]$$

Se puede extraer la información de la *VEP* de cada individuo a partir de la diagonal para expresar el grado de confianza en las valoraciones genéticas. Los resultados se presentan en las siguientes tablas:

	$\hat{u}_i$	I_{100} *para* $\sigma_u = 4$	c^{ii}
u_1	4,0411	120	1,3651
u_2	-5,3091	73	1,2879
u_3	-0,7525	96	1,2226
u_4	3,4071	117	1,2625
u_5	-5,8166	71	1,4352
u_6	-1,1834	94	1,4352

	VEP_i	ρ_i^2	ρ_i	ρ_i^2 *(sin fijos)*	ρ_i *(sin fijos)*

u_1	$1{,}3651\,\sigma_e^2$	0,0900	0,3000	0,1751	0,4184
u_2	$1{,}2879\,\sigma_e^2$	0,1414	0,3760	0,6860	0,8282
u_3	$1{,}2226\,\sigma_e^2$	0,1849	0,4300	0,6871	0,8289
u_4	$1{,}2625\,\sigma_e^2$	0,1583	0,3979	0,6565	0,8103
u_5	$1{,}4352\,\sigma_e^2$	0,0432	0,2979	0,6567	0,8104
u_6	$1{,}4352\,\sigma_e^2$	0,0432	0,2979	0,6629	0,8142

Se ha añadido en esta tabla una columna con los valores genéticos de los individuos expresados con media 100 y varianza 400 para familiarizar al lector con este tipo de índices. Para ello se ha asumido arbitrariamente una varianza genética aditiva de 16.

Varios puntos son destacables en esta tabla:

- La precisión expresada como repetibilidad del mérito genético (ρ_i) es bastante más elevada que cuando se expresa en forma de fiabilidad (ρ_i^2).
- La precisión de los individuos con más información es superior al resto. Así los individuos del centro del pedigrí utilizan para su valoración genética su propio dato, la de ascendientes y la de descendientes, tendiendo a tener mayor precisión en la valoración genética que los otros.
- La decisión sobre la selección de un individuo como reproductor puede depender de la fiabilidad. En principio, el mejor animal es el 1, pero el 4 tiene un valor genético similar y tal vez preferible ya que la precisión del primero es del 30% mientras que la del segundo es un 43%.

Según se vio en el desarrollo de los índices de selección, cuando la fuente de información utilizada como criterio es el dato del propio individuo, la precisión expresada en forma de fiabilidad (ρ_i^2) es igual a la heredabilidad del carácter, siendo la precisión mayor a medida que se incremente la información de parientes. La heredabilidad del carácter es en este caso de 0,60 y sin embargo las fiabilidades se encuentran entre 0,04 y 0,18. La razón de esta

discrepancia es la existencia de efectos fijos mal estimados. En todos los casos el dato del individuo es asumido medido sin error. En el caso de los índices de selección el ajuste de los efectos fijos se hace previamente asumiendo que se conocen los verdaderos valores, y por tanto carentes de incertidumbre. Sin embargo, en la metodología BLUP se tiene en cuenta que el ajuste se hace con un efecto fijo que se estima simultáneamente, de modo que al mismo tiempo se tiene en cuenta que, aunque el dato está medido sin error, no es éste el caso para el dato ajustado para los efectos fijos del modelo, el cual lleva el error del propio efecto fijo que se le ajusta. De esta conclusión debe extraerse la importancia de definir modelos en los que los distintos niveles de los efectos fijos cuenten con el número de observaciones suficiente como para que sea bien estimado, ya que la precisión de esta estimación repercutirá en la precisión de la valoración genética de los individuos incluidos en ese nivel del efecto fijo. En las dos últimas columnas de la tabla se ha incorporado la precisión obtenida a partir de la diagonal de la inversa de una matriz de coeficientes en la que los efectos fijos han sido eliminados. Se observa cómo en este caso, en el que la heredabilidad es alta, los individuos que tienen dato propio presentan mucha mayor precisión que el 1 que no dispone de dato propio. Por otro lado la precisión obtenida con el dato propio medida en fiabilidad sería la de la propia heredabilidad (0,60). Se observa que en un carácter de heredabilidad alta como es éste, el tener dato propio proporciona ya una buena precisión y la información de parientes añade poco a la precisión de la valoración.

CONCEPTOS CLAVE

- ¿Qué tipo de información de parentesco se utiliza en la valoración genética mediante el BLUP?
- ¿Qué propiedad estadística añadida tiene el BLUP con respecto a un índice de selección? ¿Cómo se logra?
- ¿Cómo se llama la matriz que incorpora la información de pedigrí en las ecuaciones del modelo mixto?
- ¿Cómo se define el coeficiente de consanguinidad de un individuo?
- ¿Qué valores se encuentran en las diagonales de la matriz **A**?
- ¿Qué valor encontramos en la posición de la matriz **A** donde se cruzan un padre y un hijo sin ninguna otra relación de parentesco?
- ¿Qué indica un valor negativo en la matriz **A** y en su inversa?
- ¿Cuándo las reglas de Henderson para construir la inversa de **A** son sólo aproximadas?
- ¿Qué coeficiente se utiliza para ponderar la información de parentesco? ¿Para qué valores de heredabilidad es este parámetro más elevado o más reducido?
- ¿Cómo se calculan los elementos del lado derecho de las ecuaciones del modelo mixto?
- ¿Cómo se puede interpretar cada ecuación del modelo mixto? ¿Cómo se elimina en la práctica el sesgo?
- ¿Son únicas las soluciones que se obtienen para los efectos fijos? ¿Siempre?
- ¿Para qué individuos se obtiene un valor genético? ¿Para cuáles no se obtiene?
- ¿La media de las soluciones de los valores genéticos es nula?
- ¿Cómo se puede transformar el valor genético para que tengan media 100 y varianza 400? ¿Y para que tengan media 10 y varianza 4?

- ¿Dónde encontramos la medida de la precisión de los valores genéticos?
- ¿Es independiente la precisión de la distribución de los datos en los niveles de los efectos fijos?

OCTAVA PARTE

MODELOS PARTICULARES DE VALORACIÓN GENÉTICA

RESUMEN

Este capítulo supone un acercamiento a los modelos que se usan en realidad, de manera que se introducen con ejemplos. Como ejemplo de aplicación de un nuevo efecto aleatorio como el ambiental permanente se utiliza la producción de leche, y como ejemplo de aplicación del efecto materno se emplea el peso al destete. Con la exposición de estos modelos se aprovecha para comentar cómo pueden considerarse los efectos que participan de la parte fija del modelo. Otros casos particulares habitualmente utilizados en la práctica se exponen a continuación, como el empleo de grupos genéticos para tener en cuenta la ausencia de pedigrí en animales nacidos en distintas fechas o con distintos orígenes, o la valoración genética conjunta para varios caracteres relacionados. Se concluye con la exposición del modelo padre como herramienta que puede aún ser útil en determinados contextos.

VIII-1. Modelos particulares de valoración genética.

Se presentan a continuación algunos ejemplos de modelos utilizados en circunstancias concretas. Sólo se comentan las diferencias con respecto al modelo general tratado ya con todo detalle en la definición general del modelo lineal mixto. Se acompañan de ejemplos de aplicación y se aprovecha para comentar algunas particularidades de cada situación. Por ello, no pretenden mostrarse como soluciones definitivas a problemas concretos. Al contrario, con los conocimientos adquiridos, el gestor de las poblaciones objeto de mejora deberá estudiar su situación para esmerarse en definir el modelo más apropiado para su situación particular.

En algunos de los modelos se discutirá sobre la parte fija del modelo como ejemplo de la forma de razonar en situaciones concretas pero se recuerda una vez más que cada situación debe estudiarse de forma específica.

VIII-2. El modelo padre

Este modelo, conocido también como modelo macho o modelo semental, fue ideado con la intención de reducir considerablemente las dimensiones de las matrices de coeficientes en la valoración genética. Su desarrollo se llevó a cabo en el contexto de la mejora genética del ganado vacuno lechero. En esta población el interés se centraba en la valoración genética de los machos que se encontraban en los centros de inseminación artificial, lo que suponía una importantísima reducción en el número de ecuaciones si se renunciaba a valorar a las vacas.

VIII-2.1. Definición del modelo

La base teórica del modelo consiste en asignar el dato al padre del individuo al que pertenece, es decir, a la mitad del valor genético

que lo origina. La ecuación del modelo es similar a la utilizada en el modelo animal:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_s s} + \mathbf{e}$$

Para seguir la explicación se hará uso del ejemplo que se presenta a continuación. Se trata de un carácter que sólo puede ser medido en hembras y se presenta inicialmente el pedigrí de todos los individuos:

Pedigrí original			**Pedigrí modelo macho**				
Animal	**Padre**	**Madre**	**Macho**	**Padre del macho**	**Abuelo materno del macho**	**Ganadería**	**Peso**
1	-	-	-	-	-	-	-
2	1	-	1	-	-	1	415
3	1	2	-	-	-	-	-
4	1	2	1	-	-	1	420
5	3	2	-	-	-	-	-
6	3	4	3	1	1	1	412
7	3	4	-	-	-	-	-
8	3	4	3	1	1	2	407
9	5	4	-	-	-	-	-
10	5	6	5	3	1	2	411
11	7	6	-	-	-	-	-
12	7	8	7	3	1	2	419
13	7	8	-	-	-	-	-
14	9	10	9	5	1	2	438

Originalmente el modelo presentaba un vector de 14 incógnitas del efecto genético aditivo. En el modelo macho, sin embargo, sólo hay cinco padres con hijos, de forma que el vector de incógnitas aleatorias tiene únicamente esa dimensión. Las ecuaciones quedan de la siguiente manera:

$$\begin{bmatrix} 415 \\ 420 \\ 412 \\ 407 \\ 411 \\ 419 \\ 438 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_3 \\ s_5 \\ s_7 \\ s_9 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{bmatrix}$$

La definición de las esperanzas y varianzas del modelo también es similar y también presenta diferencias:

$$E\begin{bmatrix} \mathbf{y} \\ \mathbf{b} \\ \mathbf{s} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{Xb} \\ \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \qquad Var\begin{bmatrix} \mathbf{y} \\ \mathbf{s} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{Z_s G_s Z_s' + R} & \mathbf{Z_s G_s} & \mathbf{R} \\ \mathbf{G_s Z_s'} & \mathbf{G_s} & \mathbf{0} \\ \mathbf{R} & \mathbf{0} & \mathbf{R} \end{bmatrix}$$

siendo: $\mathbf{G_s} = \mathbf{A_s}\sigma_s^2$

La diferencia con respecto al modelo animal está en la varianza del efecto aleatorio. $\mathbf{A_s}$ representa la matriz numerador de relaciones aditivas entre machos y σ_s^2 es la varianza entre los machos. Si llamamos *j* al padre del animal *i*, sólo la mitad del valor genético del individuo está ajustado en el modelo por el efecto padre, $s_i = \frac{1}{2} u_i$.

VIII-2.2. Las ecuaciones del modelo mixto del modelo padre

La resolución del modelo padre es en todo similar a la del modelo animal:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_s} \\ \mathbf{Z_s'X} & \mathbf{Z_s'Z_s} + \mathbf{A_s^{-1}}\alpha_s \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_s'y} \end{bmatrix}$$

El valor de α_s coincide, como siempre, con el cociente entre la varianza residual y la varianza del efecto, en este caso σ_s^2. Dado que s_j = ½ u_i, tomando la varianza a ambos lados se obtiene que $\sigma_{s_j}^2$ = ¼ $\sigma_{u_i}^2$, y la varianza de ambos elementos es la de la distribución correspondiente: σ_s^2 = ¼ σ_u^2. Por otro lado, según la ecuación del modelo, la varianza fenotípica sólo tiene dos componentes, la varianza entre machos y la varianza residual: $\sigma_p^2 = \sigma_s^2 + \sigma_e^2 \Rightarrow \sigma_e^2 = \sigma_p^2 - \sigma_s^2$. Podemos entonces desarrollar α_s para dejarlo en función de la heredabilidad:

$$\alpha_s = \frac{\sigma_e^2}{\sigma_s^2} = \frac{\sigma_p^2 - \sigma_s^2}{\sigma_s^2} = \frac{\sigma_p^2 - \sigma_u^2/4}{\sigma_u^2/4} = \frac{4\sigma_p^2 - \sigma_u^2}{\sigma_u^2} =$$

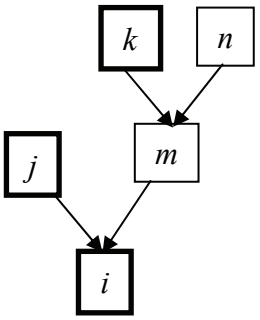
$$= \frac{\left(4\sigma_p^2 - \sigma_u^2\right)/\sigma_p^2}{\sigma_u^2/\sigma_p^2} = \frac{4\,\sigma_p^2/\sigma_p^2 - \sigma_u^2/\sigma_p^2}{\sigma_u^2/\sigma_p^2} = \frac{4 - h^2}{h^2}$$

Así pues, tampoco en el modelo padre es necesario conocer los componentes de varianza sino que el valor de α_s puede obtenerse en función de la heredabilidad. Para el ejemplo asumiremos una heredabilidad de 0,25:

$$\alpha_s = \frac{\sigma_e^2}{\sigma_s^2} = \frac{4 - h^2}{h^2} = \frac{4 - 0,25}{0,25} = 15$$

VIII-2.3. La inversa de la matriz de relaciones aditivas entre machos.

Representamos a continuación la situación en la que el individuo *i* (macho) es hijo de *j* (macho) y *m* (hembra). A su vez *m* es hija de *k* (macho) y *n* (hembra).



Cuando se utiliza un modelo en el que no existen hembras, el valor genético aditivo del individuo no puede ser descompuesto en función de sus dos padres. Originalmente esta descomposición sería de la siguiente manera:

$$u_i = \tfrac{1}{2} u_j + \tfrac{1}{2} u_m + \varphi_i$$

Ya que las hembras no van a ser tomadas en consideración, podemos escribir esta expresión para el individuo *m*, madre de *i*:

$$u_m = \tfrac{1}{2} u_k + \tfrac{1}{2} u_n + \varphi_m$$

Y sustituir el valor de u_m en la expresión anterior:

$$\begin{aligned} u_i &= \tfrac{1}{2} u_j + \tfrac{1}{2}\left(\tfrac{1}{2} u_k + \tfrac{1}{2} u_n + \varphi_m\right) + \varphi_i = \\ &= \tfrac{1}{2} u_j + \tfrac{1}{4} u_k + \tfrac{1}{4} u_n + \tfrac{1}{2} \varphi_m + \varphi_i = \\ &= \tfrac{1}{2} u_j + \tfrac{1}{4} u_k + \varphi_i^* \end{aligned}$$

Se ha establecido $\varphi_i^* = \frac{1}{4}u_n + \frac{1}{2}\varphi_m + \varphi_i$ para aproximar el valor genético aditivo de cada padre mediante la mitad del valor genético de su padre y la cuarta parte del valor genético de su abuelo materno.

Con esta premisa y siguiendo los mismos desarrollos empleados para la inversa de la matriz de relaciones aditivas se llega a una nueva tabla de coeficientes a ser aplicada para cada individuo en el contexto del modelo padres. A su vez existe también una pequeña modificación de la expresión que nos lleva a las diagonales de la matriz **D**:

	i	**j**	**k**
i	1	$-\frac{1}{2}$	$-\frac{1}{4}$
j	$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
k	$-\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$

x d_{ii}^{-1}; siendo $d_{ii} = 1 - \frac{1}{4}a_{jj} - \frac{1}{16}a_{kk}$

VIII-2.3.1. Reglas de Henderson para la construcción de $\mathbf{A_S^{-1}}$ aproximada

Como en el caso del modelo animal, en $d_{ii} = 1 - \frac{1}{4}a_{jj} - \frac{1}{16}a_{kk}$ se asume que la consanguinidad de *j* y *k* es nula en todos los casos, entonces los valores de d_{ii} y de su inverso d_{ii}^{-1} dependerán únicamente de si *j* y *k* son o no conocidos:

			d_{ii} =	d_{ii}^{-1} =
d_{ii} =	*j* y *k* desconocidos	$a_{jj} = a_{kk} = 0$	1	1
	j conocido y *k* desconocido	$a_{jj} = 1; a_{kk} = 0$	$^{3}/_{4}$	$^{4}/_{3}$
	j desconocido y *k* conocido	$a_{jj} = 0; a_{kk} = 1$	$^{15}/_{16}$	$^{16}/_{15}$
	j y *k* conocidos	$a_{jj} = a_{kk} = 1$	$^{11}/_{16}$	$^{16}/_{11}$

Utilizando estos valores de d_{ii}^{-1} en la tabla anterior, y teniendo en cuenta que los coeficientes que afectan a individuos desconocidos no pueden anotarse, se deducen las reglas de Henderson para la construcción de la inversa de la matriz de relaciones aditivas entre machos.

En primer lugar se crea una tabla de *n* filas y *n* columnas siendo en este caso *n* el número de machos. A continuación se recorre el registro de pedigrí de los machos, en el que la madre ha sido reemplazada por su padre, es decir, el abuelo materno del individuo. Se anotan los siguientes coeficientes dependiendo del caso en el que se encuentre cada uno de los machos:

- Si se desconocen el padre y el abuelo materno de *i*:
  - Se anota un 1 en la diagonal del individuo, en la posición (*i*, *i*).
- Si sólo se conoce el padre, *j*, y se desconoce el abuelo materno:
  - Se anota $^{4}/_{3}$ en la diagonal del individuo (*i*, *i*).
  - $^{-2}/_{3}$ donde se cruza el individuo *i* con el padre conocido *j*, en los dos lugares donde se da el cruce: (*i*, *j*) y (*j*, *i*).
  - $^{1}/_{3}$ en la diagonal del padre conocido: (*j*, *j*).

- Si sólo se desconoce el padre pero se conoce el abuelo materno *k*:
  - Se anota $^{16}/_{15}$ en la diagonal del individuo (*i*, *i*).
  - $^{-4}/_{15}$ donde se cruza el individuo *i* con el abuelo materno *k*, en los dos lugares donde se da el cruce: (*i*, *k*) y (*k*, *i*).
  - $^{1}/_{15}$ en la diagonal del padre conocido: (*j*, *j*).
- Si se conocen tanto el padre *j* como el abuelo materno *k* se añade:
  - $^{16}/_{11}$ en el elemento (*i*, *i*).
  - $^{-8}/_{11}$ en los cruces entre el individuo *i* y su padre *j* en los dos lugares donde se da el cruce: (*i*, *j*) y (*j*, *i*).
  - $^{-4}/_{11}$ en los cruces entre el individuo *i* y su abuelo materno *k* en los dos lugares donde se da el cruce: (*i*, *k*) y (*k*, *i*).
  - $^{4}/_{11}$ en la diagonal del padre: (*j*, *j*).
  - $^{1}/_{11}$ en la diagonal del abuelo materno: (*k*, *k*).
  - $^{2}/_{11}$ en el cruce entre el padre y el abuelo materno en los dos lugares en que se da: (*j*, *k*) y (*k*, *j*).

Se utilizan estas reglas en el pedigrí de los machos:

Pedigrí modelo macho		
Macho	**Padre del macho**	**Abuelo materno del macho**
1	-	-
3	1	1
5	3	1
7	3	1
9	5	1

Y una vez anotados todos los coeficientes la tabla utilizada para construir la inversa de la matriz de relaciones aditivas quedaría así:

	1	3	5	7	9
1	$1+\frac{4}{11}+\frac{1}{11}+\frac{2}{11}+\frac{2}{11}+\frac{1}{11}+\frac{1}{11}$	$-\frac{8}{11}\ -\frac{4}{11}+\frac{2}{11}+\frac{2}{7}$	$-\frac{4}{11}+\frac{2}{11}$	$-\frac{4}{11}$	$-\frac{4}{11}$
3	$-\frac{8}{11}\ -\frac{4}{11}+\frac{2}{11}+\frac{2}{11}$	$\frac{16}{11}+\frac{4}{11}+\frac{4}{11}$	$-\frac{8}{11}$	$-\frac{8}{11}$	
5	$-\frac{4}{11}+\frac{2}{11}$	$-\frac{8}{11}$	$\frac{16}{11}+\frac{4}{11}$		$-\frac{8}{11}$
7	$-\frac{4}{11}$	$-\frac{8}{11}$		$\frac{16}{11}$	
9	$-\frac{4}{11}$		$-\frac{8}{11}$		$\frac{16}{11}$

Y la inversa de $\mathbf{A}_s$ y su producto por a_s quedan:

$$\mathbf{A}_s^{-1}\alpha_s = \begin{bmatrix} \frac{23}{11} & -\frac{8}{11} & -\frac{2}{11} & -\frac{4}{11} & -\frac{4}{11} \\ -\frac{8}{11} & \frac{24}{11} & -\frac{8}{11} & -\frac{8}{11} & 0 \\ -\frac{2}{11} & -\frac{8}{11} & \frac{20}{11} & 0 & -\frac{8}{11} \\ -\frac{4}{11} & -\frac{8}{11} & 0 & \frac{16}{11} & 0 \\ -\frac{4}{11} & 0 & -\frac{8}{11} & 0 & \frac{16}{11} \end{bmatrix} 15 =$$

$$= \begin{bmatrix} \frac{345}{11} & -\frac{120}{11} & -\frac{30}{11} & -\frac{60}{11} & -\frac{60}{11} \\ -\frac{120}{11} & \frac{360}{11} & -\frac{120}{11} & -\frac{120}{11} & 0 \\ -\frac{30}{11} & -\frac{120}{11} & \frac{300}{11} & 0 & -\frac{120}{11} \\ -\frac{60}{11} & -\frac{120}{11} & 0 & \frac{240}{11} & 0 \\ -\frac{60}{11} & 0 & -\frac{120}{11} & 0 & \frac{240}{11} \end{bmatrix}$$

VIII-2.4. Resolución de las ecuaciones del modelo mixto del modelo padre.

Finalmente las ecuaciones del modelo mixto, a falta de sumar $\mathbf{A_s^{-1}}\alpha_s$ quedan:

	μ	g_1	g_2	s_1	s_3	s_5	s_7	s_9
μ	7	3	4	2	2	1	1	1
g_1	3	3	0	2	1	1	0	0
g_2	4	0	4	0	1	1	1	1
s_1	2	2	0	2	0	0	0	0
s_3	2	1	0	0	2	0	0	0
s_5	1	1	1	0	0	1	0	0
s_7	1	0	1	0	0	0	1	0
s_9	1	0	1	0	0	0	0	1

μ	2922
g_1	1247
g_2	1675
s_1	835
s_3	819
s_5	411
s_7	419
s_9	438

Se construyen las ecuaciones del modelo mixto:

$$\begin{bmatrix} 7 & 3 & 4 & 2 & 2 & 1 & 1 & 1 \\ 3 & 3 & 0 & 2 & 1 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 345/11 & -120/11 & -30/11 & -60/11 & -60/11 \\ 2 & 1 & 1 & -120/11 & 382/11 & -120/11 & -120/11 & 0 \\ 1 & 0 & 1 & -30/11 & -120/11 & 322/11 & 0 & -120/11 \\ 1 & 0 & 1 & -60/11 & -120/11 & 0 & 262/11 & 0 \\ 1 & 0 & 1 & -60/11 & 0 & -120/11 & 0 & 262/11 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{s}_1 \\ \hat{s}_3 \\ \hat{s}_5 \\ \hat{s}_7 \\ \hat{s}_9 \end{bmatrix} = \begin{bmatrix} 2922 \\ 1247 \\ 1675 \\ 835 \\ 819 \\ 411 \\ 419 \\ 438 \end{bmatrix}$$

Para resolver se puede hacer cero por ejemplo la media y obtener el resto. El resultado se presenta a continuación junto con el que se obtiene del modelo animal equivalente. En el segundo caso se ha dividido el valor genético de los individuos por dos para hacer los resultados comparables con los obtenidos con el modelo padre:

$$\begin{bmatrix} \hat{\mu}^0 \\ \hat{g}_1^0 \\ \hat{g}_2^0 \\ \hat{s}_1 \\ \hat{s}_3 \\ \hat{s}_5 \\ \hat{s}_7 \\ \hat{s}_9 \end{bmatrix} = \begin{bmatrix} 0 \\ 415,9798 \\ 418,9287 \\ -0,1157 \\ -0,7050 \\ -0,3087 \\ -0,3616 \\ 0,6605 \end{bmatrix} \qquad \begin{bmatrix} \hat{\mu}^0 \\ \hat{g}_1^0 \\ \hat{g}_2^0 \\ \hat{u}_{1/2} \\ \hat{u}_{3/2} \\ \hat{u}_{5/2} \\ \hat{u}_{7/2} \\ \hat{u}_{9/2} \end{bmatrix} = \begin{bmatrix} 0 \\ 415,7004 \\ 418,4146 \\ -0,0554 \\ -0,2967 \\ 0,2363 \\ -0,0051 \\ 0,8785 \end{bmatrix}$$

Obviamente los resultados son sensiblemente distintos dado que se dispone de pocos datos y todas las relaciones genéticas entre individuos que no sean establecidas a través de los machos, son ignoradas. Sin embargo, aunque no ha sido mostrado, el número de ecuaciones en el modelo animal es doble para el ejemplo mostrado. Cuando los datos proceden de un esquema basado en centros de inseminación artificial, esta diferencia en las dimensiones de las ecuaciones es mucho mayor.

El modelo padre ha caído en desuso como método de valoración genética ya que, tanto la memoria de los ordenadores como el tiempo de cálculo, han aumentado enormemente. Sin embargo, aún es muy utilizado en la estimación de parámetros genéticos cuando el carácter precisa de medidas múltiples como ocurre por ejemplo en caracteres discretos de carácter subjetivo tal como la facilidad de parto.

VIII-3. Modelos con medidas repetidas

Se llama también modelo de ambiente permanente o modelo de repetibilidad. Este modelo se aplica cuando el carácter puede medirse varias veces en cada individuo. Algunos ejemplos de caracteres en los que se ajusta este modelo son: la cantidad de leche producida, el diámetro de la fibra, los resultados de carreras de caballos, el tamaño de camada o el intervalo entre partos. Por ejemplo, no tendría sentido ajustar este modelo en un carácter

como el peso a la canal ya que el animal sólo puede ser sacrificado una vez.

El ejemplo más paradigmático de este tipo de modelo es la producción de leche de manera que se utilizará este ejemplo para formalizar el modelo.

VIII-3.1. Efectos fijos ajustados en el carácter cantidad de leche

Es importante conocer lo que supone el ajuste de efectos fijos discontinuos en el modelo. Al ajustar un efecto fijo con distintos niveles se está asumiendo que hay una diferencia real única fija entre los niveles del efecto. Por lo tanto el modelo realiza una estimación de estas diferencias, o lo que es lo mismo, cada uno de los niveles debería tener suficientes datos como los que serían necesarios para estimar una media. Por ejemplo, imaginemos un efecto fijo definido por la ganadería. Una ganadería con pocos datos llevaría a una mala estimación del efecto ganadería. Los animales pertenecientes a la misma se ajustarían mediante un efecto mal estimado. Además, al dato ajustado para la media le acompañaría la propia incertidumbre de la ganadería a la que pertenece y el animal aparecería valorado con muy mala precisión. Por tanto, debe tenerse en cuenta el número de datos dentro de cada nivel de cada efecto fijo en el momento de definir el modelo.

En todos los modelos existen efectos fijos que ajustan las coordenadas de espacio y tiempo. De hecho este efecto era el único que se ajustaba cuando para valorar genéticamente a los animales imperaba la metodología basada en índices de selección. Se restaba a cada dato la media de lo que se conocía como contemporáneas de establo, es decir, los registros que pertenecían a la misma ganadería en la misma época. Este valor representaba la superioridad (o no, cuando su valor no era positivo) que el animal presentaba en relación a los que compartían el mismo espacio al mismo tiempo.

En el caso del vacuno lechero normalmente se trabaja con la leche ordeñada en una longitud de lactación estándar como por ejemplo, 305 días. El efecto fijo que define el espacio es el rebaño. Un efecto fijo es aquél que posee pocos niveles distintos y todos ellos están incluidos en el modelo. En los comienzos de la metodología BLUP el efecto ganadería fue considerado ocasionalmente como aleatorio dado que podía asumirse que existían muchas ganaderías y que sólo unas pocas estaban en nuestros datos. Sin embargo, en la valoración genética de cada año se repetían sistemáticamente la inmensa mayoría de las ganaderías por lo que podía considerarse como fijo. Sin que este debate teórico haya conducido a un consenso se ha impuesto el ajuste de la ganadería como efecto fijo por razones bien diferentes. En concreto, la inclusión de un efecto aleatorio en el modelo implica la necesidad de conocer la varianza que define su distribución. Por ello todos los efectos de difícil asignación a una de las dos categorías se asumen como fijos.

Efectos fijos que ajustan las diferencias en el rendimiento que se dan en distinta época son dos. El primero, el año de parto, pretende ajustar las tendencias en el rendimiento que no son de origen genético. En general las prácticas de manejo suelen evolucionar llevando a animales del mismo nivel genético a consumir más. Por otro lado, los valores genéticos de una población sometida a selección artificial probablemente aumentarán también de año en año. Aunque podría pensarse que el efecto año confunde la tendencia en el manejo con la tendencia genética, la información de parentesco permite separar ambos efectos. El segundo efecto que ajusta las diferencias temporales es la estación. Habitualmente los rendimientos de los individuos son superiores en primavera e inferiores en verano por lo que debe ajustarse un efecto estación de parto.

Por tanto, el modelo que ajusta las producciones de leche tiene estos tres efectos, el rebaño, el año y la estación. Sin embargo, aunque existe una tendencia común, no todos los años afectan por igual a todos los ganaderos. Por ejemplo circunstancias personales de diversa índole pueden modificar el efecto del año en algunos ganaderos tanto a favor suyo como en su contra. Del mismo modo, aunque en verano se registran peores producciones en

general, esta disminución no afectará por igual a las ganaderías en extensivo que aquellas que tienen permanentemente estabulados a sus animales. Este efecto se conoce como interacción entre efectos y debe ajustarse también en los modelos. Para ello se crea un efecto combinado Rebaño-Año-Estación (conocido generalmente como RAE) mediante cuyo inclusión se logra el ajuste de cada uno de los tres así como de su interacción. El número de niveles del efecto combinado será el producto del número de niveles de los efectos originales, por lo que debe vigilarse el número de datos que quedan en cada uno de los niveles del efecto combinado. Por ejemplo, si existen 2 rebaños (R1 y R2), 3 años (A1, A2 y A3), y dos estaciones (E1 y E2), el efecto combinado RAE tendrá 2 x 3 x 2 = 12 niveles:

Rebaño	*Año*	*Estación*	*RAE*
1	1	1	1
1	1	2	2
1	2	1	3
1	2	2	4
1	3	1	5
1	3	2	6
2	1	1	7
2	1	2	8
2	2	1	9
2	2	2	10
2	3	1	11
2	3	2	12

Es conocida la influencia del número de lactación en la producción de leche de manera que los animales de primer parto producen menos que los de segundo y tercer parto, comenzando a reducir muy poco a poco su producción a partir del cuarto o quinto parto. Este efecto tiene dos componentes, la edad del animal y el número de lactación en sí mismo. Los modelos pueden incluir ambos efectos. El número de lactación se suele emplear como efecto fijo discontinuo con 4 o 5 niveles, uno por número de lactación incluyendo el último de ellos las lactaciones posteriores. En cambio el efecto edad del animal al parto suele ajustarse como efecto fijo continuo. Para ajustar la relación no completamente

lineal de la producción con la edad, suele incluirse también como efecto fijo continuo el cuadrado de la edad al parto.

En el caso de las pequeños rumiantes también se suele incluir como efecto fijo el número de crías al parto con tres niveles (1, 2, 3 o más), dado que el mayor número de crías estimula una mayor producción.

VIII-3.2. Definición del modelo con medidas repetidas

Ecuación del modelo:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e}$$

Se ha incluido un nuevo efecto aleatorio, el efecto ambiental permanente. Se trata de un efecto ambiental que representa la parte común que un individuo aporta a todas sus lactaciones, pero que no lo transmite a su descendencia. Un ejemplo burdo sería el de un animal que ha perdido una ubre por algún traumatismo, lo que le afecta con una menos producción a todas sus lactaciones pero que no lo heredarán sus descendientes.

El vector que contiene todos los ambientes permanentes distintos es el vector **p**, y **W** es su matriz de incidencia o matriz diseño que relaciona los datos con el individuo a quien pertenecen. Obsérvese que tanto **Z** como **W** relacionan los datos con el individuo al que pertenecen. Sin embargo, el vector **u** contiene el valor genético aditivo de todos los individuos en el pedigrí (incluidos los machos) aunque no posean dato y por tanto su dimensión es igual al número de individuos en la población. Así, los machos no tendrán producción de leche, pero sí un valor genético para producción de leche que heredaran sus hijas. Sin embargo, únicamente habrá tantos ambientes permanentes distintos como animales con dato y por ello ésta será la dimensión del vector **p**.

Esperanzas y Varianzas del modelo:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_{ep}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

A la vista de la definición de esta parte del modelo se deduce el tipo de efecto que es. Por el hecho de haber sido ajustada una varianza para este efecto se sabe que se trata de un efecto aleatorio. Obsérvese que para obtener la solución de cada nivel del efecto se dispone de pocos datos por individuo, por lo que, de haber sido definido como fijo, cada nivel se habría estimado con muy pocos datos y por tanto, con mucho error. Sin embargo, para un efecto aleatorio el modelo asume que puede obtenerse cualquier valor dentro de su distribución con la única condición de que la media de todos ellos debe valer cero y su varianza, la varianza del efecto. Para formalizar este hecho se ha definido su distribución con media cero ($E(\mathbf{p}) = \mathbf{0}$) y varianza, la varianza ambiental permanente (σ_{ep}^2). También se ha asumido que todos los ambientes permanentes son independientes por lo que todas las covarianzas son nulas. Esto se describe en la estructura de la matriz de varianzas y covarianzas del efecto $\text{Var}(\mathbf{p}) = \mathbf{I}\sigma_{ep}^2$. Finalmente, las matrices nulas que aparecen en las covarianzas entre este efecto aleatorio y el resto describe la independencia de este efecto con respecto al residuo y al valor genético aditivo.

VIII-3.3. Ecuaciones del modelo mixto en modelos de medidas repetidas

Las ecuaciones del modelo mixto a resolver tienen una estructura similar a las vistas previamente:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} & \mathbf{X'Z} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{I}\frac{\sigma_e^2}{\sigma_{ep}^2} & \mathbf{W'Z} \\ \mathbf{Z'X} & \mathbf{Z'W} & \mathbf{Z'Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{p}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Aunque las dimensiones del sistema de ecuaciones crecen, todo lo estudiado previamente para el modelo mixto es aplicado aquí sin incremento de dificultad.

En este caso disponemos de dos cocientes entre varianzas que pueden ser también obtenidos a partir de la heredabilidad (h^2) y de la repetibilidad (R), teniendo en cuenta que la varianza fenotípica (σ_p^2):

$$h^2 = \frac{\sigma_u^2}{\sigma_p^2} \qquad R = \frac{\sigma_u^2 + \sigma_{ep}^2}{\sigma_p^2} \qquad \sigma_p^2 = \sigma_u^2 + \sigma_{ep}^2 + \sigma_e^2$$

Llamaremos α_{ep} al nuevo cociente de varianzas de manteniendo α para el cociente previo. Ambos se obtienen:

$$\alpha = \frac{\sigma_e^2}{\sigma_u^2} = \frac{\sigma_p^2 - \sigma_u^2 - \sigma_{ep}^2}{\sigma_u^2} = \frac{\sigma_p^2 - (\sigma_u^2 + \sigma_{ep}^2)}{\sigma_u^2} = \frac{\sigma_p^2}{\sigma_u^2} - \frac{\sigma_u^2 + \sigma_{ep}^2}{\sigma_u^2} =$$

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

$$\alpha_{ep} = \frac{\sigma_e^2}{\sigma_{ep}^2} = \frac{\sigma_p^2 - \sigma_u^2 - \sigma_{ep}^2}{\sigma_{ep}^2 + \sigma_u^2 - \sigma_u^2} = \frac{\sigma_p^2 - \left(\sigma_u^2 + \sigma_{ep}^2\right)}{\left(\sigma_{ep}^2 + \sigma_u^2\right) - \sigma_u^2} =$$

$$= \frac{\left[\sigma_p^2 - \left(\sigma_u^2 + \sigma_{ep}^2\right)\right] / \sigma_p^2}{\left[\left(\sigma_{ep}^2 + \sigma_u^2\right) - \sigma_u^2\right] / \sigma_p^2} = \frac{1-R}{R-h^2}$$

Por tanto, tampoco es necesario conocer las varianzas de cada efecto sino simplemente los parámetros genéticos heredabilidad y repetibilidad.

Cuando se pretende predecir al rendimiento de un individuo *i* con este modelo se debe tener en cuenta el efecto animal completo que incluye su valor genético y su ambiente permanente, de manera que habrá que sumar $\hat{u}_i + \hat{p}_i$ a la combinación de efectos fijos deseada.

VIII-3.4. El modelo padre-vaca jerarquizada

El modelo padre-vaca jerarquizada merece un breve comentario por su importancia histórica. Es una particularización del modelo padre desarrollado anteriormente en este capítulo en el que se ajusta como efecto adicional el animal propietario del dato, normalmente una vaca al ser utilizado originalmente en el vacuno lechero. Dado que el valor genético del animal propietario del dato no interesa, el efecto animal se incluye aquí prescindiendo de la información de parentesco para reducir considerablemente el coste computacional. En este efecto aleatorio se incluye el ambiente permanente confundido con el valor genético del animal que produce el dato, pudiéndose obtener el primero por diferencia. Su empleo permitió abordar el estudio de la repetibilidad cuando aún existían limitaciones computacionales.

VIII-4. Modelos con efectos maternos

Debe ajustarse este modelo cuando el carácter analizado se mide a edades tempranas. Son caracteres de este estilo los pesos de los individuos jóvenes como el peso al nacimiento pero sobre todo el carácter típico es el peso al destete. La idea que subyace en este modelo es que el rendimiento de los individuos jóvenes aún depende mucho del ambiente materno, es decir el ambiente que le ha proporcionado su madre. Transcurrido suficiente tiempo desde que el individuo es separado de su madre, el peor o mejor ambiente recibido de ella es compensado con el ambiente recibido en un tiempo posterior. Por ejemplo, no tendría sentido ajustar este modelo en un carácter como el peso a la canal.

El carácter que sirve de ejemplo en este modelo es el peso al destete, así que será éste el carácter que se utilizará como ejemplo para formalizar el modelo.

VIII-4.1. Efectos fijos ajustados en el carácter peso al destete

El efecto que agrupa a los animales en el espacio y en el tiempo puede ser en este caso, igual que en el modelo anterior, el efecto combinado Rebaño-Año-Estación que suele abreviarse como RAE y que corrige la influencia que sobre el dato tienen los tres efectos así como su interacción.

También el número de parto de la madre es determinante en el tamaño de la cría. Entre otras razones, en el carácter peso al destete tiene una enorme influencia la producción de leche de la madre, y ésta, tal y como fue descrito anteriormente, no es la misma en todos los partos de la madre. Así, igualmente, el número de parto de la madre se suele emplear como efecto fijo discontinuo con 4 o 5 niveles, uno por número de parto incluyendo el último de ellos los partos posteriores. Igualmente el efecto edad de la madre al parto suele ajustarse como efecto fijo continuo, y también como covariable lineal y cuadrática.

Un efecto fijo discontinuo que no debe dejar de incluirse es el sexo del ternero ya que hay una importante diferencia en peso entre los dos sexos.

Cuando los destetes no se realizan a una edad fija, es evidente que la edad del destete tendrá una influencia sobre su peso. Por ello, debe incluirse también una covariable edad del ternero al destete. Si los distintos periodos de destete no distan mucho entre sí, el crecimiento del ternero en ese período puede ser ajustado linealmente, pero si el período completo es largo, entonces la edad del ternero debe ajustarse al mismo tiempo tanto como covariable lineal como cuadrática.

VIII-4.2. Definición del modelo con efectos maternos

Ecuación del modelo:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Mm} + \mathbf{e}$$

En este caso se ha incluido como nuevo efecto aleatorio, el efecto materno. Desde el punto de vista del ternero se trata de un efecto ambiental, pero desde el punto de vista de la madre es un efecto genético, ya que las hijas de hembras con elevado efecto materno también poseerán un buen carácter materno; basta pensar que gran parte del efecto materno se debe al carácter lechero de la madre. Por tanto suele ajustarse como un efecto genético con tantos niveles como individuos.

El vector que contiene todos los efectos maternos distintos es el vector **m**, y **M** es su matriz de incidencia o matriz diseño que relaciona los datos con su madre. Obsérvese que en este caso u y m tienen el mismo número de niveles pero **Z** y **M**, sus matrices diseño, difieren notablemente.

Esperanzas y Varianzas del modelo:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{A}\sigma_{um} & \mathbf{0}\\ \mathbf{A}\sigma_{um} & \mathbf{A}\sigma_m^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

El efecto materno se clasifica entonces como un efecto aleatorio ya que se le ha incluido en la matriz de varianzas y covarianzas. Al igual que el efecto ambiental permanente definido en el modelo anterior, se ha definido su distribución con media cero ($E(\mathbf{m}) = \mathbf{0}$) y varianza, la varianza ambiental permanente (σ_m^2). Sin embargo en este caso se ha establecido una estructura de covarianzas entre niveles como proporcional a la matriz de relaciones aditivas dado que se trata de un efecto genético. Es más, se ha asumido también

que puede existir una covarianza entre ambos efectos genéticos. De hecho es muy común que esta covarianza sea negativa, es decir, que las hembras con buena capacidad de crecimiento suelen ser luego malas madres.

Aunque ésta es la estructura más habitual asumida para la matriz de varianzas y covarianzas de los efectos maternos, existen otras posibilidades:

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\sigma_m^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix} \quad \text{o}$$

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_m^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

En estos dos casos se ha asumido que no existe covarianza entre los dos efectos genéticos del modelo. Obsérvese que esto no significa que no se vaya a tener en cuenta. Al contrario, se está definiendo que tal covarianza vale cero, lo que podría resultar incorrecto. Esta asunción se suele realizar cuando se posee una elevada incertidumbre sobre el parámetro σ_{um}. Esto sucede cuando este parámetro se ha estimado en una población donde existen pocas madres con datos de su descendencia que dispongan al mismo tiempo de su dato de crecimiento. En tal caso puede ser preferible asumir que esta relación no existe a asumir el valor procedente de la estimación. En el segundo de los dos casos el efecto materno se ha asumido exclusivamente como ambiental al ignorar las relaciones de parentesco entre las madres ($Var(\mathbf{m})=\mathbf{I}\sigma_m^2$).

Una interesante variante de este modelo es el que se ajusta en especies prolíficas que propone diferentes niveles para partos distintos de la misma madre. Es el efecto aleatorio camada o efecto de ambiente común, que permite además equilibrar el tamaño de

camada de las madres fomentando la adopción por parte de las hembras menos prolíficas, ya que en este caso el efecto sólo se define como ambiental. Incluso, puede definirse un modelo que posea un efecto materno y un efecto de ambiente común (que podemos llamar **c** con matriz de incidencia V, con valor esperado nulo y varianza σ_c^2. La ecuación, esperanza y varianzas del modelo serían en este caso las siguientes:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Mm} + \mathbf{Vc} + \mathbf{e}$$

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{m}\\ \mathbf{c}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{c}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{A}\sigma_{um} & \mathbf{0} & \mathbf{0}\\ \mathbf{A}\sigma_{um} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_c^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

Se deja al lector la interpretación de esta definición.

VIII-4.3. Ecuaciones del modelo mixto en efectos maternos

También en este caso las ecuaciones del modelo mixto a resolver tienen una estructura similar a las vistas anteriormente:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'M} & \mathbf{X'Z} \\ \mathbf{M'X} & \mathbf{M'M} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_m^2} & \mathbf{M'Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_{um}} \\ \mathbf{Z'X} & \mathbf{Z'M} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_{um}} & \mathbf{Z'Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'y} \\ \mathbf{Z'y} \end{bmatrix}$$

En este caso a los bloques **M'Z** y **Z'M** hay que añadir información al existir una covarianza definida entre los dos efectos genéticos cuya estructura definen las dos matrices diseño **Z** y **M**.

La traducción de los cocientes de varianza que multiplican a la inversa de la matriz de relaciones aditivas en los cuatro bloques no es tan inmediata por lo que se dejará expresado en su forma actual.

Obsérvese que en este caso disponemos del predictor del valor genético de los individuos tanto para crecimiento ($\hat{\mathbf{u}}$) como para efecto materno ($\hat{\mathbf{m}}$). Ello permitiría seleccionar los animales para cualquiera de los dos objetivos, el de crecimiento o para seleccionar animales como madres, incluso podrían separarse dos líneas en función de los objetivos de selección de la población, o también definir un índice de selección con diferentes pesos económicos que permitan buscar un óptimo económico del ganadero.

VIII-4.4. El modelo con efecto materno y ambiente permanente materno

El modelo con efecto ambiental permanente materno se desarrolla con el objeto de diferenciar la parte del efecto materno que no es de origen genético por lo que su ecuación incluye los dos efectos aleatorios adicionales vistos, el efecto materno y el ambiente permanente, en este caso materno:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Mm} + \mathbf{Wp} + \mathbf{e}$$

La distinción entre la parte del efecto materno que es genética y la parte que no lo es, la hace el modelo sobre la base de la existencia

de medidas repetidas del efecto materno, es decir, aprovechando la existencia de diversos hijos de las mismas madres. Así pues, mientras que **m** contiene tantos niveles como individuos, en **p** hay tantos niveles como madres. En este caso, **p**, el ambiente permanente materno es la parte del efecto materno común a todos los hijos de la misma madre que no tiene componente genético, y es un efecto que no se suele incluir en los modelos con demasiada frecuencia.

La definición del modelo se completaría con las esperanzas y varianzas de los efectos:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{m}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{A}\sigma_{um} & \mathbf{0} & \mathbf{0}\\ \mathbf{A}\sigma_{um} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_{ep}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

La resolución del modelo se llevaría a cabo resolviendo las ecuaciones siguientes:

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'M} & \mathbf{X'W}\\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} & \mathbf{Z'M}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_{um}} & \mathbf{Z'W}\\ \mathbf{M'X} & \mathbf{M'Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_{um}} & \mathbf{M'M}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_m^2} & \mathbf{M'W}\\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'M} & \mathbf{W'W}+\mathbf{I}\frac{\sigma_e^2}{\sigma_{ep}^2}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\\ \hat{\mathbf{m}}\\ \hat{\mathbf{p}}\end{bmatrix}=\begin{bmatrix}\mathbf{X'y}\\ \mathbf{Z'y}\\ \mathbf{M'y}\\ \mathbf{W'y}\end{bmatrix}$$

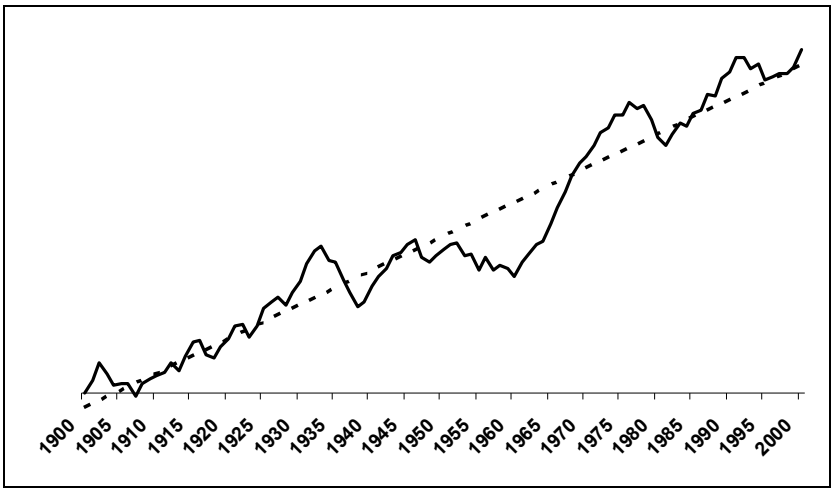
Dos observaciones merecen ser hechas en este momento. La primera, que incrementar el número de efectos aleatorios del modelo no incrementa más su complejidad teórica, aunque sí sus dimensiones. La segunda, que la definición de muchos efectos aleatorios implica la necesidad de conocer más parámetros por lo que se tiende a definir un número bajo de los mismos.

VIII-4.5. El modelo padre-abuelo materno

El modelo padre-abuelo materno también merece un breve comentario por su importancia histórica. Es otra particularización del modelo padre desarrollado anteriormente en este capítulo en el que se ajusta como efecto adicional el abuelo materno del animal propietario del dato. En este efecto aleatorio se incluye la cuarta parte del efecto genético aditivo y la mitad del efecto genético materno. Su empleo permitió abordar el estudio de componentes maternos cuando aún no se disponía de las capacidades computacionales que exigía el modelo completo. Obsérvese que en este modelo el número de ecuaciones para los efectos genético aditivo y materno se corresponde con el número de machos existentes, un número habitualmente muy inferior al número total de animales.

VIII-5. Modelos con grupos genéticos

Una de las definiciones utilizadas en el modelo es $E(\mathbf{u}) = 0$. Con ello forzamos al modelo a que todos los individuos sin padres conocidos pertenezcan a una misma distribución de media cero. Esta asunción tiene coherencia en poblaciones cerradas no sometidas a selección, pero puede no tener tanta cuando no es el caso. Imaginemos una población sometida a selección desde el año 1900 y disponemos de su información desde el año 2000. Se produce entonces una tendencia más o menos irregular en los valores genéticos de forma que el valor genético de un animal nacido en el año 2000 no es comparable con el de un animal nacido en el año 1900 según se deduce de la siguiente figura:



Sin embargo, tal y como tenemos definido el modelo, el valor genético esperado de cualquier individuo sin padres conocidos será el mismo, cero, independientemente de su año de nacimiento. La misma situación se da en los casos de inmigración de animales de otra población. Por ejemplo, si un ganadero compra un semental de una población con supuesto mayor valor genético que la propia, el animal aparece en el registro genealógico como fundador y se le asigna el mismo valor genético esperado que cualquier otro individuo fundador de la propia población.

Para corregir este efecto se define un nuevo efecto fijo conocido como grupo genético. Con este modelo ningún individuo en la población aparece como fundador. Aquellos padres desconocidos son reemplazados con animales llamados padres fantasmas que poseen ese individuo como único hijo, y se asigna a cada uno de ellos un grupo genético diferente, de manera que ahora todos los fundadores son animales fantasmas cuyo valor esperado se corresponde con la media del grupo genético al que se le asigna. Así para cada individuo fundador, $E(u_i) = g_j$, siendo g_j la media del grupo genético al que se le asignado.

La manera de agrupar los animales fantasmas para definir los grupos genéticos no sigue una norma fija y exige un análisis preliminar de los datos. Deben vigilarse dos cuestiones:

- La definición de pocos grupos genéticos corrige débilmente el problema. En el extremo, la definición de un único grupo genético equivale a no ajustar el efecto. Por tanto, en la figura presentada como ejemplo, podría definirse un grupo para cada 50 años, lo que supondría dos grupos, cada década, cada año, incluso cada mes. Lo último llevaría a una elevada corrección del efecto.
- La definición de muchos grupos genéticos podría conllevar grupos con pocos animales. Dado que se trata de un efecto fijo, cada nivel ha de contar con un suficiente número de datos que nos permita una buena estimación.

Por tanto, no existen normas fijas. No es necesario tampoco que se hagan grupos equilibrados en el tiempo. Por ejemplo, podría decidirse que todos los padres de animales nacidos antes de 1950 pertenecen a un mismo grupo genético, mientras que entre 1950 y 1980 se define un nivel cada 5 años, un grupo distinto para los nacidos a partir de 1980 y de 1995 al 2000 dos grupos cada año separando por ejemplo los padres fantasmas de los machos y de las hembras. Incluso en el último año podría haber un grupo genético adicional para los padres fantasmas de animales importados. Esta flexibilidad en la definición de los grupos genéticos es igualmente su mayor crítica ya que diferentes definiciones de grupos genéticos podrían conducir a diferentes soluciones en los valores genéticos.

En la definición del modelo, la formulación de la ecuación se complica ligeramente ya que se define un efecto fijo en individuos sin datos, apareciendo en la ecuación de cada individuo una fracción de varios niveles del mismo efecto. Se escribe a continuación la ecuación del modelo habitual pero en álgebra de escalares. Dado que la parte considerada hasta ahora fija en el modelo no se verá afectada, en la ecuación se escribe *FIJOS* para representar lo que en álgebra de matrices se ha llamado **Xb**:

$$y_i = FIJOS + u_i + e_i$$

El modelo con grupos genéticos en álgebra de escalares se representaría así:

$$y_i = FIJOS + u_i^* + \sum_{j=1}^{f} t_{ij} g_j + e_i$$

En esta expresión *f* es el número de padres fantasmas, u_i^* es la parte del valor genético que excede a la aportada por los fundadores fantasmas y t_{ij} el parentesco del individuo con el fundador fantasma. Obsérvese que finalmente el coeficiente que multiplica al grupo g_j se corresponde con la suma de los parentescos del individuo con todos los fundadores fantasmas asignados al grupo *j*.

Se va a desarrollar la resolución el modelo de grupos genéticos utilizando el ejemplo previo para resolver el BLUP. En este caso, se han identificado los padres desconocidos como fundadores fantasmas numerados de F1 a F5, asignando los dos primeros a un grupo y los otros tres a otro grupo genético diferente:

Animal	Padre	Madre
1	F1	F2
2	F3	F4
3	F5	2
4	1	2
5	3	2
6	3	4

Fantasma	Grupo
F1	1
F2	1
F3	2
F4	2
F5	2

Obsérvese que los elementos t_{ij} se corresponden con los coeficientes de la matriz **T** definida como $\mathbf{T} = \left(\mathbf{I} - \tfrac{1}{2}\mathbf{P}\right)^{-1}$ en el desarrollo de la inversa de **A**. En este caso la matriz de **P** de padres incluye también a los fundadores fantasmas:

$$\mathbf{P} = \left[\begin{array}{cccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{array}\right]$$

$$\mathbf{T} = \left[\begin{array}{cccccc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{8} & \frac{3}{8} & \frac{1}{4} & 0 & \frac{3}{4} & \frac{1}{2} & 0 & 1 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 1 & 0 \end{array}\right]$$

Obsérvese que el bloque inferior izquierdo de **T** presenta el porcentaje de genes de cada individuo procedente de cada fundador fantasma.

Se va a definir **Q** como la matriz con tantas filas como individuos en el pedigrí (incluyendo los fundadores fantasmas) y tantas columnas como grupos genéticos, que representa el porcentaje de genes que posee cada individuo procedente de cada grupo genético (como suma de los genes que recibe de los fundadores fantasmas asignados a cada grupo genético). Para ello es preciso definir previamente también **Q*** con las mismas dimensiones que **Q**, como la matriz que define qué individuos del pedigrí son

fundadores fantasma y a qué grupo se les asigna anotando un 1 en la posición correspondiente y manteniendo nulos el resto de los elementos. Así definidas, **Q** se obtiene mediante **Q** = **TQ***:

$$\mathbf{Q}^* = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right] \Rightarrow$$

$$\mathbf{Q} = \mathbf{TQ}^* = \left[\begin{array}{ccccc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{8} & \frac{3}{8} & \frac{1}{4} & 0 & \frac{3}{4} & \frac{1}{2} & 0 & 1 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 1 \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ \hline 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \\ \frac{1}{4} & \frac{3}{4} \end{array}\right]$$

Se define ahora el vector de incógnitas grupos genéticos **g** como $\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$, se incrementa el vector de valores genéticos aditivos con los fundadores fantasmas y se aumenta la matriz de incidencia de los efectos aleatorios **Z** en cinco columnas de ceros para incorporar los fundadores fantasmas. Además se separa del valor genético del individuo la parte que corresponde a los grupos **u** = **u*** + **Qg**, de manera que la parte aleatoria del modelo se puede desglosar **Zu** = **Z**(**u*** + **Qg**) = **Zu*** + **ZQg**. Puede observarse que

la matriz de incidencia de los grupos genéticos se obtendrá mediante el producto de dos matrices (**Qg**), siendo en este caso una matriz de incidencia compuesta por fracciones y no por ceros o unos, ya que en determinado individuo puede haber un porcentaje de genes procedente de un grupo y otro porcentaje de otro. Esta matriz se muestra más abajo. La nueva ecuación del modelo queda de la siguiente manera:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu}^{*} + \mathbf{ZQg} + \mathbf{e}$$

Y las diversas matrices introducidas son de la siguiente manera:

$$\mathbf{Zu} = \left[\begin{array}{ccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} u_{F1} \\ u_{F2} \\ u_{F3} \\ u_{F4} \\ u_{F5} \\ \hline u_{1} \\ u_{2} \\ u_{3} \\ u_{4} \\ u_{5} \\ u_{6} \end{array}\right] =$$

$$\mathbf{Zu}^{*} + \mathbf{ZQg} = \left[\begin{array}{ccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} u^{*}_{F1} \\ u^{*}_{F2} \\ u^{*}_{F3} \\ u^{*}_{F4} \\ u^{*}_{F5} \\ \hline u^{*}_{1} \\ u^{*}_{2} \\ u^{*}_{3} \\ u^{*}_{4} \\ u^{*}_{5} \\ u^{*}_{6} \end{array}\right] + \left[\begin{array}{cc} 0 & 1 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \\ \frac{1}{4} & \frac{3}{4} \end{array}\right] \left[\begin{array}{c} g_{1} \\ g_{2} \end{array}\right]$$

Con esta definición de la ecuación del modelo, y recordando que el grupo genético es un efecto fijo, las ecuaciones del modelo mixto a resolver son las siguientes:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'ZQ} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} & \mathbf{Z'ZQ} \\ \mathbf{Q'Z'X} & \mathbf{Q'Z'Z} & \mathbf{Q'Z'ZQ} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}^* \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{Q'Z'y} \end{bmatrix}$$

La inversa de **A** se construye como siempre pero para los once individuos, los seis que había originalmente en el pedigrí y los cinco fundadores fantasmas con las reglas de Henderson, dando lugar a:

$$\mathbf{A}^{-1} = \left[\begin{array}{cccccc|cccccc} 3/2 & 1/2 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 3/2 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/2 & 1/2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 3/2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3/2 & 0 & 1/2 & -1 & 0 & 0 & 0 \\ \hline -1 & -1 & 0 & 0 & 0 & 5/2 & 1/2 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 1/2 & 1/2 & 7/2 & -1/2 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1/2 & 3 & 1/2 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1/2 & 5/2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 2 \end{array}\right]$$

Finalmente, usando la misma heredabilidad que en el ejemplo (h^2 = 0,60; $\alpha = 2/3$) se construyen las ecuaciones del modelo mixto presentadas:

$$\left[\begin{array}{cccccccc|cccccc|cc|c}
5 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 3/4 & 17/4 & \hat{\mu} & 2070 \\
2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 2 & \hat{b}_1 & 845 \\
3 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 3/4 & 9/4 & \hat{b}_2 & 1225 \\
\hline
0 & 0 & 0 & 1 & 1/3 & 0 & 0 & 0 & -2/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{u}^*_{F1} & 0 \\
0 & 0 & 0 & 1/3 & 1 & 0 & 0 & 0 & -2/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{u}^*_{F2} & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1/3 & 0 & 0 & -2/3 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{u}^*_{F3} & 0 \\
0 & 0 & 0 & 0 & 0 & 1/3 & 1 & 0 & 0 & -2/3 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{u}^*_{F4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/3 & -2/3 & 0 & 0 & 0 & 0 & 0 & \hat{u}^*_{F5} & 0 \\
\hline
0 & 0 & 0 & -2/3 & -2/3 & 0 & 0 & 0 & 5/3 & 1/3 & 0 & -2/3 & 0 & 0 & 0 & 0 & \hat{u}^*_1 & 0 \\
1 & 1 & 0 & 0 & 0 & -2/3 & -2/3 & 1/3 & 1/3 & 10/3 & -1/3 & -2/3 & -2/3 & 0 & 0 & 1 & \hat{u}^*_2 & 415 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & -2/3 & 0 & -1/3 & 3 & 1/3 & -2/3 & -2/3 & 0 & 1 & \hat{u}^*_3 & 430 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -2/3 & -2/3 & 1/3 & 8/3 & 0 & -2/3 & 1/2 & 1/2 & \hat{u}^*_4 & 420 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -2/3 & -2/3 & 0 & 7/3 & 0 & 0 & 1 & \hat{u}^*_5 & 400 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2/3 & -2/3 & 0 & 7/3 & 1/4 & 3/4 & \hat{u}^*_6 & 405 \\
\hline
3/4 & 0 & 3/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/4 & 5/16 & 7/16 & \hat{g}_1 & 311,25 \\
17/4 & 2 & 9/4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1/2 & 1 & 3/4 & 7/16 & 61/16 & \hat{g}_2 & 1758,75
\end{array}\right]$$

Es preciso recordar que el grupo genético es un efecto fijo discontinuo, y por lo tanto, la suma de sus dos ecuaciones vuelve a ser igual que la de la media. Así pues, la inversa generalizada a obtener debe eliminar también una de estas dos ecuaciones. Resolveremos aquí anulando la ecuación de la media y la del grupo genético 2. Las soluciones que se obtienen son:

$$\begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 0,0000 \\ 422,5000 \\ 397,6792 \end{bmatrix} \begin{bmatrix} \hat{u}^*_{F1} \\ \hat{u}^*_{F2} \\ \hat{u}^*_{F3} \\ \hat{u}^*_{F4} \\ \hat{u}^*_{F5} \end{bmatrix} = \begin{bmatrix} 0,0000 \\ 0,0000 \\ -1,4651 \\ -1,4651 \\ 2,9301 \end{bmatrix} \begin{bmatrix} \hat{u}^*_1 \\ \hat{u}^*_2 \\ \hat{u}^*_3 \\ \hat{u}^*_4 \\ \hat{u}^*_5 \\ \hat{u}^*_6 \end{bmatrix} = \begin{bmatrix} 0,0000 \\ -2,9301 \\ 2,9301 \\ -1,4651 \\ 0,9946 \\ -1,2567 \end{bmatrix}$$

$$\begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} 44,9194 \\ 0,0000 \end{bmatrix}$$

Dos cuestiones merecen ser destacadas:

- La solución de los grupos genéticos no es función estimable pero sí su diferencia. Sabemos entonces que la media de los animales fundadores asignados al grupo genético 1 son superiores a los del grupo genético 2 en casi 45 kilogramos.
- Las soluciones proporcionadas para los valores genéticos de los individuos, tanto para los fundadores fantasmas como para el resto, no contienen todo el valor genético del individuo, sino que hay que sumarle la parte del efecto fijo correspondiente: $\hat{\mathbf{u}}^0 = \hat{\mathbf{u}}^* + \mathbf{Q}\hat{\mathbf{g}}$. Las soluciones definitivas son:

$$\begin{bmatrix} \hat{u}^0_{F1} \\ \hat{u}^0_{F2} \\ \hat{u}^0_{F3} \\ \hat{u}^0_{F4} \\ \hat{u}^0_{F5} \\ \hline \hat{u}^0_{1} \\ \hat{u}^0_{2} \\ \hat{u}^0_{3} \\ \hat{u}^0_{4} \\ \hat{u}^0_{5} \\ \hat{u}^0_{6} \end{bmatrix} = \begin{bmatrix} 0,0000 \\ 0,0000 \\ -1,4651 \\ -1,4651 \\ 2,9301 \\ \hline 0,0000 \\ -2,9301 \\ -2,9301 \\ -2,9301 \\ 0,9946 \\ -1,2567 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ ½ & ½ \\ 0 & 1 \\ ¼ & ¾ \end{bmatrix} \begin{bmatrix} 44,9194 \\ 0,0000 \end{bmatrix} = \begin{bmatrix} 44,9194 \\ 44,9194 \\ -1,4651 \\ -1,4651 \\ 2,9301 \\ \hline 44,9194 \\ -2,9301 \\ -2,9301 \\ 20,9946 \\ 0,9946 \\ 9,9731 \end{bmatrix}$$

Obsérvese que se ha utilizado la notación $\hat{\mathbf{u}}^0$ porque los valores genéticos definitivos no son soluciones únicas. Las soluciones en $\hat{\mathbf{u}}^*$ sí son únicas pero desafortunadamente no representan el valor genético total de los individuos, sino su superioridad o inferioridad con respecto al valor genético medio de los animales de su grupo. Las soluciones de los grupos genéticos han sido forzadas a ser referidas a uno de los grupos, en este caso el grupo 2. Esta propiedad puede ser empleada para decidir cual será la base con la que se comparan los individuos valorados forzando ese grupo a valer cero.

Se pueden manipular las ecuaciones para encontrar otras equivalentes que proporcionen las soluciones de los valores genéticos con las medias de sus grupos ya incorporadas y sin necesidad de incluir los fundadores fantasmas cuyo valor no tiene ningún interés. En las nuevas ecuaciones **Z** vuelve a tener tantas columnas como individuos en el pedigrí desapareciendo las correspondientes a los padres fantasmas. Para ello se absorben las ecuaciones de estos fundadores fantasmas en el resto y se obtienen las siguientes:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{0} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}_{nn}\frac{\sigma_e^2}{\sigma_u^2} & \mathbf{A}^{-1}_{np}\frac{\sigma_e^2}{\sigma_u^2} \\ \mathbf{0} & \mathbf{A}^{-1}_{pn}\frac{\sigma_e^2}{\sigma_u^2} & \mathbf{A}^{-1}_{pp}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}^* + \mathbf{Q}\hat{\mathbf{g}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{0} \end{bmatrix}$$

En estas ecuaciones la matriz $\mathbf{A}^{-1}$ no corresponde a la inversa de la matriz de relaciones aditivas de todos los individuos sino que los valores correspondientes a animales del mismo grupo genético son sumados en una única incógnita. Podemos entonces particionar la inversa de **A** separando los individuos de los grupos:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{-1}_{nn} & \mathbf{A}^{-1}_{ng} \\ \mathbf{A}^{-1}_{gn} & \mathbf{A}^{-1}_{gg} \end{bmatrix}$$

El nuevo pedigrí, en el que los padres fantasmas son reemplazados por el grupo genético al que pertenecen es el siguiente:

Animal	Padre	Madre
1	7	7
2	8	8
3	8	2
4	1	2
5	3	2
6	3	4

Y la inversa de **A** se construye con reglas muy parecidas a las utilizadas en condiciones normales, y que pueden extraerse del siguiente cuadro:

	i	**j**	**k**
i	1	–½	–½
j	–½	¼	¼
k	–½	¼	¼

$$\text{x } b_i \text{ siendo } b_i = \frac{4}{2 + n^{o}\ padres\ desconocidos}$$

En el ejemplo, la inversa de **A** con sus cuatro bloques queda de la siguiente manera:

$$\mathbf{A^{-1}} = \left[\begin{array}{ccccc|cc} 1+½ & ½ & & -1 & & -1 & \\ ½ & 1+½+⅓+⅓ & -⅔+½ & -1 & -1 & & -1 \\ & -⅔+½ & ⅓+½+½ & ½ & -1 & -1 & -⅔ \\ -1 & -1 & ½ & 2+½ & & -1 & \\ & -1 & -1 & & 2 & & \\ & & -1 & -1 & & 2 & \\ \hline -1 & -1 & & & & ¼+¼+¼+¼ & \\ & -1 & -⅔ & & & & ¼+¼+¼+¼+⅓ \end{array}\right] =$$

$$\mathbf{A^{-1}} = \left[\begin{array}{cccccc|cc} 3/2 & ½ & 0 & -1 & 0 & 0 & -1 & 0 \\ ½ & 7/3 & -⅙ & -1 & -1 & 0 & 0 & -1 \\ 0 & -⅙ & 7/3 & ½ & -1 & -1 & 0 & -⅔ \\ -1 & -1 & ½ & 5/2 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 & 0 & 0 \\ \hline -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & -⅔ & 0 & 0 & 0 & 0 & 4/3 \end{array}\right]$$

Y las nuevas ecuaciones del modelo mixto son las siguientes:

$$\left[\begin{array}{ccc|c|ccccc|cc}
5 & 2 & 3 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 3 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
\hline
0 & 0 & 0 & 1 & 1/3 & 0 & -2/3 & 0 & 0 & -2/3 & 0 \\
1 & 1 & 0 & 1/3 & 23/9 & -1/9 & -2/3 & -2/3 & 0 & 0 & -2/3 \\
1 & 1 & 0 & 0 & -1/9 & 23/9 & 1/3 & -2/3 & -2/3 & 0 & -4/9 \\
1 & 0 & 1 & -2/3 & -2/3 & 1/3 & 8/3 & 0 & -2/3 & 0 & 0 \\
1 & 0 & 1 & 0 & -2/3 & -2/3 & 0 & 7/3 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & -2/3 & -2/3 & 0 & 7/3 & 0 & 0 \\
\hline
0 & 0 & 0 & -2/3 & 0 & 0 & 0 & 0 & 0 & 2/3 & 0 \\
0 & 0 & 0 & 0 & -2/3 & -4/9 & 0 & 0 & 0 & 0 & 8/9
\end{array}\right]
\left[\begin{array}{c}
\hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hline \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hline \hat{g}_1 \\ \hat{g}_2
\end{array}\right]
=
\left[\begin{array}{r}
2070 \\ 845 \\ 1225 \\ \hline 0 \\ 415 \\ 430 \\ 420 \\ 400 \\ 405 \\ \hline 0 \\ 0
\end{array}\right]$$

Aunque no se aprecia a simple vista, las ecuaciones para los grupos genéticos pueden obtenerse como combinación lineal de las anteriores y debe anularse uno de los grupos para obtener la solución, por ejemplo el segundo. Las soluciones que se obtienen son exactamente las presentadas anteriormente con la excepción de no proporcionar valores para los padres fantasmas. Es importante recordar que los valores genéticos siguen sin ser soluciones únicas y representan la superioridad genética de cada animal con respecto a la media de los animales del grupo genético tomado como referencia.

VIII-6. Modelos multicarácter

Los modelos multicarácter permiten evaluar genéticamente todos los individuos para más de un carácter de forma simultánea con el objetivo de aprovechar la información que aporta cada carácter para evaluar los otros con los que tiene correlación genética.

Si consideramos dos caracteres conjuntamente la ecuación del modelo sería la siguiente:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \end{bmatrix} + \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix}$$

En esta ecuación los subíndices representan el carácter. El resto de los elementos son los mismos que los definidos en modelos unicarácter. Las matrices de varianzas y covarianzas de los vectores de efectos aleatorios del modelo son también similares al modelo unicarácter.

PRODUCTO DE KRONECKER

Si A es una matriz de dimensiones (m x n) y B es una matriz de dimensiones (p x q), entonces el producto de Kronecker $A \otimes B$ es la matriz bloque de dimensiones (mp x nq): $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}b_{11} & .. & \mathbf{A}b_{n1} \\ : & : & : \\ \mathbf{A}b_{1n} & .. & \mathbf{A}b_{nn} \end{bmatrix}$.

Se define a continuación en primer lugar las matrices $\mathbf{G_0}$ y $\mathbf{R_0}$ de la siguiente manera:

$$\mathbf{G_0} = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_1u_2} \\ \sigma_{u_1u_2} & \sigma^2_{u_2} \end{bmatrix} \qquad \mathbf{R_0} = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_1e_2} \\ \sigma_{e_1e_2} & \sigma^2_{e_2} \end{bmatrix}$$

En estas matrices todos los elementos han sido ya definidos en el modelo unicarácter. La única novedad es que los subíndices diferencian caracteres. Con ayuda de estas matrices de definen las matrices de varianzas y covarianzas de los efectos aleatorios:

$$\mathbf{G} = \mathbf{A} \otimes \mathbf{G_0} = \begin{bmatrix} \mathbf{A}\sigma^2_{u_1} & \mathbf{A}\sigma_{u_1u_2} \\ \mathbf{A}\sigma_{u_1u_2} & \mathbf{A}\sigma^2_{u_2} \end{bmatrix}$$

$$\mathbf{R} = \mathbf{I} \otimes \mathbf{R_0} = \begin{bmatrix} \mathbf{I}\sigma^2_{e_1} & \mathbf{I}\sigma_{e_1e_2} \\ \mathbf{I}\sigma_{e_1e_2} & \mathbf{I}\sigma^2_{e_2} \end{bmatrix}$$

Se definen los elementos de las inversas de $\mathbf{G_0}$ y $\mathbf{R_0}$ de la siguiente manera:

$$\mathbf{G_0^{-1}} = \begin{bmatrix} g^{11} & g^{12} \\ g^{12} & g^{22} \end{bmatrix} \qquad \mathbf{R_0^{-1}} = \begin{bmatrix} r^{11} & r^{12} \\ r^{12} & r^{22} \end{bmatrix}$$

Se exponen a continuación las ecuaciones del modelo mixto en el caso en que ambos caracteres presenten el mismo modelo con todos los caracteres medidos en todos los individuos que posean dato ($\mathbf{X_1} = \mathbf{X_2} = \mathbf{X}$ y $\mathbf{Z_1} = \mathbf{Z_2} = \mathbf{Z}$):

$$\left[\begin{array}{cc|cc} \mathbf{X'X}r^{11} & \mathbf{X'Z}r^{11} & \mathbf{X'X}r^{12} & \mathbf{X'Z}r^{12} \\ \mathbf{Z'X}r^{11} & \mathbf{Z'Z}r^{11} + \mathbf{A^{-1}}g^{11} & \mathbf{Z'X}r^{12} & \mathbf{Z'Z}r^{12} + \mathbf{A^{-1}}g^{12} \\ \hline \mathbf{X'X}r^{12} & \mathbf{X'Z}r^{12} & \mathbf{X'X}r^{22} & \mathbf{X'Z}r^{22} \\ \mathbf{Z'X}r^{12} & \mathbf{Z'Z}r^{12} + \mathbf{A^{-1}}g^{12} & \mathbf{Z'X}r^{22} & \mathbf{Z'Z}r^{22} + \mathbf{A^{-1}}g^{12} \end{array}\right] \left[\begin{array}{c} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{u}}_1 \\ \hline \hat{\mathbf{b}}_2 \\ \hat{\mathbf{u}}_2 \end{array}\right] = \left[\begin{array}{c} \sum_{b_i} \mathbf{y}_{1i} r^{11} \mathbf{y}_{2i} r^{12} \\ \sum_{u_i} \mathbf{y}_{1i} r^{11} \mathbf{y}_{2i} r^{12} \\ \hline \sum_{b_i} \mathbf{y}_{1i} r^{11} \mathbf{y}_{2i} r^{12} \\ \sum_{u_i} \mathbf{y}_{1i} r^{11} \mathbf{y}_{2i} r^{12} \end{array}\right]$$

En el vector del lado derecho se ha escrito por ejemplo $\sum_{b_i} \mathbf{y}_{1i} r^{11} \mathbf{y}_{2i} r^{12}$ para expresar la suma de los datos de cada carácter dentro de cada nivel del efecto fijo multiplicado las sumas para el primer carácter por el elemento r^{11} y las sumas para el segundo carácter por el elemento r^{22}.

La gran ventaja de los modelos multicarácter es el aumento de la precisión de la valoración genética para cada uno de los caracteres por la incorporación de los demás, especialmente si las correlaciones genéticas entre los caracteres son elevadas. Sin embargo presentan dos inconvenientes importantes:

- Las dimensiones de las ecuaciones crecen de forma cuadrática. Obsérvese cómo la matriz de coeficientes presenta 4 bloques equivalentes al único que existía en el modelo unicarácter. Es fácil razonar que de haber habido 3 caracteres el número de bloques habría sido de 9.

- Es preciso conocer muchos más parámetros. Si en un modelo unicarácter con el mismo número de efectos

aleatorios habría bastado con una varianza por efecto y carácter, en el modelo multicarácter el número de parámetros que es necesario conocer crece también de forma cuadrática con el número de caracteres.

En cualquier caso, dado que la selección sobre muchos caracteres de forma simultánea es poco aconsejable por la pérdida de respuesta en cada uno de ellos, debe procurarse seleccionar los menos posibles para las valoraciones genéticas multicarácter.

CONCEPTOS CLAVE

- ¿En qué tipo de caracteres se ajustan efectos maternos?
- ¿Es el efecto materno un efecto genético o ambiental?
- ¿Cuándo no puede ajustarse el efecto ambiental permanente?
- ¿Cómo se tienen en cuenta las interacciones entre efectos fijos?
- ¿Qué efectos fijos deberían ajustarse en un carácter como el peso a la canal en cerdos que se sacrifican cuando alcanzan un peso fijo?
- ¿Cómo podría ajustarse un efecto fijo continuo que no tiene una relación estrictamente lineal con la variable para la que se lleva a cabo la valoración genética?
- ¿Son el efecto ambiental permanente, el efecto materno y el grupo genético efectos fijos o aleatorios?
- ¿Qué estructura tiene la matriz de varianzas y covarianzas de los efectos maternos? ¿Siempre?
- ¿Cuándo debe ajustarse el efecto grupo genético?
- ¿Es mejor definir muchos o pocos niveles del efecto grupo genético?
- ¿Cuál es la utilidad de los modelos multicarácter?
- ¿Qué inconveniente tienen los modelos multicarácter?
- ¿En qué contexto es de utilidad el modelo padre?

NOVENA PARTE

EJERCICIOS

EJERCICIOS

1- Construya la matriz de relaciones aditivas y su inversa según el siguiente pedigree:

Individuo	Padre	Madre
1	**0**	**0**
2	**1**	**0**
3	**1**	**2**

Comente si hay algún individuo consanguíneo, y dar el valor de su consanguinidad si lo hay así como la explicación de su consanguinidad.

Construimos **A** siguiendo el método tabular:

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{4} \\ \frac{1}{2} & 1 & \frac{3}{4} \\ \frac{3}{4} & \frac{3}{4} & \frac{5}{4} \end{bmatrix}$$

Y su inversa siguiendo las reglas de Henderson:

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{11}{6} & -\frac{1}{6} & -1 \\ -\frac{1}{6} & \frac{11}{6} & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

Mirando las diagonales de la matriz A, la única que excede de uno es la del animal 3; por tanto sólo éste es consanguíneo y su valor es $F_3 = 5/4 - 1 = ¼$. La explicación de su consanguinidad es que el individuo 1, además de su padre es su abuelo materno.

2- En cuál de los dos siguientes casos se obtiene mayor precisión:

- **Un índice de selección individual con un solo dato en un carácter de heredabilidad 0,25.**
- **Un índice de selección en el que el dato a utilizar como criterio de selección pertenece a uno de los padres, en un carácter con heredabilidad 0,40.**

Deducir el valor de la precisión en cada uno de los casos para razonar la respuesta.

Hay que calcular la precisión mediante la expresión:

$$\rho = \sqrt{\frac{\mathbf{C'V^{-1}C}}{\mathbf{G}}}$$

- **En el caso de un índice de selección individual con heredabilidad 0,25:**

  $C' = C = Cov(u_i, y_i) = Cov(u_i, u_i + e_i) = Cov(u_i, u_i) = \sigma_u^2$

  $V = Var(y_i) = \sigma_p^2$

  $G = Var(u_i) = \sigma_u^2$

  $\varrho_1 = h = 0{,}50$

- **En el caso de un índice de selección sobre un ascendiente con heredabilidad 0,40:**

  $C' = C = Cov(u_i, y_j) = Cov(u_i, u_j + e_j) =^{*}$

  $= Cov(\frac{1}{2} u_j + \frac{1}{2} u_k + \Phi_i, u_j) = \frac{1}{2} \sigma_u^2$

  * (se ha considerado a i como hijo de j y k, $Cov(u_i, e_j) = 0$ por definición del modelo y $Cov(u_j, u_k) = 0$ por no existir relación de parentesco entre j y k)

  $V = Var(y_j) = \sigma_p^2$

  $G = Var(u_i) = \sigma_u^2$

  $\varrho_2 = \frac{1}{2} h = 0{,}32$

$\varrho_1 > \varrho_2$, y por lo tanto el primer índice sería preferible.

3- Construir las ecuaciones del modelo mixto necesarias para resolver el BLUP (no se pide resolverlo) a partir de los siguientes datos:

Individuo	Padre	Madre	Rebaño	Peso
1	-	-	-	-
2	1	-	1	100
3	1	2	1	150
4	-	2	2	120
5	3	4	2	140

Los efectos fijos son la media general y el rebaño.
Heredabilidad = 0.5

Se trata de construir las siguientes ecuaciones:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Las matrices de coeficientes, a excepción de A-1α se hace

	μ	R_1	R_2	u_1	u_2	u_3	u_4	u_5		Lado derecho
M	**4**	**2**	**2**	**0**	**1**	**1**	**1**	**1**		**510**
R_1	**2**	**2**	**0**	**0**	**1**	**1**	**0**	**0**		**250**
R_2	**2**	**0**	**2**	**0**	**0**	**0**	**1**	**1**		**260**
U_1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**		**0**
U_2	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**		**100**
U_3	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**0**		**150**
U_4	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**		**120**
U_5	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**		**140**

contando o sumando datos:

El parámetro alfa lo obtengo así:

$$\alpha = \frac{1-h^2}{h^2} = \frac{0.5}{0.5} = 1$$

Y $\mathbf{A}^{-1}$ se obtiene con las reglas de Henderson a partir del pedigree:

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{11}{6} & -\frac{1}{6} & -1 & 0 & 0 \\ -\frac{1}{6} & \frac{13}{6} & -1 & -\frac{2}{3} & 0 \\ -1 & -1 & \frac{5}{2} & \frac{1}{2} & -1 \\ 0 & -\frac{2}{3} & \frac{1}{2} & \frac{11}{6} & -1 \\ 0 & 0 & -1 & -1 & 2 \end{bmatrix}$$

Una vez calculado todo sólo hay que integrarlo todo:

$$\begin{bmatrix} 4 & 2 & 2 & 0 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & \frac{11}{6} & -\frac{1}{6} & -1 & 0 & 0 \\ 1 & 1 & 0 & -\frac{1}{6} & \frac{19}{6} & -1 & -\frac{2}{3} & 0 \\ 1 & 1 & 0 & -1 & -1 & \frac{7}{2} & \frac{1}{2} & -1 \\ 1 & 0 & 1 & 0 & -\frac{2}{3} & \frac{1}{2} & \frac{17}{6} & -1 \\ 1 & 0 & 1 & 0 & 0 & -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{R}_1 \\ \hat{R}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \end{bmatrix} = \begin{bmatrix} 510 \\ 250 \\ 260 \\ 0 \\ 100 \\ 150 \\ 120 \\ 140 \end{bmatrix}$$

4- Se desea evaluar genéticamente a un individuo para un determinado carácter mediante la utilización de un índice de selección. Se proponen los dos siguientes índices:

a) **$\hat{u}_i$ = 0,30 y_i. La covarianza entre el valor genético del individuo y la fuente de información es σ_u^2.**

b) **$\hat{u}_i$ = 0,50 y_i. La covarianza entre el valor genético del individuo y la fuente de información es ½ σ_u^2.**

¿Cuál de los dos índices escogerías? ¿Por qué?

Se escogería el índice que me diera una mejor valoración y ésta medida me la proporciona la precisión. El cuadrado de la precisión se obtiene mediante la expresión:

$$\rho^2 = \frac{C'V^{-1}C}{G}$$

siendo

C = Covar (**y**, **u'**); **C'** = Covar (**u**, **y'**); **G** = Var(**u**); **V** = Var(**y**); $\mathbf{V^{-1}}$ = Inversa de **V**

Puesto que el individuo a evaluar es el mismo y los índices son diferentes, las fuentes de información son distintas y, por tanto, también el valor de V será diferente en cada caso. Sin embargo, G tiene el mismo valor en ambos casos:

$$\mathbf{G} = \text{Var}(\mathbf{u}) = \text{Var}(u_i) = \sigma_u^2$$

En el caso de una única fuente de información y un único valor genético a predecir, las matrices se convierten en escalares y, por tanto, C = C'.

Como el índice de selección se obtiene mediante la expresión $\hat{u}_i = C'V^{-1}y_i$, se puede conocer el valor de $C'V^{-1}$ en cada caso y en consecuencia el valor de la precisión:

a) $C'V^{-1} = 0,30$ y por tanto:

$$\rho^2 = \frac{0,30\sigma_u^2}{G} = \frac{30\sigma_u^2}{\sigma_u^2} = 0,30$$

b) $C'V^{-1} = 0,50$ y en este caso: $\rho^2 = \frac{0,50C}{G} = \frac{0,50\frac{1}{2}\sigma_u^2}{G} = \frac{0,25\sigma_u^2}{\sigma_u^2} = 0,25$

Por tanto es preferible el primer índice.

5- Construir, sin resolver, las ecuaciones del modelo mixto necesarias para evaluar dos individuos para un carácter mediante metodología BLUP con producciones 100 Kg y 200 Kg. No se conocen ascendientes de ninguno de ellos y la heredabilidad del carácter es $h^2 = 0.50$.

- **Realizar la valoración genética también mediante un índice de selección.**
- **¿Qué tipo de índice de selección sería?**

¿Qué diferencias hay entre los dos métodos?

Se trataría de construir las siguientes ecuaciones del modelo lineal mixto con la media general como único efecto fijo y los valores genéticos de dos animales como único efecto aleatorio:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Parte de la matriz de coeficientes $\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}$, y el vector del lado derecho (RHS) $\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$, se rellenan haciendo una tabla de doble entrada con todas las incógnitas y contando el número de observaciones de los cruces (matriz de coeficientes), y sumando observaciones (vector del lado derecho):

	Media(μ)	Animal 1 (u_1)	Animal 2 (u_2)		RHS
Media (μ)	**2**	**1**	**1**		**300**
Animal 1 (u_1)	**1**	**1**	**0**		**100**
Animal 2(u_2)	**1**	**0**	**1**		**200**

El parámetro α se obtiene como $(1-h^2)/h^2 = \mathbf{1}$. A^{-1} se construye con las reglas de Henderson; en este caso ambos animales son fundadores (ningún padre conocido), de manera que al aplicar las reglas de Henderson nos queda una matriz identidad de tamaño 2: $A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Las ecuaciones del modelo mixto que se pedían son:

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 300 \\ 100 \\ 200 \end{bmatrix}$$

Y aunque no se pedía en el examen, las soluciones al modelo se pueden obtener:

$$\begin{bmatrix} \hat{\mu} \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 300 \\ 100 \\ 200 \end{bmatrix} = \begin{bmatrix} 150 \\ -25 \\ 25 \end{bmatrix}$$

- Realizar la valoración genética también mediante un índice de selección

El único índice posible con la información disponible por animal es un índice de selección individual ya que sólo se dispone de un dato de cada animal. Según este índice:

$$\hat{u}_i = h^2 y_i$$

siendo y_i el dato del animal i desviado respecto de la media. Siguiendo la metodología BLUP, es necesario obtener previamente el valor de los efectos fijos promediando los datos presentes

dentro de cada nivel de efecto fijo. En este caso el único efecto fijo es la media general y su estimación se obtendría directamente promediando los datos que hay:

$$\hat{\mu} = \frac{100 + 200}{2} = 150$$

Conocida la media se asume que el valor estimado es su verdadero valor ($\boldsymbol{\mu}$ = **150**). A continuación se desvían los datos respecto de este valor:

$$y_1 = 100 - 150 = -50$$
$$y_2 = 200 - 150 = 50$$

Y aplicando el índice de selección, los valores genéticos de los animales serían:

$$\hat{\mathbf{u}}_1 = 0.5\ (y_1) = 0.5\ (-50) = \mathbf{-25}$$
$$\hat{\mathbf{u}}_2 = 0.5\ (y_2) = 0.5\ (50) = \mathbf{25}$$

- ¿Qué tipo de índice de selección sería?

Sería un índice de selección individual o fenotípico porque se utiliza el dato del propio individuo.

¿Qué diferencias hay entre los dos métodos?

El BLUP es un método de valoración que presenta la propiedad de ser insesgado en comparación con un índice de selección (BLP) que es un método sesgado. El sesgo se produce por el hecho de estimar los efectos fijos previamente y asumir que son los verdaderos valores para ajustar las observaciones. Posteriormente, con las observaciones ya ajustadas se realiza la valoración genética.

Si los animales no se distribuyen aleatoriamente entre los distintos niveles de efectos fijos, las estimaciones de los efectos fijos están incorporando parte de las soluciones para los efectos aleatorios (los valores genéticos), lo que provoca el sesgo. La otra diferencia fundamental entre los índices de selección y el BLUP consiste en que este último incorpora toda la información de parentesco presente en los datos, mientras que en la metodología de los índices de selección se suele restringir la información.

En este ejemplo el único efecto fijo es la media general por lo que es imposible la distribución desigual de los animales por efectos fijos. Además no existe información de parentesco por lo que en este ejemplo ambos métodos son equivalentes, como lo muestran los resultados obtenidos.

6- Deducir el valor del coeficiente de un índice de selección en los dos casos siguientes:

I) El dato a utilizar es el del propio individuo

II) El dato a utilizar es el del padre del individuo

Comentar cuál de los dos índices proporcionaría una mejor valoración si la heredabilidad (h^2) en el primer caso fuera de 0.10 y en el segundo caso 0.40.

Hay que predecir el valor genético aditivo mediante:

$$\hat{u}_i = C'V^{-1}y$$

- **En el caso de un índice de selección individual:**

  $C' = Cov(u_i, y_i) = Cov(u_i, u_i + e_i) = Cov(u_i, u_i) = \sigma_u^2$

  $V = Var(y_i) = \sigma_p^2$

  $\hat{u}_i = h^2 y_i$

- **En el caso de un índice de selección sobre un ascendiente:**

  $C' = Cov(u_i, y_j) = Cov(u_i, u_j + e_j) = Cov(u_i, u_j) = ½ \sigma_u^2$

  $V = Var(y_j) = \sigma_p^2$

  $\hat{u}_i = ½ h^2 y_j$

Para comparar los índices hay que calcular la precisión.
En ambos casos: $G = Var(u_i) = \sigma_u^2$:

$$\rho_1 = \sqrt{\frac{\mathbf{C'V^{-1}C}}{\mathbf{G}}} = h = \sqrt{0.10}$$

$$\rho_2 = \sqrt{\frac{\mathbf{C'V^{-1}C}}{\mathbf{G}}} = \sqrt{\frac{1}{4}h^2} = \sqrt{\frac{1}{4}0,40} = \sqrt{0,10}$$

Por tanto, ambos índices serían equivalentes en términos de precisión.

7- Para un carácter de heredabilidad $h^2 = 0.25$, un índice de selección individual produce una precisión del 50% ($\rho = \sqrt{0,25} = 0,5$).

a) Calcular el número de datos (m) que sería necesario para obtener la misma precisión si la heredabilidad fuese $h^2 = 0.1$, el criterio de selección fuese la media de los caracteres del individuo y la varianza ambiental permanente fuera nula ($h^2 = t$, siendo t la repetibilidad del carácter). La precisión de un índice de selección basado en la media de las producciones de un individuo es $\rho = \sqrt{\frac{mh^2}{1+t(m-1)}}$.

b) ¿Y si en las condiciones del apartado a) se desease una precisión del 80%?

a) Igualando el cuadrado de la precisión (0.25) a su valor en la expresión en la que $h^2 = t = 0.1$:

$$0.25 = \frac{mh^2}{1+t(m-1)} = \frac{0.1m}{1+0,1(m-1)}$$

Y despejando m:

$$\boxed{m = 3}$$

b) En este segundo apartado lo que cambia es la precisión siendo $\rho = 0.8$

$$0,8^2 = 0,64 = \frac{mh^2}{1+t(m-1)} = \frac{0.1m}{1+0,1(m-1)}$$

Y despejando m:

$$\boxed{m = 16}$$

8- Definir (a) ecuación, b) esperanzas y varianzas, c) asunciones, restricciones y limitaciones) razonadamente un modelo lineal mixto, para realizar una evaluación genética para el carácter producción de leche estandarizada a 150 días en una población de ovino lechero.

La definición de un modelo lineal mixto implica los tres apartados considerados en el enunciado del problema. Aunque no existe un único modelo válido, para el caso particular que se pide aquí, un modelo podría ser el siguiente:

a) Ecuación:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e},$$

siendo **y** el vector que contiene las producciones de las ovejas normalizadas a 150 días, **b** el vector de efectos fijos que contiene los niveles de tres efectos fijos que son: el efecto combinado Rebaño-Año_de_Parto-Estación_de_Parto, el efecto número de parto, y el efecto número de corderos (1, 2, 3 y 4 o más); **u** sería el vector que contiene los niveles de los efectos genéticos aditivos de todos los animales, **p** el vector que contiene los efectos ambientales permanentes de las ovejas con registro de producción, e el vector de residuos, y **X**, **Z**, y **W**, las matrices de incidencia o matrices diseño de los vectores **b**, **u** y **p** respectivamente.

b) Esperanzas y varianzas del modelo:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_{ep}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix},$$

siendo σ_u^2, σ_{ep}^2 y σ_e^2 las varianzas genética aditiva, ambiental permanente y residual, respectivamente, y **A** la matriz de relaciones aditivas entre individuos.

c) Asunciones restricciones y limitaciones del modelo.
Se asume que todas las ovejas tienen el primer parto a la misma edad y lo mismo sucede con sucesivos partos.

NOTA: Se ha utilizado un modelo de repetibilidad que incluye el efecto ambiental permanente como efecto aleatorio, al tratarse de un carácter que puede ser medido varias veces en un mismo individuo.

9- Un animal es valorado genéticamente para dos caracteres obteniendo respectivamente los valores de +10 y +4.5 respectivamente para cada uno de ellos. La varianza genética aditiva del primer carácter es 100 y la del segundo carácter es 9. ¿Para cuál de los dos caracteres es mejor? Razonar la respuesta utilizando para los dos caracteres un índice con media 100 y varianza 400.

Los valores genéticos que se obtienen tras las valoraciones genéticas cumplen las asunciones del modelo por lo que tienen media cero y varianza σ_u^2, la varianza genética aditiva del carácter. Los valores genéticos para distintos caracteres no pueden ser comparados tal cuál resultan tras la valoración genética, por depender de la heredabilidad del carácter y de la escala de medida de cada uno. Sin embargo, pueden ser transformados a variables con misma media y varianza para poder establecer la comparación. Para ello basta con dividir el valor genético por la desviación típica genética del carácter (σ_u) con lo que el valor resultante tendría desviación típica 1 (σ_u =1) y ya podría establecerse la comparación.

El enunciado solicita que la comparación se establezca con varianza 400 (σ_u^2 = 400), es decir desviación típica 20 (σ_u = 20) para lo cuál habrá que multiplicar por este valor el dato que tenía desviación típica uno. Como además la media debe establecerse en 100, basta sumar este valor:

- Carácter 1 ($\hat{u}_1$ = + 10 con σ_u^2 = 100 y por tanto σ_u = 10)

$$I_1 = \hat{u}_1 \frac{20}{\sigma_{u_1}} + 100 = 10\frac{20}{10} + 100 = 120$$

- Carácter 2 ($\hat{u}_2$ = + 4.5 con σ_u^2 = 9 y por tanto σ_u = 3)

$$I_2 = \hat{u}_2 \frac{20}{\sigma_{u_2}} + 100 = 4.5 \frac{20}{3} + 100 = 130$$

Como I_2 es mayor que I_1, el individuo es mejor para el carácter 2.

10- Se pretende valorar dos individuos para un carácter determinado, cuya varianza genética aditiva es 80, mediante un índice de selección basado en el rendimiento de los individuos para otros dos caracteres con los que está correlacionado. Las covarianzas genéticas aditivas entre el carácter objetivo de selección y los dos utilizados como criterio son $\sigma_{u,y1} = 50$ y $\sigma_{u,y2} = -10$. Los dos caracteres que se registran no están correlacionados y presentan varianzas fenotípicas $\sigma_{y1}^2 = 200$ y $\sigma_{y2}^2 = 5$.

a) **Construir el índice de selección.**
b) **¿Cuál es la precisión del índice?**
c) **Valorar los dos individuos y decir cuál de los dos es mejor según sus registros productivos:**

ANIMAL	Y1	Y2
1	500	20
2	400	10

a) Basta con sustituir los valores correspondientes en la expresión del índice de selección: $\hat{u} = \mathbf{C'V^1y}$. Es importante destacar que **V** es en este caso una matriz diagonal por lo que su inversa se obtiene directamente elemento a elemento

$$\hat{u} = \mathbf{C'V^1y} = \begin{bmatrix} 50 & -10 \end{bmatrix} \begin{bmatrix} 200 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} =$$

$$= \begin{bmatrix} 50 & -10 \end{bmatrix} \begin{bmatrix} \frac{1}{200} & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} =$$

$$= 0{,}25\ y_1 - 2\ y_2$$

b) $$\rho = \sqrt{\frac{C'V^{-1}C}{G}} = \sqrt{\frac{\begin{bmatrix} \frac{1}{4} & -2 \end{bmatrix}\begin{bmatrix} 50 \\ -10 \end{bmatrix}}{80}} = \sqrt{\frac{32,5}{80}} = 0,64$$

c) Utilizando el índice del apartado a) podemos obtener una predicción del valor genético aditivo de cada individuo utilizando los datos desviados de la media:

$\hat{u}_1 = 0,25\ y_1 - 2\ y_2 = 0,25$ x (500-450) – 2 x (20-15) = 2,5

$\hat{u}_2 = 0,25\ y_1 - 2\ y_2 = 0,25$ x (400-450) – 2 x (10-15) = -2,5

El animal 1 es mejor que el animal 2 porque $\hat{u}_1 = 2,5 > \hat{u}_2 = -2,5$

11- La ecuación del modelo que se define a continuación para valorar animales para el peso al destete presenta algunos errores. ¿Cuáles son? ¿Por qué? Comentar además algunas posibles mejoras del modelo y escribir las esperanzas y varianzas del modelo:

$\boldsymbol{y = Xb + Zu + Wp + e}$

siendo y el vector de pesos al destete de los terneros, b el vector de efectos fijos del modelo incluyendo el sexo con 2 niveles, el efecto combinado rebaño-año de parto con 10 niveles y la edad al destete del ternero como variable fija continua; u es el vector de efectos genéticos aditivos con n niveles, p el vector de ambientes permanentes con n niveles y e el vector de residuos con na niveles.

(Hay muchas soluciones correctas a este problema. La que aparece aquí es sólo una de las posibles)

Errores del modelo:

- El efecto ambiental permanente se ajusta en los modelos que pretenden analizar caracteres medidos de forma repetida en los mismos individuos, como por ejemplo la producción de leche o el resultado de una prueba deportiva. El peso al destete es un dato único y por tanto no debe ajustarse este efecto. Una salvedad habría que hacer a esto: podría ajustarse el efecto ambiental permanente materno, que se definiría con distintos niveles para cada madre. En este caso el número de niveles debería coincidir con el número de madres, número que será siempre inferior al de animales y no igual como dice el enunciado.
- Un carácter medido a edades tempranas debería incluir el efecto materno, incluyendo el vector de incógnitas **m** y su matriz diseño **M**. Dado que la capacidad materna también se hereda, se tratará de un efecto genético, es decir, con el mismo número de niveles que el efecto genético aditivo, y con varianza $\text{Var}(\mathbf{m})=\mathbf{A}\sigma_m^2$.

Y el modelo quedaría: $\boldsymbol{y = Xb + Zu + Mm + e}$

Mejoras del modelo: Cualquier inclusión de efectos ambientales razonados como efecto fijo sería una mejora razonable del modelo. Por ejemplo, el efecto nº de parto de la madre (las vacas de primer parto darán terneros más pequeños, y eso hay que tenerlo en cuenta). La inclusión del efecto aleatorio ambiental permanente materno (cada vaca tiene más de un hijo), es otra posible mejora del modelo.

Esperanzas y Varianzas del Modelo: Una vez corregidos los errores anteriores el modelo quedaría de la siguiente manera:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad V\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{A}\sigma_{um} & \mathbf{0}\\ \mathbf{A}\sigma_{mu} & \mathbf{A}\sigma_m^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

Siendo **A** la matriz de relaciones aditivas, **I** la matriz identidad, σ_u^2 la varianza genética aditiva, σ_m^2 la varianza genética materna y $\sigma_{um}=\sigma_{mu}$ la covarianza entre efectos genéticos aditivo directo y materno.

12- Tras una valoración genética BLUP en la que se han ajustado dos efectos fijos (a y e) además de la media general (μ), y el efecto genético aditivo (u, σ_u = 10), se obtienen los siguientes valores:

$\hat{\mu} = 0$ $\hat{a}_1$ = 0, $\hat{a}_2$ = -5 $\hat{e}_1$ = 18, $\hat{e}_2$ = 16, $\hat{e}_3$ = 14 $\hat{u}_1$ = -10, $\hat{u}_2$ = 10, $\hat{u}_3$ = -5, $\hat{u}_4$ = 15

a) ¿Cómo se ha resuelto la dependencia lineal que existía entre los efectos fijos?

b) ¿Se puede saber cuánto es mejor el nivel de mayor producción de cada uno de los niveles de los efectos fijos en relación al resto?

c) Transformar los valores genéticos en otros de media 100 y varianza 400.

a) Cuando además de la media general, los modelos incluyen efectos fijos discontinuos, nos encontramos ante modelos de rango no completo. En ellos va a aparecer información redundante de manera que sobrará una ecuación por cada nuevo efecto fijo discontinuo que se incluya. En este problema hay dos efectos fijos por lo que este fenómeno se da por partida doble, de manera que la ecuación de la media general explica lo mismo que las dos del primer efecto fijo o las tres del segundo efecto fijo. Es preciso recurrir a una inversa generalizada para resolver el modelo. Una inversa generalizada sencilla se obtiene eliminando ecuaciones que contengan información redundante e invirtiendo el resto de la matriz de coeficientes. Las ecuaciones eliminadas darán lugar a soluciones nulas. A la vista de los resultados en este caso se ha eliminado la ecuación correspondiente a la media y la primera del primer nivel del efecto fijo.

b) El empleo de inversas generalizadas lleva a obtención de soluciones no únicas. Sin embargo determinadas combinaciones de las incógnitas sí que son únicas independientemente de la inversa generalizada que se

utilice. Entre estas combinaciones están las diferencias entre los niveles del mismo efecto fijo por lo que SÍ se puede conocer las diferencias entre los niveles de los efectos fijos. Así, por ejemplo, un individuo que registra su dato en el primer nivel del primer efecto fijo producirá 5 unidades más del carácter que en el segundo, y uno que registre su dato en el primer nivel del segundo efecto fijo producirá 2 unidades más que si lo hubiera registrado en el segundo nivel o 4 unidades más que en el tercero.

c) Basta con tipificar u_i dividiendo por la desviación típica y reescalarla a la media y desviación típica deseada: $I_i = \frac{\hat{u}_i}{\sigma_u} 20 + 100 = 2\hat{u}_i + 100$:

$I_1 = 80$; $I_2 = 120$; $I_3 = 90$; $I_4 = 130$

13- Deducir la genealogía a partir de la cual se ha obtenido la inversa de la matriz de relaciones aditivas al margen (Se sugiere empezar por el último individuo e ir deduciendo hacia atrás descontando coeficientes. El sexo de los individuos que actúen como padre/madre no es importante). Con la genealogía obtenida construir la matriz de relaciones aditivas original A.

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{11}{6} & -\frac{2}{3} & \frac{1}{2} & -1 \\ -\frac{2}{3} & \frac{5}{3} & -\frac{2}{3} & 0 \\ \frac{1}{2} & -\frac{2}{3} & \frac{11}{6} & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

Si comenzamos por el individuo 4, a partir del valor *'2'* de la diagonal y de los *'-1'* de su fila y columna, se concluye que los padres del 4 son el 1 y el 3. Descontando todos los coeficientes que corresponden según las reglas de Henderson se llega a una inversa parcial de A-1. Procediendo recursivamente con todos los individuos se obtiene:

$$\mathbf{A}_{3}^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} & 0 \\ -\frac{2}{3} & \frac{5}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{bmatrix}$$ -(3 hijo de 2) ->

$$\mathbf{A}_{2}^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{bmatrix}$$ -(2 hijo de 1) -> $\mathbf{A}_{1}^{-1} = [1]$

De modo que la genealogía y la matriz de relaciones aditivas obtenida con el método tabular son:

Individuo	*Padre*	*Madre*
1	0	0
2	1	0
3	0	2
4	1	3

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} & \frac{5}{8} \\ \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 & \frac{5}{8} \\ \frac{5}{8} & \frac{1}{2} & \frac{5}{8} & \frac{9}{8} \end{bmatrix}$$

14- Dadas las siguientes matrices: $\mathbf{C} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$,

$\mathbf{V} = \begin{bmatrix} 1/10 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ **y ,** $\widehat{H} = I = \mathbf{v'C'V^{-1}y}$ **siendo H el agregado genético-económico e I el índice de selección, se pide:**

a) Número de caracteres en el objetivo y en el criterio de selección.

b) Covarianza entre el primer objetivo y el segundo criterio.

c) Pesos económicos de cada uno de los caracteres del agregado genético-económico.

d) Ponderaciones de cada uno de los caracteres presentes en el índice de selección.

A la vista de la expresión $\widehat{H} = I = \mathbf{v'C'V^{-1}y}$ se pueden identificar los distintos componentes del índice de selección. **v'** es el vector de pesos económicos de los caracteres del objetivo, cuyo número es la dimensión del vector, 2. **C'**, además de ser la transpuesta de **C**, es la matriz de covarianzas entre los caracteres del objetivo y los caracteres del índice de selección, **C'** = cov (**u** , **y'**), y tendrá tantas filas como caracteres en el objetivo y tantas columnas como caracteres en el criterio; las correspondientes dimensiones de **C'** (al contrario que las de **C**), son 2 y 3. Finalmente **V** (y también su inversa $\mathbf{V^{-1}}$), es la matriz de varianzas y covarianzas de los criterios, por lo que será una matriz cuadrada con tamaño igual al número de criterios considerados, en este caso, 3. Las respuestas a los apartados del problema son, por tanto:

a) 2 en el objetivo y 3 en el criterio.

b) Si

$$\mathbf{C}' = \text{cov}(u, y') = \text{cov}\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}\right) = \begin{bmatrix} \sigma_{u_1 y_1} & \sigma_{u_1 y_2} & \sigma_{u_1 y_3} \\ \sigma_{u_2 y_1} & \sigma_{u_2 y_2} & \sigma_{u_2 y_3} \end{bmatrix}$$

y $\mathbf{C}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, entonces el valor pedido, $\sigma_{u_1 y_2} = 2$.

c) Directamente de **v**, el peso económico del primero es 3 y el del segundo es 1.

d) Es preciso calcular los coeficientes del índice de selección:

$$I = \mathbf{v'C'V^{-1}y} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}' \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1/10 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} =$$

$$= \begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} =$$

$$= \begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 10 & 10 & 6 \\ 40 & 25 & 12 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} =$$

$$= \begin{bmatrix} 70 & 55 & 30 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 70y_1 + 55y_2 + 30y_3 .$$

Por tanto los coeficientes de los caracteres son 70, 55 y 30.

15- a) Construir <u>la inversa</u> de la matriz de relaciones aditivas, b) completar la tabla de datos y c) deducir el valor de la heredabilidad a partir de las ecuaciones del modelo mixto:

$$\begin{bmatrix} 4 & 2 & 2 & 0 & 0 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & 4 & 1 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 & 1 & 4 & -1 & -1 \\ 1 & 0 & 1 & 0 & 0 & -1 & -1 & 3 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 & -1 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{R}_1 \\ \hat{R}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \end{bmatrix} = \begin{bmatrix} 70 \\ 25 \\ 45 \\ 0 \\ 0 \\ 10 \\ 15 \\ 20 \\ 25 \end{bmatrix}$$

TABLA DE DATOS

Individuo	Padre	Madre	Rebaño	Dato
1	**0**	**0**	**-**	**-**
2	**0**	**0**	**-**	**-**
3	**1**	**2**		**10**
4	**1**	**2**		
5	**3**	**4**		**20**
6	**3**	**4**		

a) Directamente se aplican las reglas de Henderson y se construye la matriz:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1+½+½ & ½+½ & -1 & -1 & & \\ ½+½ & 1+½+½ & -1 & -1 & & \\ -1 & -1 & 2+½+½ & ½+½ & -1 & -1 \\ -1 & -1 & ½+½ & 2+½+½ & -1 & -1 \\ & & -1 & -1 & 2 & \\ & & -1 & -1 & & 2 \end{bmatrix} =$$

$$= \begin{bmatrix} 2 & 1 & -1 & -1 & 0 & 0 \\ 1 & 2 & -1 & -1 & 0 & 0 \\ -1 & -1 & 3 & 1 & -1 & -1 \\ -1 & -1 & 1 & 3 & -1 & -1 \\ 0 & 0 & -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix}$$

b) De las ecuaciones del modelo mixto, el bloque de la matriz de coeficientes correspondiente a los efectos aleatorios representa $\mathbf{Z'Z + A^{-1}}\alpha$, luego:

$$\mathbf{Z'Z} + \mathbf{A^{-1}}\alpha = \begin{bmatrix} 2 & 1 & -1 & -1 & 0 & 0 \\ 1 & 2 & -1 & -1 & 0 & 0 \\ -1 & -1 & 4 & 1 & -1 & -1 \\ -1 & -1 & 1 & 4 & -1 & -1 \\ 0 & 0 & -1 & -1 & 3 & 0 \\ 0 & 0 & -1 & -1 & 0 & 3 \end{bmatrix} =$$

$$= \mathbf{Z'Z} + \begin{bmatrix} 2 & 1 & -1 & -1 & 0 & 0 \\ 1 & 2 & -1 & -1 & 0 & 0 \\ -1 & -1 & 3 & 1 & -1 & -1 \\ -1 & -1 & 1 & 3 & -1 & -1 \\ 0 & 0 & -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix} \alpha$$

Como los elementos de fuera de la diagonal de **Z'Z** deben ser nulos, se deduce que el valor de α es 1, y que se ha añadido un 1 en las 4 últimas diagonales, lo que indica que los individuos del 3 al 6 tienen un dato. Por las posiciones de los unos en la matriz de coeficientes (fila 2, columnas 6 y 7; fila 3, columnas 8 y 9) se deduce que los individuos 3 y 4 son del rebaño 1, y los otros 2 son del rebaño 2. En el lado derecho de las ecuaciones están las suma de los datos para cada rebaño en las posiciones 2 y 3, por lo que se puede deducir el rendimiento por diferencia. La tabla queda así:

TABLA DE DATOS

Individuo	Padre	Madre	Rebaño	Dato
1	0	0	-	-
2	0	0	-	-
3	1	2	1	10
4	1	2	1	15
5	3	4	2	20
6	3	4	2	25

c) Dado que $\alpha = 1$, se puede despejar y obtener su valor:

$$\alpha = 1 = \frac{1-h^2}{h^2} \Rightarrow h^2 = 1-h^2 \Rightarrow 1 = 2h^2 \Rightarrow h^2 = 0,5$$

16- Se desea valorar genéticamente un padre y un hijo aprovechando para ambos su propio dato y el del otro pariente. Por tanto, se busca el valor genético de dos individuos u_h y u_p a partir de sus datos y_h y y_p. $\left(\mathbf{u}=\begin{bmatrix} u_p \\ u_h \end{bmatrix}; \mathbf{y}=\begin{bmatrix} y_p \\ y_h \end{bmatrix}\right)$

a) Construir la matriz C' = *Cov*(u, y') de la expresión del índice $\hat{\mathbf{u}} = \mathbf{C'V^{-1}y}$.

b) Construir $\mathbf{G^{-1}}$, siendo G = *Var* (u) = $\mathbf{A}\sigma_u^2$

c) Obtener razonadamente el resultado del producto $\mathbf{C'G^{-1}}$ sin llevarlo a cabo.

Dejar los apartados a) y b) en función de los parámetros σ_p^2, σ_u^2 y/o h^2.

a) Construcción de los elementos del índice:

- $\mathbf{C'} = Cov(\mathbf{u}, \mathbf{y'}) = Cov\left(\begin{bmatrix} u_p \\ u_h \end{bmatrix}, \begin{bmatrix} y_p \\ y_h \end{bmatrix}'\right) = \begin{bmatrix} \sigma_{u_p y_p} & \sigma_{u_p y_h} \\ \sigma_{u_h y_p} & \sigma_{u_h y_h} \end{bmatrix}$

- $\sigma_{u_p y_p} = \sigma_{u_h y_h} = Cov(u_p, y_p) = Cov(u_p, u_p + e_p) = Cov(u_p, u_p) = Var(u_p) = \sigma_u^2$
- $\sigma_{u_p y_h} = Cov(u_p, y_h) = Cov(u_p, u_h + e_h) = Cov(u_p, \frac{1}{2}u_p + \frac{1}{2}u_m + \varphi_h) = \frac{1}{2} Var(u_p) = \frac{1}{2}\sigma_u^2$
- $\sigma_{u_h y_p} = Cov(u_h, y_p) = Cov(u_h, u_p + e_p) = Cov(\frac{1}{2}u_p + \frac{1}{2}u_m + \varphi_h, u_p) = \frac{1}{2} Var(u_p) = \frac{1}{2}\sigma_u^2$

$$\mathbf{C'} = \begin{bmatrix} \sigma_u^2 & \frac{1}{2}\sigma_u^2 \\ \frac{1}{2}\sigma_u^2 & \sigma_u^2 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}\sigma_u^2 = \mathbf{C}$$

b) $\mathbf{G^{-1}} = \mathbf{A^{-1}} \, {}^{1}\!/_{\sigma_u^2}$ y la inversa da **A** por las reglas de Henderson es:

$$\begin{bmatrix} {}^{4}\!/_{3} & {}^{-2}\!/_{3} \\ {}^{-2}\!/_{3} & {}^{4}\!/_{3} \end{bmatrix}:$$

$$\mathbf{G}^{-1} = \begin{bmatrix} {}^{4}\!/_{3} & {}^{-2}\!/_{3} \\ {}^{-2}\!/_{3} & {}^{4}\!/_{3} \end{bmatrix} \frac{1}{\sigma_u^2}$$

c) Dado que los coeficientes de **A** representan el porcentaje de genes que comparten los individuos, la matriz **C'** coincide $\mathbf{A}\sigma_u^2$, y por lo tanto con **G**. Así pues:

$$\mathbf{C'G^{-1}} = \mathbf{GG^{-1}} = \mathbf{I}$$

17.- Dados los datos de la tabla, construir <u>únicamente la ecuación</u> de dos modelos fijos especificando cada uno de los elementos de los vectores y matrices que la componen en cada uno de los siguientes casos:
a) Además de la media se pretende ajustar la influencia de cada uno de los efectos ganadería y año pero no la de su interacción.
b) Además de la media y de los efectos se pretende ajustar también la interacción entre ganadería y año.

	Ganadería A	Ganadería B
Año 2008	**5** **4**	**2** **3**
Año 2009	**3** **4**	**4** **6**

Se pide la ecuación de un modelo fijo, luego será de la forma **y** = **Xb** + **e**, siendo **y** el vector de datos, **X** la matriz de incidencia de los efectos fijos, **b** el vector de incógnitas fijas y **e** el vector de residuos.

a) El vector de efectos fijos contiene la media general μ, y dos niveles de cada uno de los efectos fijos Ganadería y Año:

$$\begin{bmatrix}5\\4\\2\\3\\3\\4\\4\\6\end{bmatrix} = \left[\begin{array}{c|cc|cc} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{array}\right] \begin{bmatrix}\mu\\G_A\\G_B\\A_{2008}\\A_{2009}\end{bmatrix} + \begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\\e_7\\e_8\end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}$$

b) En este caso en lugar de dos efectos fijos además de la media general, habrá un único efecto fijo con 4 niveles

$$\begin{bmatrix}5\\4\\2\\3\\3\\4\\4\\6\end{bmatrix} = \left[\begin{array}{c|cccc}1&1&0&0&0\\1&1&0&0&0\\1&0&1&0&0\\1&0&1&0&0\\1&0&0&1&0\\1&0&0&1&0\\1&0&0&0&1\\1&0&0&0&1\end{array}\right] \begin{bmatrix}\mu\\G_A A_{2008}\\G_B A_{2008}\\G_A A_{2009}\\G_B A_{2009}\end{bmatrix} + \begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\\e_7\\e_8\end{bmatrix}$$

$$\mathbf{y} = \mathbf{X} \quad \mathbf{b} + \mathbf{e}$$

18.- Consideremos dos individuos, uno hijo del otro. El padre tienes dos registros del carácter, 25 y 20; el hijo sólo tiene un dato, 22. En la evaluación genética se ajusta un modelo de repetibilidad con ecuación: $\mathbf{y = Xb + Zu + Wp + e}$. El único efecto fijo en b es la media general, u es el vector de efectos genéticos aditivos, p el vector de efectos ambientales permanentes, e el vector de residuos, y X, Z, y W, las correspondientes matrices diseño. La definición del modelo se completa con las esperanzas y varianzas:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{b}\\ \mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ \mathbf{b}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_{ep}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2\end{bmatrix}$$

Construir las ecuaciones del modelo mixto sabiendo que $\sigma_u^2 = 1$, $\sigma_{ep}^2 = 3$ y $\sigma_e^2 = 3$. Obsérvese que en este modelo α no puede obtenerse directamente de la heredabilidad; de hecho habrá que considerar uno para cada efecto aleatorio: $\alpha_u = \sigma_e^2 / \sigma_u^2$ y $\alpha_p = \sigma_e^2 / \sigma_{ep}^2$

Se trataría de construir las siguientes ecuaciones del modelo lineal mixto con la media general como único efecto fijo y los valores genéticos de dos animales como único efecto aleatorio:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + A^{-1}\alpha_u & Z'W \\ W'X & W'Z & W'W + I\alpha_{ep} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$

Parte de la matriz de coeficientes $\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z & Z'W \\ W'X & W'Z & W'W \end{bmatrix}$, y el vector del lado derecho (LD) $\begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$, se rellenan haciendo una tabla de doble entrada con todas las incógnitas y contando el número de observaciones de los cruces (matriz de coeficientes), y sumando observaciones (vector del lado derecho):

	Media(μ)	Anim 1 (u_1)	Anim 2 (u_2)	Perm 1 (p_1)	Perm 2 (p_2)		LD
Media (μ)	**3**	**2**	**1**	**2**	**1**		**67**
Animal 1 (u_1)	**2**	**2**	**0**	**2**	**0**		**45**
Animal 2(u_2)	**1**	**0**	**1**	**0**	**1**		**22**
Perman. 1 (p_1)	**2**	**2**	**0**	**2**	**0**		**45**
Perman. 2 (p_2)	**1**	**0**	**1**	**0**	**1**		**22**

Los parámetro α se obtienen como $\alpha_u = \sigma_e^2 / \sigma_u^2 = 3/1 = 3$ y $\alpha_p = \sigma_e^2 / \sigma_{ep}^2 = 3/3 = 1$. A^{-1} se construye con las reglas de Henderson; en este caso el primero es fundador por lo que habrá un 1 en su diagonal y el segundo es hijo del 1 por lo que se anota 4/3 en su diagonal, −2/3 en el cruce con el 1 y 1/3 en la diagonal del 1, de manera que al aplicar las reglas de Henderson nos queda una matriz identidad de tamaño 2: $A^{-1} = \begin{bmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{bmatrix}$, y al multiplicar por $\alpha_u = 3$ nos queda $A^{-1}\alpha_u = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$. $I\alpha_{ep}$ resulta finalmente una matriz identidad. Se reúne todo y se obtienen las ecuaciones del modelo mixto que se pedían:

$$\begin{bmatrix} 3 & 2 & 1 & 2 & 1 \\ 2 & 6 & -2 & 2 & 0 \\ 1 & -2 & 5 & 0 & 1 \\ 2 & 2 & 0 & 3 & 0 \\ 1 & 0 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} 67 \\ 45 \\ 22 \\ 45 \\ 22 \end{bmatrix}$$

Y aunque no se pedía en el examen, las soluciones al modelo se pueden obtener:

$$\begin{bmatrix} \hat{\mu} \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 & 2 & 1 \\ 2 & 6 & -2 & 2 & 0 \\ 1 & -2 & 5 & 0 & 1 \\ 2 & 2 & 0 & 3 & 0 \\ 1 & 0 & 1 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 67 \\ 45 \\ 22 \\ 45 \\ 22 \end{bmatrix} = \begin{bmatrix} 22{,}2826 \\ 0{,}0217 \\ -0{,}0217 \\ 0{,}1304 \\ -0{,}1304 \end{bmatrix}$$

Obsérvese que se ha empleado la misma información para obtener los valores genéticos de los individuos y los ambientes permanentes asociados a cada uno de ellos. Sin embargo el modelo ha sido capaz de distinguir entre ellos gracias a la información de parentesco. Obsérvese también que la media de los efectos aleatorios es cero tal y como se especifica en la definición del modelo.

19.- Del cruce de un individuo con su hija nace un tercer individuo. De los tres se tiene registrado un único dato de un carácter de interés. Concretar todos y cada uno de los elementos de las matrices de varianzas y covarianzas de los efectos aleatorios de un modelo de evaluación genética para estos datos en el que el único efecto aleatorio además del residuo es el efecto genético aditivo, la varianza residual es y la heredabilidad 0,40.

Si la heredabilidad es 0,40, entonces:

$$h^2 = 0,40 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \frac{\sigma_u^2}{\sigma_u^2 + 150}$$

$$0,40\left(\sigma_u^2 + 150\right) = \sigma_u^2$$

$$0,40\left(\sigma_u^2\right) + 0,40\left(150\right) = \sigma_u^2$$

$$0,60\sigma_u^2 = 60$$

$$\sigma_u^2 = 100$$

- La varianza del vector de efectos genéticos aditivos es la matriz **G** que se corresponde con el producto de la matriz de relaciones aditivas **A** y la varianza genética aditiva (σ_u^2).

El pedigrí descrito en el enunciado se puede escribir:

Individuo	Padre	Madre
1	0	0
2	1	0
3	1	2

Y la matriz **A** se construye con el método tabular:

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 & 3/4 \\ 1/2 & 1 & 3/4 \\ 3/4 & 3/4 & 5/4 \end{bmatrix}$$

y

$$\mathbf{G} = \mathbf{A}\sigma_u^2 = \begin{bmatrix} 1 & 1/2 & 3/4 \\ 1/2 & 1 & 3/4 \\ 3/4 & 3/4 & 5/4 \end{bmatrix} \sigma_u^2 = \begin{bmatrix} 1 & 1/2 & 3/4 \\ 1/2 & 1 & 3/4 \\ 3/4 & 3/4 & 5/4 \end{bmatrix} 100 = \begin{bmatrix} 100 & 50 & 75 \\ 50 & 100 & 75 \\ 75 & 75 & 125 \end{bmatrix}$$

- La varianza del vector de residuos aditivos es la matriz **R** que se corresponde con el producto de la matriz identidad **I** y la varianza genética aditiva (σ_e^2)

$$\mathbf{R} = \mathbf{I}\sigma_e^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \sigma_e^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} 150 = \begin{bmatrix} 150 & 0 & 0 \\ 0 & 150 & 0 \\ 0 & 0 & 150 \end{bmatrix}$$

- Finalmente, todas las covarianzas entre todos los efectos genéticos aditivos y residuos son considerados nulos. Por tanto, las varianzas y covarianzas de los efectos aleatorios se concretan:

$$Var\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = Var\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} = \left[\begin{array}{ccc|ccc} 100 & 50 & 75 & 0 & 0 & 0 \\ 50 & 100 & 75 & 0 & 0 & 0 \\ 75 & 75 & 125 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 150 & 0 & 0 \\ 0 & 0 & 0 & 0 & 150 & 0 \\ 0 & 0 & 0 & 0 & 0 & 150 \end{array}\right]$$

BIBLIOGRAFÍA

Sólo soy consciente de alguna de la bibliografía consultada para realizar este texto, únicamente porque pertenecen a la lista de mis libros más visitados. No sería justo dejar de citarlos, porque en ellos encuentro siempre los detalles que necesito. Sin embargo, la justicia que se hace con ellos se convierte en injusticia para otros autores que de una u otra manera han contribuido a mis conocimientos sobre valoración genética a lo largo de mi experiencia profesional. Como ilustración, en estos apuntes pueden encontrarse algunos ejemplos prácticos que han llegado hasta mí durante la docencia que he recibido en mi formación. Al tiempo que agradecerles su contribución, quiero pedir disculpas a todos ellos por no haberlos podido citar propiamente. Se presentan a continuación mis textos de consulta habitual en relación a la valoración genética:

FALCONER, D.S. y MACKAY, T.F.C. 1996. *Introducción a la Genética Cuantitativa*. Ed. Acribia, S.A.

MRODE, R.A. 1996. *Linear models for the prediction of animal breeding values*. CAB International. 1996

NICHOLAS, F.W., 1996. *Introducción a la Genética Veterinaria*. Ed. Acribia.

RICO, M., 1999. *Los Modelos Lineales En La Mejora Genética Animal*. Ed. Marcos Rico Gutiérrez.

SCHAEFFER, L. R., 1993. *Linear models and computing strategies in animal breeding*. Handout, CGIL, University of Guelph, Guelph, ON, Canada.