UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"ESTIMACIÓN DE COMPONENTES DE VARIANZA UTILIZANDO LOS MÉTODOS BAYESIANO Y MÁXIMA VEROSIMILITUD RESTRINGIDA PARA EL ESTUDIO DE LA HEREDABILIDAD"

Presentada por:
ANA CECILIA VARGAS PAREDES

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTIAE EN:
ESTADÍSTICA APLICADA

Lima – Perú
2017

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
ESCUELA DE POSGRADO
MAESTRÍA EN ESTADÍSTICA APLICADA

"ESTIMACIÓN DE COMPONENTES DE VARIANZA UTILIZANDO LOS MÉTODOS BAYESIANO Y MÁXIMA VEROSIMILITUD RESTRINGIDA PARA EL ESTUDIO DE LA HEREDABILIDAD"

Presentada por:

ANA CECILIA VARGAS PAREDES

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

SUSTENTADA Y APROBADA POR EL SIGUIENTE JURADO

Mg.Sc. César Higinio Menacho Chiok
PRESIDENTE

Mg.Sc. Víctor Manuel Maehara Oyata
PATROCINADOR

Mg.Sc. Jesús Salinas Flores
MIEMBRO

Mg.Sc. Rino Nicanor Sotomayor Ruíz
MIEMBRO

AGRADECIMIENTOS

Quiero empezar agradeciendo al profesor Mg.Sc. Víctor Maehara por brindarme su apoyo como patrocinador y guía, con sus preguntas, críticas, sugerencias y observaciones, las cuales hicieron posible el desarrollo de esta tesis.

Quiero agradecer también al Ph.D. Gustavo Gutiérrez del departamento Académico de Producción Animal de la facultad de Zootecnia por haberme permitido asistir a algunas clases del curso Modelos Lineales y su aplicación en Ciencia Animal de la Escuela de Posgrado y trabajar en una investigación similar la cual aportó en la comprensión de algunos modelos estadísticos que involucran información genealógica, así también a la Mg.Sc. María Elisa García del Departamento Académico de Producción Animal de la facultad de Zootecnia por brindarme la oportunidad de iniciarme en el uso de algunos software especializados de evaluación genética.

Asimismo, agradezco a los profesores miembros del jurado Mg.Sc. Rino Sotomayor, Mg.Sc. César Menacho y MgSc. Jesús Salinas, por su tiempo y disposición para revisar esta tesis.

A mis familiares, amigos y amigas por el apoyo brindado.

RESUMEN

Se estimó mediante un modelo lineal mixto los componentes de varianza y heredabilidad de la producción de leche, a partir de los registros de 3397 lactaciones, provenientes de 1359 vacas de raza Holsteins, de 57 rebaños con información genealógica de 5 generaciones, utilizando máxima verosimilitud restringida conocida como REML y muestreo de Gibbs basado en procedimientos bayesianos. Con ambas metodologías se obtuvo una heredabilidad, en sentido amplio, moderada de 0.135 vía REML y una media de 0.318 vía muestreo de Gibbs. Para realizar el análisis exploratorio de residuales (en función de los tres tipos: marginal, residual condicional y efectos aleatorios) del modelo lineal mixto estimado vía REML, se adaptó funciones en R para incorporar la información genealógica o pedigrí al modelo. Como resultado de esto se verificó la linealidad de los efectos fijos y la normalidad del componente genético del animal. No se encontró normalidad para el efecto aleatorio del rebaño ni para los residuales condicionales. Para estos últimos tampoco se observó homocedasticidad. Además, se encontró que para 132 animales la estructura de covarianza considerada en el modelo no es adecuada. También, se observó 215 animales y 7 rebaños con efectos atípicos. En el diagnóstico del procedimiento de simulación del muestreo de Gibbs desde la perspectiva bayesiana no se encontró problemas de convergencia. Se obtuvieron errores de Montecarlo bajos y tamaños efectivos de muestra mayores a 1000 para cada componente del modelo.

Palabras clave: Modelo lineal mixto, REML, residuales, análisis bayesiano.

ABSTRACT

The components of variance and heritability for milk production were estimated using a mixed linear model from the records of 3397 lactations from 1359 Holsteins cows in 57 herds with genealogical information of 5 generations using maximum restricted verisimilitude known as REML and Gibbs sampling based on Bayesian procedures. Both methodologies estimated moderate broad heritability of 0.135 via REML and average of 0.318 via Gibbs sampling. The exploratory analysis of residuals was done according to three types of residuals (marginal residual, conditional residual and random effects) of the mixed linear model estimated via REML. R functions were adapted to incorporate genealogical or pedigree information to the model. As a result of this, linearity of the fixed effects and normality of genetic component of the animal were verified. However, normality of random effects and conditional residuals were not found, neither homoscedasticity for the latter. In addition, it was found that the covariance structure considered in the model is not adequate for 132 animals. It was observed 215 animals and 7 herds with atypical effects. In the diagnosis of the Gibbs sampling simulation procedure from the Bayesian perspective, there were not problems of chain convergence, low Montecarlo errors and effective sample sizes greater than 1000 were obtained for each component of the model.

Keywords: Mixed linear model, REML, residuals, Bayesian analysis.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

I. INTRODUCCIÓN

Los componentes de varianza son parámetros correspondientes a las varianzas de los efectos aleatorios de un modelo. La estimación de estos componentes se ha convertido en una metodología de análisis muy utilizada en diferentes áreas, tales como: mejoramiento animal, biología en general, ensayos clínicos, proceso de manufacturación, psicología, sociología, etc. En el campo de mejoramiento genético animal, la estimación de parámetros genéticos como la heredabilidad (proporción de la varianza fenotípica atribuida a factores genéticos aditivos) es usado para predecir los valores de cría a partir de los cuales se realiza la selección de los animales. Los procedimientos estadísticos utilizados para realizar estas estimaciones se basan principalmente en dos grandes metodologías, una basada en máxima verosimilitud restringida conocida como REML (Thompson, 2005) y otra basada en procedimientos bayesianos (Sorensen y Gianola, 2002).

La primera metodología utiliza la estimación máxima verosímil cuya idea fundamental consiste en tomar como estimación de los parámetros de interés los valores que hagan máxima la probabilidad de obtener la muestra observada, pero requiere el supuesto de normalidad de la variable respuesta (Searle *et al*,1992). Además, se sabe que estos estimadores son sesgados (Searle *et al* 1992), por lo que se propuso la estimación Máxima Verosimilitud Restringida (REML), la cual considera la pérdida de grados de libertad resultante de estimar los efectos fijos. Como respuesta para evitar este sesgo, Patterson y Thompson (1971) dan una descripción general de estos estimadores. Esta metodología continúa siendo la más utilizada entre los investigadores, principalmente por la disponibilidad de *softwares* que implementan este procedimiento.

En la aproximación bayesiana se combina lo que se conoce de los parámetros (distribuciones a priori) con la información que proporciona los datos para obtener la distribución a posteriori, la cual representa la incertidumbre sobre los parámetros después de que se ha tomado en cuenta la información de los datos (Blasco, 2001). El procedimiento computacional estándar, en la estimación bayesiana de parámetros

genéticos, es el método de Montecarlo y cadenas de Markov (MCMC) para obtener muestras de la distribución a posteriori, a partir de diferentes algoritmos, entre ellos el algoritmo de muestreo de Gibbs y el algoritmo de Metropolis-Hastings, los cuales son muy populares y han sido implementados en diferentes softwares. Esta metodología está siendo aplicada en muchas de las áreas de interés en el mejoramiento genético animal. Uno de los primeros estudios fue el trabajado por Wang, Rutledge y Gianola (1993, 1994), quienes aplicaron análisis bayesiano vía muestreo de Gibbs, para estimar parámetros genéticos relacionados con el tamaño de la camada de cerdos ibéricos a partir de un modelo univariado.

Blasco (2001) señala también que ambas escuelas frecuentista y bayesiana, están bien establecidas y no es necesario justificar por que, se prefiere una sobre la otra. Existen *softwares* disponibles para analizar una gran variedad de problemas desde ambos puntos de vista. En general, los algoritmos para REML son más complicados de programar computacionalmente que los métodos vía muestreo de Gibbs (Misztal, 2008), sobre todo cuando los modelos son más complejos e involucran la estimación de varios parámetros y de varios caracteres.

Este trabajo tiene como objetivo principal estimar el parámetro genético: heredabilidad de la producción de leche mediante un modelo animal unicarácter a partir de registros de 3397 lactaciones de ganado lechero de raza Holsteins con información genealógica que comprende 6547 animales descargados desde United State Department of Agriculture, USDA, utilizando Máxima Verosimilitud restringida (REML) e inferencia bayesiana.

Además, tiene como objetivos específicos:

1. Describir ambas metodologías y obtener las estimaciones de los componentes de varianza con cada una de ellas.
2. Realizar el análisis exploratorio de residuales del modelo lineal mixto estimado vía REML, siguiendo la propuesta dada por Singer *et al*. (2013) en función a los tres tipos de residuales (marginal, condicional y efectos aleatorios), pero incorporando la información genealógica o pedigrí en el modelo lineal mixto.

3. Realizar el diagnóstico del procedimiento de simulación del muestreo de Gibbs desde la perspectiva bayesiana.
4. Comparar numéricamente las estimaciones de heredabilidad obtenidas con estos procedimientos.

II. REVISIÓN DE LITERATURA

2.1 Conceptos genéticos básicos relacionados a Heredabilidad

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

- **Genética**

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

- **Medio ambiente**

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

- **Gen**

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

- **ADN**

Acido desoxirribonucleico, molécula que conforma el código genético.

- **Cromosoma**

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

- **Locus**

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

- **Alelo**

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

- **Gameto**

Célula reproductora, masculina o femenina, cuyo núcleo solamente contiene un cromosoma de cada par, y que puede unirse a otro gameto de sexo opuesto, en la fecundación, pero no multiplicarse por sí sola.

- **Genotipo y fenotipo**

El genotipo de un animal representa el gen o grupo de genes responsable por un rasgo o carácter en particular. En un sentido más general, el genotipo describe todo el grupo de genes que un individuo ha heredado. El fenotipo es el valor que toma un rasgo; es decir, es lo que puede ser observado o medido. Por ejemplo, el fenotipo puede ser la producción individual de leche de una vaca, el porcentaje de grasa en la leche o el porcentaje de proteína en la leche.

Existe una diferencia importante entre genotipo y fenotipo. El genotipo es esencialmente una característica fija del organismo; permanece constante a lo largo de la vida del animal y no es modificado por el medio ambiente. Cuando solamente uno o un par de genes son responsables por un rasgo, el genotipo permanece generalmente sin cambios a lo largo de

la vida del animal (ejemplo color de pelo). En este caso, el fenotipo otorga una buena indicación de la composición genética del individuo. Sin embargo, para algunos rasgos, el fenotipo cambia constantemente a lo largo de la vida del individuo como respuesta a factores ambientales. En este caso, el fenotipo no es un indicador directo confiable del genotipo. Esto generalmente se presenta cuando muchos genes se encuentran involucrados en la expresión de un rasgo tal como producción de leche.

2.2 Mejoramiento genético

El ser humano ha observado variabilidad en los rendimientos de animales y plantas de los que ha venido obteniendo recursos para su propio bienestar. Asimismo, ha observado cómo los animales o plantas con mejores rendimientos tendían a transmitir esa superioridad a sus descendientes, por lo que la selección artificial de animales o plantas (individuos) para ser los reproductores de la siguiente generación ha sido de importancia para agricultores y criadores desde antaño. Esta variabilidad se debe al efecto que tienen las condiciones en que los individuos se desarrollan, típicamente descrita como el efecto ambiente, así como también por las características genéticas heredadas de sus ancestros (Gutiérrez, 2010).

Los caracteres o características evaluadas en los animales son variables observables o medibles de los individuos que se quieren optimizar en el proceso productivo con la finalidad de aumentar su rendimiento. Estos caracteres no son afectados de la misma forma por el ambiente. Hay caracteres que tienen mayor determinación genética que ambiental. Por ejemplo, el tamaño y color de pelaje. También, hay caracteres que tiene mayor influencia ambiental que genética, como los caracteres de tipo reproductivo, por ejemplo: producción de leche, porcentaje de proteína, porcentaje de grasa, etc (Gutiérrez, 2010).

El mejoramiento genético de animales o plantas consiste en la aplicación de principios biológicos, estadísticos y económicos, con la finalidad de encontrar estrategias óptimas para aprovechar la variación genética que existe en una especie de animales, en particular para maximizar el mérito genético, habilidad de un determinado progenitor de producir descendencia con rasgos deseables superiores comparado con otros padres (Montaldo y Barría, 1998). Esto implica estimar o predecir el mérito genético, también llamado valor de la cría de los animales, para seleccionar a los futuros progenitores a través de programas de

selección genética. La estimación de este valor genético está sujeta a error por lo que se justifica el uso de procedimientos estadísticos como la estimación de componentes de varianza.

Por tanto, en los programas de selección genética, lo que se busca es maximizar la tasa de crecimiento de algún carácter que se piensa tiene una base genética. Típicamente, los animales con el mérito genético esperado más alto se conservan para ser los padres de la siguiente generación, mientras que aquellos con el mérito genético más bajo son desechados, (Gianola, 2000. En el mejoramiento genético para producción de leche de vacuno se busca incrementar los genes que maximizan la producción dado el medio ambiente (clima, alimentación, manejo, etc.) en el que la vaca expresa su potencial.

Gianola (2000) señala que el mérito puede representarse formalmente mediante una función lineal o no lineal de los valores genéticos para varias características que se consideran importantes desde el punto de vista de generar ganancias económicas o bien de aportar algún beneficio. El componente genético del mérito no puede ser observado, así que tiene que inferirse a partir de las observaciones hechas a los candidatos de selección o en sus parientes. Esto presenta al menos tres problemas: (a) determinar si las características que forman parte de la función de mérito tienen una base genética; (b) obtener métodos razonablemente precisos para inferir el mérito ("evaluación genética") y (c) decidir qué hacer con los animales que tengan las mejores evaluaciones. Los dos primeros son problemas estadísticos, (a) se conoce comúnmente como "estimación de parámetros genéticos" y (b) se conoce como "estimación (predicción) de valores de cría" (mérito genético) conceptualmente inseparables.

Para estimar los parámetros genéticos y el mérito genético se realiza observaciones de los rasgos (o características) de una especie, tales como producción de leche, porcentaje de grasa, porcentaje de proteína, peso al nacimiento, color del pelo, tamaño de camada, etc. Estas características son métricas (medibles) y son generalmente poligénicos. Esto quiere decir determinados por varios genes, que además tienen una fuerte influencia ambiental (Gianola, 2000).

Para el estudio de estos caracteres métricos, es necesario utilizar métodos y modelos estadísticos que permitan separar el fenotipo en efecto ambiente y genotipo. Después de

esto, estimar los componentes de varianza y la heredabilidad asociada al efecto genético; es decir, cuantificar la variación observada que se debe a los genes sin considerar el efecto ambiental (Gutiérrez, 2010).

2.3 Modelo estadístico para la evaluación genética

Fisher (1918) sentó las bases para el modelo infinitesimal. Describió las consecuencias de la herencia mendeliana en el fenotipo (expresión del carácter observado) y planteando el modelo:

Fenotipo (P) = Genotipo (G)+ Ambiente (E)

Esto muestra que el genotipo no se expresa en su totalidad en el fenotipo, sino que se ve modificado por el ambiente. En algunos casos es posible encontrar interacción entre genotipo y ambiente; sin embargo, es muy difícil de tener en cuenta en la valoración genética, por lo que suele incorporarse en la parte no genética del fenotipo (ambiente, E).

Gutiérrez (2010) señala que el genotipo de un individuo es consecuencia de multitud de genes y se descompone en:

- 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 forma siguiente: P=A+D+I+E.

La variabilidad de un carácter en un conjunto de individuos es cuantificada en la varianza, la cual es el material de trabajo de las valoraciones genéticas. Por tanto, la varianza fenotípica puede descomponerse en varianzas de diferente origen. Bajo el supuesto de no interacción entre los componentes anteriores, la varianza fenotípica puede descomponerse de la siguiente manera:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2$$

En esta expresión $\sigma_G^2, \sigma_E^2, \sigma_A^2, \sigma_D^2$ y σ_I^2 son, respectivamente, las varianzas genotípicas, residual, genética aditiva, dominante y epistática.

2.3.1 Heredabilidad

Gutiérrez (2010) explica que la variabilidad de un carácter es imprescindible para llevar a cabo la selección, asimismo, muestra que esta variabilidad presenta varios orígenes. Pero que únicamente es de interés la que tiene base genética. El conocimiento de la proporción de la variabilidad que es de origen genético es un parámetro de mucho interés, es conocido como heredabilidad y tiene dos acepciones:

- **Heredabilidad en sentido amplio.** Definido como la proporción de la variabilidad fenotípica que es de origen genético:

$$H^2 = \frac{\sigma_A^2 + \sigma_D^2 + \sigma_I^2}{\sigma_P^2}$$

- **Heredabilidad en sentido estrecho.** Definido como la proporción de la variabilidad fenotípica que es de origen genético aditivo:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

  Esto permite medir cuánto de la variabilidad fenotípica de un carácter en una población dada, es probable que se transmita a sus descendientes, puesto que se considera que la parte aditiva es la que se hereda de forma directa, ya que los otros componentes (dominancia, epistasis) son interacciones.

La heredabilidad no es herencia. La herencia es la transmisión de un valor fenotípico de un pariente a su descendencia, mientras que la heredabilidad es la trasmisión de la variabilidad fenotípica, dentro de una población, de generación en generación. Es importante señalar que la heredabilidad de un carácter es definida para una población dada en un tiempo dado, además esta cantidad puede variar entre poblaciones y de tiempo en tiempo (Mousseau y Roff, citado por De Villemereuil, 2012). La estimación de este parámetro se realiza como paso previo a la valoración genética.

Gutiérrez (2010) resalta que, aunque la variabilidad 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 0.4). Caracteres relacionados con el tamaño.
- Heredabilidad moderada (de 0.15 a 0.4). 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 la prolificidad.

En general, se puede resumir que los componentes de un modelo genético estadístico son: (a) función matemática (por lo general lineal) que relaciona el o los caracteres medibles con efectos fijos (parámetros de localización) y aleatorios, (b) parámetros de dispersión genética y ambiental, etc. y (c) supuestos sobre la distribución de las observaciones y de los efectos aleatorios (Gianola, 2000).

En programas de mejoramiento genético animal, la estimación de parámetros genéticos a partir de estos modelos estadísticos se ha aplicado tanto dentro de un contexto bayesiano, utilizando principalmente el método de muestreo de Gibbs, como en un contexto frecuentista, considerando principalmente la metodología de modelos lineales mixtos (Blasco, 2001).

2.3 El modelo animal

Cuando se realiza un análisis genético con individuos relacionados por parentesco, el efecto genético entre individuos está correlacionado. Estas correlaciones genéticas aditivas pueden ser derivadas desde el registro de descendencia de los animales (pedigrí) y arregladas en una matriz que recibe el nombre de matriz de relaciones genéticas aditivas.

Vásquez *et al*. (2010) indican que $\mathbf{A}$, matriz de relaciones genéticas aditivas, es una matriz simétrica definida positiva (a menos que haya gemelos idénticos o clones en el pedigrí) de dimensión igual al número de individuos en la pedigrí.

Wright, citado por Mrode (2014), describe a esta matriz con elementos en la diagonal igual a: $a_{ii} = 1 + F_i$ para el animal i, donde F_i es el coeficiente de consanguinidad del animal i, cada elemento de la diagonal representa dos veces la probabilidad que dos gametos tomados al azar del animal i aporten alelos idénticos a su descendencia. Cada elemento fuera de la diagonal de la matriz $\mathbf{A}$ es el numerador del coeficiente de parentesco entre el animal i y el animal j. Así, cuando se multiplica estos elementos por la varianza genética aditiva (σ_u^2), $\mathbf{A}\sigma_u^2$ es la matriz de covarianza entre los valores genéticos de los animales. Luego si u_i es el valor genético para el animal i, su varianza está dada por $v(u_i) = a_{ii}\sigma_u^2 = (1 + F_i)\sigma_u^2$. La matriz de relaciones aditivas $\mathbf{A}$ puede calcularse mediante el método tabular, procedimiento recursivo (Henderson, citado por Mrode 2014).

El modelo animal es un modelo lineal mixto como en (1) con al menos dos componentes aleatorios: $\mathbf{u}$ que corresponde al efecto genético de cada animal evaluado, con una estructura de varianzas y covarianzas: $\mathbf{G} = \mathbf{A}\sigma_u^2$, donde $\mathbf{A}$ es la matriz de relaciones genéticas aditivas que se construye a partir de la información genealógica de los individuos descrita anteriormente, y el componente del error $\mathbf{e}$, cuya estructura de varianzas y covarianzas se asume como: $V(\mathbf{e}) = \mathbf{R} = \mathbf{I}\sigma_e^2$. Además en este modelo se supone independencia entre estos componentes (Mrode 2014).

Elzo (2012) indica que en 1989, se implementó el modelo animal para la evaluación genética de vacas y toros en Estados Unidos. La estructura matemática del modelo animal se basa en las ecuaciones del modelo mixto (2) desarrolladas por Henderson, citado por Elzo (2012) quien señala que las ecuaciones para el modelo animal proporcionan soluciones directas para los individuos que tienen registros de producción en términos de valor genético aditivo, así como también para individuos sin registro de producción, los cuales pueden ser estimados mediante la inclusión de la matriz de parentesco, por lo que el modelo permite valorar animales jóvenes o sin información de producción. Por tanto la valoración genética de un individuo estimada a través del modelo animal dependerá de su propia información productiva, ajustada por efectos fijos (como rebaño, año, época de parto, etc.), y de su información genealógica. La inclusión de estos efectos genéticos aumentará la confiabilidad de las predicciones de los valores genéticos de los individuos

con producción y permitirá predecir los valores genéticos de aquellos individuos que carecen de registros productivos.

2.3.1. Formulación matricial del modelo animal

Henderson (1959) formuló el problema de predicción del mérito genético a través de un modelo de efectos mixtos cuya ecuación es:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

donde:

$\mathbf{y}$: vector de observaciones de dimensión $N \times 1$

$\boldsymbol{\beta}$: vector de dimensión $p \times 1$ correspondientes a los efectos fijos.

$\mathbf{u}$: vector de dimensión $q \times 1$ correspondientes a los efectos aleatorios de poblaciones aleatorias con estructuras de varianzas y covarianzas.

$\mathbf{X}$: matriz diseño conocida de dimensión $N \times p$, relacionado con los elementos de $\boldsymbol{\beta}$.

$\mathbf{Z}$: matriz diseño conocida de dimensión $N \times q$, relacionado con los elementos de $\mathbf{u}$.

$\mathbf{e}$: vector de los efectos aleatorios (error) de dimensión $N \times 1$.

La estructura de los esperados, varianzas y covarianzas de los efectos aleatorios, son:

$E(\mathbf{u}) = \mathbf{0}$,

$E(\mathbf{e}) = \mathbf{0}$,

$V\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$, donde $\mathbf{G}$ y $\mathbf{R}$ son matrices definidas positivas, luego:

$E(\mathbf{y}) = \mathbf{X\beta}$

$V(\mathbf{y}) = \mathbf{ZGZ^T} + \mathbf{R}$

$\text{cov}(\mathbf{y}, \mathbf{u}) = \mathbf{ZG}$

$\text{cov}(\mathbf{y}, \mathbf{e}) = \mathbf{R}$

En el mejoramiento genético de animales, el problema de predicción del valor genético se centra en la "estimación" de los efectos aleatorios que fue desarrollada por Henderson buscando derivar la mejor predicción lineal insesgada óptima. Dícese *lineal*

porque es una función lineal de los datos; *insesgada* en el sentido que el valor promedio de los estimados es igual al valor promedio de la cantidad a ser estimada; *mejor* en el sentido que tiene mínimo error cuadrático medio entre todos los estimadores lineales insesgados y *predictor* para distinguir de los estimadores de efectos fijos, conocida como BLUP (1950-1984) como lo señala Robinson (1991).

Para estimar $\boldsymbol{\beta}$ y $\mathbf{u}$ matemáticamente Henderson (1950) asumió normalidad para $\boldsymbol{\beta}$ y $\mathbf{u}$, y las estructuras de las matrices $\mathbf{G}$ y $\mathbf{R}$ conocidas, luego minimizando la función de densidad conjunta de $\mathbf{y}$ y $\mathbf{u}$ con respecto a $\boldsymbol{\beta}$ y $\mathbf{u}$, obtuvo las siguientes ecuaciones simultáneas llamadas ecuaciones de modelos mixtos - MME, (Robinson, 1991):

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \qquad (2)$$

Sin embargo técnicamente $\mathbf{u}$ no puede estimarse por tratarse de un vector aleatorio. Gianola (2000) señala que la función objetivo, maximizada por Henderson (1950), es una densidad posterior conjunta, bajo un enfoque bayesiano, o bien una verosimilitud *penalizada* o *extendida ad hoc.*

Gianola (2000) señala también que las MME han sido usadas principalmente para la evaluación genética de ganado en todo el mundo. Una dificultad que se encuentran para resolver estas ecuaciones es que el sistema presentado en (2) puede ser de orden de varios millones de ecuaciones en modelos univariados (unicarácter) o multivariados (multicarácter). Otra dificultad adicional que encuentra en el desarrollo de MME es la inversión de $\mathbf{G}$ cuando el orden de $\mathbf{u}$ es grande (miles o millones), en cambio la inversión de $\mathbf{R}$ no es un problema puesto que normalmente en el modelo lineal mixto se asume $\mathbf{R}=\mathbf{I}\sigma_e^2$, por tanto la estimación del sistema (2) se realiza asumiendo homogeneidad e independencia en el error.

Henderson (1950) simplificó el procedimiento de inversión de $\mathbf{G}$ haciendo $\mathbf{G}=\mathbf{G}_0 \otimes \mathbf{A}$, donde $\mathbf{G}_0$ es de orden igual al número de características, $\otimes$ es el producto directo (o producto Kronecker), definido como:

$$\mathbf{A}\otimes\mathbf{B}=\begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1c}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2c}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1}\mathbf{B} & a_{r2}\mathbf{B} & \dots & a_{rc}\mathbf{B} \end{pmatrix}$$, para $\mathbf{A}$ $r\times c$ con elementos a_{ij} y $\mathbf{B}$ una matriz cualquiera y $\mathbf{A}$ es una matriz de "relaciones genéticas aditivas", cuyo orden es igual al número de animales en el pedigrí (información genealógica, la cual refleja las probabilidades de que individuos emparentados porten copias idénticas del mismo alelo). Esta matriz contiene toda la información sobre parentescos y consanguinidad de los individuos. Henderson descubrió que $\mathbf{A}^{-1}$ puede obtenerse directamente a partir de una lista de los progenitores de los animales con sus relaciones de parentesco. Esto permite usar todas las relaciones disponibles en la evaluación genética, lo cual produce inferencias más precisas sobre los valores genéticos y también permite la posibilidad de corregir sesgos (Gianola, 2000).

2.4 El modelo lineal mixto

El modelo lineal animal es un modelo lineal mixto, es decir un modelo en cuyos componentes se encuentran factores de efectos fijos y factores de efectos aleatorios. Searle Casella & McCulloch (1992) definen que un factor puede entenderse como una categorización de los datos observados, que surge con el interés de atribuir a estas categorizaciones la variabilidad de los datos. A las clases individuales de un factor se les llama niveles. Desde el punto de vista frecuentista un factor puede ser fijo o aleatorio.

2.4.1 Factor de efectos fijos

Un factor fijo contiene un conjunto finito de niveles que ocurren en los datos y de los que se tiene, por lo general, especial interés. Según el propósito del estudio, si se desea conocer las diferencias entre niveles específicos de un factor se considera como fijo.

2.4.2 Factor de efectos aleatorios

Un factor aleatorio se refiere a una muestra aleatoria de niveles que son obtenidos, por lo general, de un conjunto infinito de niveles. Si el interés del estudio recae en conocer que

tan grande es la variación entre las diferencias de los niveles de un factor se lo considera aleatorio.

Un modelo que contiene solo factores fijos se denomina modelo de efectos fijos o modelo fijo, un modelo que contienen solo factores aleatorios se llama modelo de efectos aleatorio o modelo aleatorio y un modelo que contiene ambos efectos se le denomina modelo mixto. Blasco (2001) indica que desde el punto de vista bayesiano no es necesaria esta distinción entre factores aleatorios o fijos, puesto que la perspectiva de expresar lo incierto es trazando funciones de densidad que pretenden describir eso desconocido, por tanto todo lo desconocido es considerado aleatorio.

Los modelos mixtos son también llamados modelos multinivel (Bates, 2014), porque los efectos aleatorios agregan niveles de variación a las observaciones que previamente se van incorporando al modelo estadístico común, como son los modelos de regresión, modelos lineales generalizados y modelos de regresión no lineales.

2.4.3 Formulación del modelo mixto

Bates (2010) y Bates (2014) señala que en un modelo lineal, la distribución de $\mathbf{y}$, variable respuesta es un vector aleatorio normal multivariante

$$\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}) \tag{3}$$

donde n es la dimensión del vector respuesta, $\mathbf{X}$ es una matriz diseño de dimensión $n \times p$, los parámetros del modelo son los coeficientes $\boldsymbol{\beta}$ y σ. Un modelo mixto incorpora dos vectores aleatorios, el de la variable respuesta $\mathbf{y}$ y el vector de efectos aleatorios B. En este modelo la distribución condicional de $\mathbf{y}$ dado $B = b$, es tal que:

$$(\mathbf{y} / B = b) \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b, \sigma^2\mathbf{I}) \tag{4}$$

donde $\mathbf{Z}$ es matriz diseño para un vector de dimensión q de variables aleatorias B cuyos valores son fijados a b. La distribución incondicional de B es también normal multivariante de la forma:

$$B \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \tag{5}$$

$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ es la matriz de varianzas y covarianzas y debe ser semidefinida positiva. Por razones computacionales es conveniente expresar el modelo en términos del factor de covarianza relativo $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, el cual es una matriz de dimensión q que depende de los parámetros componentes de covarianza $\boldsymbol{\theta}$, y satisface:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}} \tag{6}$$

Por lo que $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ depende tanto de $\boldsymbol{\theta}$ como de σ.

Desde una perspectiva computacional, el modelo es reformulado tal que $\boldsymbol{\theta}$ aparece solo en la distribución condicional para el vector respuesta dado los efectos aleatorios. La cual permite trabajar con matrices de covarianza singulares pues no requieren la evaluación de $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ sino que son basado en $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, esta reformulación es hecha definiendo una variable aleatoria con distribución esférica normal:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{7}$$

donde

$$B = \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \mathbf{u} \tag{8}$$

Por lo que el modelo de la ecuación (4) queda reformulado como:

$$(\mathbf{y} / \mathbf{u} = u) \sim N(\mathbf{X}\beta + \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u, \sigma^2 \mathbf{I}) \tag{9}$$

El vector $E(\mathbf{y} / \mathbf{u} = u) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u$ es interpretado como media condicional (o moda) de las variable respuesta dado valores de $\mathbf{u} = u$.

Debido a que se observa $\mathbf{y}$ y no b o u la distribución condicional de interés para propósito de inferencia estadística es $(\mathbf{u} / \mathbf{y} = y)$. Esta distribución condicional es siempre una distribución continua con densidad de probabilidad condicional $f_{\mathbf{u}/\mathbf{y}}(u / y)$.

Se puede evaluar $f_{\mathbf{u}/\mathbf{y}}(u / y)$ como el producto de densidad incondicional $f_{\mathbf{u}}(u)$ y la densidad condicional, $f_{\mathbf{y}/\mathbf{u}}(y / u)$, esta densidad condicional no normalizada se escribe como:

$$h(u / y, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma) = f_{\mathbf{y}/\mathbf{u}}(y / u, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)\, f_{\mathbf{u}}(u / \sigma) \tag{10}$$

Se dice que h es la densidad condicional no normalizada porque es proporcional a $h(u / y, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)$. Para obtener la densidad condicional se debe normalizar h dividiendo por el valor de la integral:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma / y) = \int_{\mathbb{R}^q} h(u / y, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma) du \tag{11}$$

El valor de esta integral $L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma / y)$ es exactamente la verosimilitud de los parámetros $\boldsymbol{\theta},\boldsymbol{\beta},$ y σ, dado los valores observados y, los estimadores máximo verosímiles (MV) de estos parámetros son los valores que maximizan L.

La integral definida como en la verosimilitud dada en (11) tiene una forma cerrada en un modelo lineal mixto, esta integral puede ser evaluada usando factor de descomposición de Choleski, $\mathbf{L}_{\boldsymbol{\theta}}$, y la moda condicional.

$$\tilde{u} = \arg\max_u h(u / y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma) = \arg\max_u f_{\mathbf{y}/\mathbf{u}}(y / u,\boldsymbol{\theta},\boldsymbol{\beta},\sigma) f_{\mathbf{u}}(u) \tag{12}$$

La notación $\arg\max_u$ significa que $\tilde{u}$ es el valor de u que maximiza la expresión.

En general la moda de una distribución continua es el valor de la variable aleatoria que maximiza la densidad. El valor de $\tilde{u}$ es llamado la moda condicional de u dado $\mathbf{y} = y$, porque $\tilde{u}$ maximiza la densidad condicional de $\mathbf{u}$ dado $\mathbf{y} = y$. La ubicación del máximo puede ser determinado maximizando la densidad condicional normalizada porque h es solo una constante múltiplo de $f_{\mathbf{u}/\mathbf{y}}(u / y)$. Luego, en un modelo lineal mixto las densidades descritas en la última parte de (12) son:

$$f_{\mathbf{y}/\mathbf{u}}(y / u,\boldsymbol{\theta},\boldsymbol{\beta},\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp - \left(\frac{\left\| y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u \right\|^2}{2\sigma^2} \right) \tag{13}$$

$$f_{\mathbf{u}}(u / \sigma) = \frac{1}{(2\sigma^2)^{q/2}} \exp - \left(\frac{\left\| u \right\|^2}{2\sigma^2} \right) \tag{14}$$

Con producto:

$$h(u / y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma) = \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp - \left(\frac{\left\| y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u \right\|^2 + \left\| u \right\|^2}{2\sigma^2} \right) \tag{15}$$

Considerado la devianza:

$$-2\log(h(u / y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma)) = (n+q)\log(2\pi\sigma^2) + \left(\frac{\left\| y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u \right\|^2 + \left\| u \right\|^2}{\sigma^2} \right) \tag{16}$$

Como (16) describe la densidad negativa del logaritmo, $\tilde{u}$ será el valor de u que maximice la expresión del lado derecho de (16).

La única parte del lado derecho de (16) que depende de u es el numerador del segundo término, así:

$$\tilde{u} = \arg\min_u \left\| y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} u \right\|^2 + \left\| u \right\|^2 \tag{17}$$

La expresión a ser minimizada, llamada la función objetivo, es descrita como una suma de cuadrados residuales penalizadas (PRSS siglas en inglés) y el minimizador $\tilde{u}$ es llamado la solución mínimo cuadrado penalizado (PLS, siglas en inglés). Estos son llamados de esta forma porque el primer término de la función objetivo, $\|y - \mathbf{X\beta} - \mathbf{Z\Lambda_\theta} u\|^2$, es una suma de cuadrados residual y el segundo término $\|u\|^2$, es una penalidad de la longitud $\|u\|$, de u. Valores grandes de u (es decir longitudes grandes como vectores) incurren a una alta penalidad.

El criterio PRSS determina la moda condicional balanceando la fidelidad a los datos observados (es decir producir pequeño residual de suma de cuadrados) frente a la simplicidad del modelo (pequeño $\|u\|$), a este tipo de criterio se refiere como un objetivo suavizado, es decir que busca suavizar la respuesta ajustada reduciendo la complejidad del modelo mientras se busca también una razonable fidelidad de los datos.

Para evaluar la verosimilitud se considera al criterio PRSS como una función de parámetros, dado los datos, y su valor mínimo se escribe como:

$$r^2_{\boldsymbol{\theta},\boldsymbol{\beta}} = \arg\min_u \|y - \mathbf{X\beta} - \mathbf{Z\Lambda_\theta} u\|^2 + \|u\|^2 \tag{18}$$

Es posible observar que $\tilde{u}$ puede ser determinado con cálculos directos (no iterativamente), de hecho se puede minimizar el criterio PRSS con respecto a u y $\boldsymbol{\beta}$ simultáneamente sin iterar, por lo que el valor mínimo es:

$$r^2_{\boldsymbol{\theta}} = \arg\min_{u,\boldsymbol{\beta}} \|y - \mathbf{X\beta} - \mathbf{Z\Lambda_\theta} u\|^2 + \|u\|^2 \tag{19}$$

El valor mínimo de $\boldsymbol{\beta}$ es llamado el estimado condicional de $\boldsymbol{\beta}$ dado $\boldsymbol{\theta}$.

El problema general de maximizar $L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma / y)$ con respecto a $\boldsymbol{\theta}, \boldsymbol{\beta}$, y σ puede ser enorme, pues cada evaluación de esta función involucra integrales de dimensión alta, y porque la dimensión de $\boldsymbol{\beta}$ puede ser grande. Sin embrago este problema de optimización general puede ser dividido en subproblemas manejables. Dado un valor de $\boldsymbol{\theta}$ se puede determinar la moda condicional, $\tilde{u}(\boldsymbol{\theta})$ de u y la estimación condicional $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ simultáneamente usando mínimos cuadrados penalizados iterativos re-ponderados (PIRLS siglas en inglés). La moda condicional y la estimación condicional son definidas como:

$$\begin{bmatrix} \tilde{u}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \arg\max_{u,\boldsymbol{\beta}} h(u / y, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma) \tag{20}$$

Es común en problemas de optimización re-expresar la densidad condicional sobre la devianza, la cual es el negativo de dos veces el logaritmo de la densidad, por lo que la optimización se convierte en:

$$\begin{bmatrix} \tilde{u}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \arg\max_{u,\boldsymbol{\beta}} -2\log(h(u/y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma)) \tag{21}$$

Este problema de optimización puede solucionarse muy eficientemente usando PIRLS, es más para modelos lineales mixtos $\tilde{u}(\boldsymbol{\theta})$ y $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ pueden ser directamente evaluados

La serie de expansión de Taylor de segundo orden de $-2\log(h)$ a $\tilde{u}(\boldsymbol{\theta})$ y $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ proporciona la aproximación de Laplace para el perfil de la devianza. Optimizando esta función con respecto a $\boldsymbol{\theta}$ proporciona los estimadores MV de $\boldsymbol{\theta}$, desde el cual son derivados los estimados MV de $\boldsymbol{\beta}$, y σ.

2.4.4 Métodos de estimación para Modelos lineales mixtos

Bates (2014) indica que los métodos en modelos lineales mixtos que resultan en un problema de mínimos cuadrados penalizados llegan a (16), donde la función de discrepancia es:

$$d(u/y,\boldsymbol{\theta},\boldsymbol{\beta}) = \left\|y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}u\right\|^2 + \left\|u\right\|^2 \tag{22}$$

Una forma de expresar (20) como un problema de mínimos cuadrados penalizados es incorporando la penalidad como "pseudo datos" en un problema de mínimos cuadrados ordinarios, es decir extendiendo $y - \mathbf{X}\boldsymbol{\beta}$ con q respuestas cero cuando se minimiza con respecto a u, por tanto (22) puede escribirse como una suma de cuadrados residual que es lineal en ambos u y $\boldsymbol{\beta}$.

$$d(u/y,\boldsymbol{\theta},\boldsymbol{\beta}) = \left\| \begin{bmatrix} y \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} & \mathbf{X} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} u \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 \tag{23}$$

En (23) se observa que la discrepancia es una forma cuadrática en ambos u y $\boldsymbol{\beta}$.

Además como $\mathbf{X}$ es una matriz de rango columna completo, la discrepancia es una forma cuadrática definida positiva en u y $\boldsymbol{\beta}$ que es minimizado en $\tilde{u}(\boldsymbol{\theta})$ y $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ satisfaciendo:

$$\begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} + \mathbf{I} & \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^T\mathbf{Z}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}} & \mathbf{X}^T\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{u}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^T\mathbf{Z}^T y \\ \mathbf{X}^T y \end{bmatrix} \tag{24}$$

El sistema de ecuaciones (24) es simétrico, definido positivo y dispersa (muchos ceros). Una forma de determinar una solución es aplicando la descomposición de Cholesky.

Así, si una matriz $\mathbf{A}$ es simétrica, definida positiva y dispersa, entonces el factor disperso de Cholesky con la matriz de permutación $\mathbf{P}$ es la matriz triangular inferior $\mathbf{L}$ tal que:

$$\mathbf{L}\mathbf{L}^T = \mathbf{P}\mathbf{A}\mathbf{P}^T \tag{25}$$

Esta matriz de permutación $\mathbf{P}$ es determinada desde el patrón de no ceros en $\mathbf{A}$, pero no dependen de estos valores particulares, sin embargo tienen impacto sobre el número de no ceros en $\mathbf{L}$ y por tanto sobre la velocidad con la cual $\mathbf{L}$ puede ser calculado desde $\mathbf{A}$.

En la mayoría de modelos lineales mixtos la matriz $\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ es dispersa mientras que $\mathbf{X}$ es densa o casi densa, por tanto la matriz $\mathbf{P}$ puede ser restringida a la forma:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{X}} \end{bmatrix} \tag{26}$$

sin perder eficiencia, de hecho en mucho de los casos $\mathbf{P}_{\mathbf{X}} = \mathbf{I}$

Asumiendo que la matriz de permutación que se requiere es de la forma (26) entonces se puede escribir la factorización de Choleski para el sistema definido positivo del sistema (24) como:

$$\begin{bmatrix} \mathbf{L}_{\mathbf{Z}} & \mathbf{0} \\ \mathbf{L}_{\mathbf{XZ}} & \mathbf{L}_{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{\mathbf{Z}} & \mathbf{0} \\ \mathbf{L}_{\mathbf{XZ}} & \mathbf{L}_{\mathbf{X}} \end{bmatrix}^T = \begin{bmatrix} \mathbf{P}_{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^T \mathbf{Z}^T \mathbf{Z} \boldsymbol{\Lambda}_{\boldsymbol{\theta}} + \mathbf{I} & \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^T \mathbf{Z}^T \mathbf{X} \\ \mathbf{X}^T \mathbf{Z} \boldsymbol{\Lambda}_{\boldsymbol{\theta}} & \mathbf{X}^T \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{X}} \end{bmatrix}^T \tag{27}$$

La discrepancia puede ser escrito en su forma canónica como:

$$d(u/y, \boldsymbol{\theta}, \boldsymbol{\beta}) = \tilde{d}(y, \boldsymbol{\theta}) + \left\| \begin{bmatrix} \mathbf{L}_{\mathbf{Z}}^T & \mathbf{L}_{\mathbf{XZ}}^T \\ \mathbf{0} & \mathbf{L}_{\mathbf{X}}^T \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathbf{Z}}(u - \tilde{u}) \\ \mathbf{P}_{\mathbf{X}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \end{bmatrix}^T \right\|^2 \tag{28}$$

donde $$\tilde{d}(y, \boldsymbol{\theta}) = d(\tilde{u}(\boldsymbol{\theta})/y, \boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})) \tag{29}$$

es la discrepancia mínima, dado $\boldsymbol{\theta}$.

2.4.5 Verosimilitud perfilada para modelos lineales mixtos

Sustituyendo (28) en (16) proporciona la densidad condicional no normalizada sobre la devianza escalada como:

$$
\begin{aligned}
&-2\log(h(u/y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma)) \\
&= (n+q)\log(2\pi\sigma^2) + \frac{\tilde{d}(y,\boldsymbol{\theta}) + \left\| \begin{bmatrix} \mathbf{L}_{\mathbf{Z}}^T & \mathbf{L}_{\mathbf{XZ}}^T \\ \mathbf{0} & \mathbf{L}_{\mathbf{X}}^T \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathbf{Z}}(u-\tilde{u}) \\ \mathbf{P}_{\mathbf{X}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \end{bmatrix}^T \right\|^2}{\sigma^2}
\end{aligned} \tag{30}
$$

La integral de la forma cuadrática sobre la devianza escalada de (30) proporciona la log-verosimilitud de $l(\boldsymbol{\theta},\boldsymbol{\beta},\sigma/y)$ como:

$$
\begin{aligned}
&-2l(\boldsymbol{\theta},\boldsymbol{\beta},\sigma/y) = -2\log(L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma/y)) \\
&= n\log(2\pi\sigma^2) + \log(|\mathbf{L}_{\mathbf{Z}}|^2) + \frac{\tilde{d}(y,\boldsymbol{\theta}) + \left\| \mathbf{L}_{\mathbf{X}}^T \mathbf{P}_{\mathbf{X}} (\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T \right\|^2}{\sigma^2}
\end{aligned} \tag{31}
$$

desde el cual se observa que el estimado de $\boldsymbol{\beta}$, dado $\boldsymbol{\theta}$, es $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ y el estimado condicional de σ, dado $\boldsymbol{\theta}$, es:

$$
\tilde{\sigma}^2(\boldsymbol{\theta}) = \frac{\tilde{d}(\boldsymbol{\theta}/y)}{n} \tag{32}
$$

Sustituyendo este estimador condicional en (31) se obtiene la verosimilitud perfilada $\tilde{L}(\boldsymbol{\theta}/y)$, como:

$$
-2\tilde{l}(\boldsymbol{\theta}/y) = \log(|\mathbf{L}_{\mathbf{Z}}|^2) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(y,\boldsymbol{\theta})}{n}\right)\right) \tag{33}
$$

El estimador máximo verosímil $\boldsymbol{\theta}$ puede ser expresado como:

$$
\hat{\boldsymbol{\theta}}_{\mathbf{L}} = \arg\min(-2\tilde{l}(\boldsymbol{\theta}/y)) \tag{34}
$$

desde el cual los estimadores máximo verosímiles de σ y $\boldsymbol{\beta}$ son evaluados como:

$$
\hat{\sigma}^2_{\mathbf{L}} = \frac{\tilde{d}(\hat{\boldsymbol{\theta}}_{\mathbf{L}}/y)}{n} \tag{35}
$$

$$
\hat{\boldsymbol{\beta}}_{\mathbf{L}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_{\mathbf{L}}) \tag{36}
$$

Normalmente la optimización de la verosimilitud perfilada (33) es de dimensión pequeña.

2.4.6 El criterio de Máxima Verosimilitud Restringida – REML

El criterio REML para determinar $\hat{\boldsymbol{\theta}}$ y $\hat{\sigma}^2$ en un modelo lineal mixto puede ser expresado como:

$$c_R(\boldsymbol{\theta},\sigma / y) = -2\log\int_{\mathbb{R}^p} L(u / y,\boldsymbol{\theta},\boldsymbol{\beta},\sigma)d\boldsymbol{\beta} \tag{37}$$

sobre la devianza escalada. El estimador REML de $\hat{\boldsymbol{\theta}}_R$ y $\hat{\sigma}_R^2$ minimiza $c_R(\boldsymbol{\theta},\sigma / y)$.

El criterio REML perfilado, función de $\boldsymbol{\theta}$, es:

$$\tilde{c}_R(\boldsymbol{\theta} / y) = \log(\left|\mathbf{L}_{\mathbf{Z}}\right|^2 \left|\mathbf{L}_{\mathbf{X}}\right|^2) + (n-p)\left(1+\log\left(\frac{2\pi\tilde{d}(y,\boldsymbol{\theta})}{n-p}\right)\right) \tag{38}$$

y el estimador REML de $\boldsymbol{\theta}$ es:

$$\hat{\boldsymbol{\theta}}_R = \arg\min_{\boldsymbol{\theta}} \tilde{c}_R(\boldsymbol{\theta} / y) \tag{39}$$

El estimador REML de σ^2 es:

$$\hat{\sigma}_R^2 = \frac{\tilde{d}(\hat{\boldsymbol{\theta}}_R / y)}{n-p} \tag{40}$$

Como el criterio REML no depende de $\boldsymbol{\beta}$ es usual utilizar $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_{\mathbf{L}})$ como estimador REML de $\boldsymbol{\beta}$.

El criterio REML puede ser evaluado desde una descomposición de Cholesky dispersa como en (27), pero sin el requerimiento que la permutación puede ser aplicada a las columnas de $\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ separadamente de las columnas de $\mathbf{X}$. Es decir se puede usar la matriz de permutación no como en (26) sino con permutaciones separadas representadas por $\mathbf{P}_{\mathbf{X}}$ y $\mathbf{P}_{\mathbf{Z}}$. Esto es útil principalmente donde ambos $\mathbf{Z}$ y $\mathbf{X}$ son grandes y dispersas (Bates, 2014).

2.4.7 Comparación con formulaciones previas

El problema PLS (mínimo cuadrados penalizados) descrito en 2.4.1 fue comparado con las ecuaciones del modelo mixto de Henderson MME descritas en (2) por Bates y DebRoy (2004).

Las MME pueden ser expresadas como:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X}/\sigma^2 & \mathbf{X}'\mathbf{Z}/\sigma^2 \\ \mathbf{Z}'\mathbf{X}/\sigma^2 & \mathbf{Z}'\mathbf{Z}/\sigma^2 + \boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{u}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}'y \\ \mathbf{Z}'y \end{pmatrix} \tag{41}$$

Bates y DebRoy (2004) modificaron las ecuaciones PLS a las siguientes ecuaciones:

$$\begin{pmatrix} \mathbf{Z}'\mathbf{Z} + \boldsymbol{\Omega}^{-1} & \mathbf{Z}'\mathbf{X} \\ \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{X} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \\ \tilde{u}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbf{Z}'y \\ \mathbf{X}'y \end{pmatrix} \tag{42}$$

donde $\boldsymbol{\Omega}_{\boldsymbol{\theta}} = \left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\right)^{-1} = \sigma^2\boldsymbol{\Sigma}^{-1}$ es la matriz de precisión relativa para un valor de $\boldsymbol{\theta}$. Ellos también mostraron que la log-verosimilitud perfilada puede ser expresada como:

$$-2\tilde{l}(\boldsymbol{\theta}) = \log\left(\left|\mathbf{Z}^T\mathbf{Z} + \boldsymbol{\Omega}\right|\right) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(y,\boldsymbol{\theta})}{n}\right)\right) \tag{43}$$

Bates *et al* (2014) señalan que la diferencia entre la ecuación (42) y (43) es el orden de los bloques en el sistema de matrices. El problema PLS puede ser solucionado usando el factor de Cholesky del sistema de matrices con otro orden. La ventaja de usar la ecuación (43) es que permite la evaluación de log-verosimilitud perfilada. El principal cambio de la formulación (43) a la formulación más reciente es el uso del factor de covarianza relativa, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, en lugar de la matriz de precisión relativa $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ y solucionar para la media $\mathbf{u}$ dado $\mathbf{y} = y$ en vez de la media para B dado $\mathbf{y} = y$. Este cambio mejora la estabilidad, puesto que la solución del problema PLS descrito en 2.4.1 está bien definida cuando $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ es singular en contraste con la formulación (43) que no puede ser usada en estos casos por que $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ no existe.

Es importante considerar $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ ser singular puesto que en la práctica puede ocurrir que los estimados de los parámetros de $\boldsymbol{\theta}$ producen una matriz singular. Aun así si estos parámetros estimados no corresponden a una matriz singular $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, es necesario evaluar el criterio de estimación para posibles situaciones que se puedan presentar en el proceso de optimización numérica.

2.4.8 Prueba de la razón de verosimilitudes

Hartley y Rao, citado por Jiang (2007), señalan que esta prueba es bien conocida y fue desarrollada en el contexto de modelos lineales mixtos por primera vez en 1967.

Gilmour *et al*. (2009) indican que es un método general para comparar el ajuste de los modelos jerárquicos (anidados) ajustados vía REML para parámetros aleatorios (componentes de varianza).

Esta prueba compara dos modelos jerárquicos, es decir cuando un modelo 1 restringido o reducido con k_1 parámetros es anidado dentro de otro modelo 2 no restringido con k_2 parámetros. La hipótesis nula de esta prueba es que el modelo 1, restringido es estadísticamente mejor que el modelo 2, no restringido.

Bates y Pinheiro (2000) resumen el procedimiento indicando que un modelo estadístico se dice que es anidado dentro del otro si este representa un caso particular del otro modelo. Así entonces si l_{R2} es el log-verosímil de un modelo general o modelo 2, y l_{R1} es el log-verosímil de un modelo restringido (o reducido) o modelo 1, se tiene entonces que $l_{R2} > l_{R1}$, por tanto el estadístico de prueba de la razón de verosimilitudes es:

$$D = 2\log(l_{R2} / l_{R1}) = 2(\log l_{R2} - \log l_{R1}) \tag{44}$$

el cual es positivo. Si k_i es el número de parámetros a ser estimados en el modelo i, entonces la distribución asintótica de D bajo la hipótesis nula que el modelo restringido es adecuado en comparación con el modelo general, se ajusta a una distribución χ^2 con $k_2 - k_1$ grados de libertad. Este prueba es aplicada en modelos que tienen la misma estructura de efectos fijos y solo cambian en la estructura de efectos aleatorios.

Para comparar dos (o más modelos) no jerárquicos se puede utilizar el Akaike Information Criteria (AIC) o el Bayesian Information Criteria (BIC) para cada modelo. Estos están dados por:

$$AIC = -2\log(l_{Ri}) + 2t_i \tag{45}$$

$$BIC = -2\log(l_{Ri}) + t_i \log \nu \tag{46}$$

Donde t_i es el número de parámetros de varianza en el modelo i y $\nu = n - p$ es los grados de libertad del residual. AIC y BIC son calculados para cada modelo y se prefiere el modelo con valores más pequeños de estos indicadores.

2.4.9 Prueba de hipótesis para efectos fijos

Bates y Pinheiro (2000) no recomiendan utilizar la prueba de razón de verosimilitudes para probar significancia de los efectos fijo porque suele ser "anticonservador" haciendo que la estimación del estadístico de prueba D no sea muy bueno. Una forma de ejecutar una prueba de hipótesis que involucre efectos fijos es utilizar los estimados REML de los parámetros de efectos aleatorios, varianzas y/o covarianzas, en las pruebas convencionales F y t de un modelo de regresión, sin embargo los autores indican que estás pruebas son sólo aproximaciones y requiere de muestras grandes.

2.5 Diagnóstico del modelo: análisis de residuales

Los residuales son frecuentemente usados para evaluar la validez de los supuestos de un modelo mixto como homocedasticidad, linealidad, normalidad, presencia de outliers, etc. Hilden-Minton (1995) extendió el concepto de residual de un modelo lineal a un modelo lineal mixto, definiendo tres tipos de residuales que Nobre y Singer (2007) resumieron y los cuales se describen a continuación:

1. Residuales marginales: $\hat{\boldsymbol{\xi}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ que predice el error marginal
2. Residuales condicionales: $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{b}$ que predice el error condicional
3. El BLUP $\mathbf{Z}\hat{b}$ que predice el efecto aleatorio

Hilden-Minton (1995) define a un residual confundido para un tipo específico de error si este depende de otros errores además del que supuestamente está prediciendo, en particular encontró que los residuales condicionales y BLUP están confundidos, por lo que $\hat{\mathbf{e}}$ no es adecuado para evaluar normalidad de $\mathbf{e}$ cuando b es gravemente no normal, $\hat{\mathbf{e}}$ puede no presentar un comportamiento normal aun cuando $\mathbf{e}$ lo es.

Los diferentes usos para los tres tipos de residuales son resumidos por Singer *et al*. (2013), quienes lo adaptaron de Nobre y Singer (2007) y son presentados en el Cuadro siguiente:

Cuadro1. Usos de residuales para propósitos de diagnóstico

Diagnóstico para	Tipo de residual	Gráfico
Linealidad de efectos fijos	Marginal	$\hat{\boldsymbol{\xi}}_{ij}^{*}$ vs valores fijos de las variables explicativas
Presencia de observaciones atípicas	Marginal	$\hat{\boldsymbol{\xi}}_{ij}^{*}$ vs índices de las observaciones.
Matriz de covarianzas dentro de las unidades	Marginal	$\mathbf{V}_i^{*}$ vs índices de unidades
Presencia de observaciones atípicas	Condicional	$\hat{\mathbf{e}}_{ij}^{*}$ vs índices de las observaciones
Homocedasticidad de errores condicionales	Condicional	$\hat{\mathbf{e}}_{ij}^{*}$ vs valores ajustados
Normalidad de errores condicionales	Condicional	QQ plot gaussiano para $\mathbf{c}_k^T\hat{\mathbf{e}}_{ij}^{*}$
Presencia de sujetos atípicos	Efectos aleatorios	$\mathbf{M}_i$ vs índices de unidades
Normalidad de los efectos aleatorios	Efectos aleatorios	χ_q^2 QQ plot para $\mathbf{M}_i$

Lesaffre y Verbeke, citado por Singer *et al*. (2013), comentaron que cuando la estructura dentro de las unidades es adecuada, $\mathbf{V}_i = \left\| I_{m_i} - \mathrm{R}_i\mathrm{R}_i^T \right\|^2$, donde $\mathrm{R}_i = \boldsymbol{\Omega}_i^{-1/2}\hat{\boldsymbol{\xi}}_i$ con $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\theta) = \mathbf{Z}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{Z}^T\sigma_{\boldsymbol{\epsilon}}^2\mathbf{I} = \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{Z}^T\sigma^2 + \sigma_{\boldsymbol{\epsilon}}^2\mathbf{I}$ debe ser cercana a cero. Unidades con valores grandes de $\mathbf{V}_i$ indicaría que la estructura de covarianza puede no ser adecuada para dichas observaciones. Singer *et al.* (2013) recomienda reemplazar R_i en $\mathbf{V}_i$ con el residual marginal estandarizado $\hat{\boldsymbol{\xi}}_i^{*} = \left[V(\hat{\boldsymbol{\xi}}_i)\right]^{-1/2}\hat{\boldsymbol{\xi}}_i$, donde $\hat{\boldsymbol{\xi}}_i$ corresponden al elemento de la diagonal $\boldsymbol{\Omega} - \mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$ asociado con la i-ésima unidad. Además recomendaron utilizar $\mathbf{V}_i^{*} = \sqrt{\mathbf{V}_i / m_i}$ como una medida estandarizada de adecuación de la estructura de covarianza dentro de las unidades.

Para evaluar la linealidad de los efectos mixtos *Singer et al.* (2013) sugieren graficar los residuales marginales estandarizados dados por $\hat{\boldsymbol{\xi}}_{ij}^{*} = \hat{\boldsymbol{\xi}}_{ij} / diag[V(\hat{\boldsymbol{\xi}}_{ij})^{1/2}]$, donde $diag[V(\hat{\boldsymbol{\xi}}_{ij})]$ es el j-ésimo elemento de la diagonal principal versus los valores de cada variable exploratoria como también versus los valores ajustados.

Nobre y Singer (2007) observaron que los residuales condicionales pueden tener varianzas diferentes, por lo que sugirieron graficar los residuales estandarizados

condicionales $\hat{\mathbf{e}}_{ij}^{*} = \hat{\mathbf{e}}_{ij} / diag(\mathbf{Q})^{1/2}$, donde $\mathbf{Q} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}^{T}\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Omega}^{-1}$ versus los valores ajustados para chequear homocedasticidad de los errores condicionales o versus índice de observaciones para chequear observaciones atípicas.

Hilden-Minton (1995) resaltó que la habilidad de chequear normalidad de los errores condicionales se incrementa cuando se minimiza la fracción de confundido para los residuales condicionales, el abogó entonces por el uso de residuales mínimos confundidos, es decir una transformación lineal de los residuales condicionales que minimizan la fracción de confundido. Los residuales mínimos confundidos son dados por:

$$\mathbf{c}_k^{T}\hat{\mathbf{e}}_{ij}^{*} = \lambda^{-1/2} l_k^{T}\hat{\mathbf{e}} = \lambda^{-1/2} l_k^{T} y \qquad k = 1,..., N - p$$

donde $1 \geq \lambda_1 \geq ... \geq \lambda_{N-p}$ son valores ordenandos de $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ obtenidos de la descomposición del valor singular $\mathbf{Q} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^{T}$, $\mathbf{L}^{T}\mathbf{L} = \mathbf{I}$, y l_k representa la k-ésima columna de $\mathbf{L}$. Los residuales mínimos confundidos estandarizados $\mathbf{c}_k^{T}\hat{\mathbf{e}}_{ij}^{*}$ pueden ser obtenidos dividiendo $\mathbf{c}_k^{T}\hat{\mathbf{e}}_{ij}$ por la raíz cuadrada de los elementos correspondiente en $\mathbf{CQC}^{T}$ donde $\mathbf{C} = (\mathbf{c}_1,..., \mathbf{c}_{N-p})^{T}$. El gráfico QQ-plot de los residuales mínimos confundidos estandarizados, $\mathbf{c}_k^{T}\hat{\mathbf{e}}_{ij}^{*}$, se emplea para chequear normalidad.

Cuando no hay y efectos confundidos y los efectos aleatorios siguen una distribución q-dimensional gaussiana, $\mathbf{M}_i = \hat{b}_i^{T}\left\{V\left[\hat{b}_i - b_i\right]\right\}\hat{b}_i$, distancia de Mahalanobis entre $\hat{b}_i$ y $E(\hat{b}_i) = 0$, debería tener una distribución chi-cuadrada con *q* grados de libertad, por lo que Nobre y Singer (2007) sugieren utilizar la gráfica QQ chi-cuadrada para $\mathbf{M}_i$ para verificar si los efectos aleatorios tienen una distribución gaussiana, asimismo $\mathbf{M}_i$ puede ser empleado para detectar valores atípicos.

2.6 Estimación bayesiana

Para León (2004), la estimación REML presenta algunas limitaciones, como estimar componentes de varianza cuando desconocemos la media de la distribución, o cuando la distribución de los estimadores no es de forma conocida o solo lo es

asintóticamente; es entonces que bajo estas condiciones los estimadores REML no siempre son fáciles de deducir y cuando los modelos son más complejos, es decir involucran estimación de muchos parámetros relacionados entre sí con una gran cantidad de información, la tarea se vuelve más complicada. Por el contrario, el enfoque bayesiano proporciona un marco de trabajo más flexible y general que con la incorporación de las técnicas de muestreo de Monte Carlo, la estimación suele ser más fáciles de obtener computacionalmente que con las técnicas máximo verosímiles.

Blasco (2001) proporciona una breve sinopsis de cómo los métodos bayesianos fueron introducidos, principalmente por Gianola (2000), en el mejoramiento genético animal en el contexto de predicción de características umbrales (variables categóricas). Sin embargo inicialmente no fueron muy utilizados, pese a la flexibilidad y potencia de estos métodos, porque presentaban problemas computacionales debido a las múltiples integrales que tenían que ser resueltas para obtener las distribuciones marginales posteriores. No fue sino hasta que cuando Wang *et al* (1994) introdujeron los métodos de Montecarlo y Cadenas de Markov (MCMC) para estimar las distribuciones marginales posteriores, que estos métodos empezaron a aplicarse con más frecuencia.

La idea base de la teoría bayesiana consiste en considerar que tanto los parámetros como los efectos aleatorios y los datos tienen distribuciones asociadas. De esta forma, dado los datos se busca describir la incertidumbre de los parámetros de interés usando probabilidades, por lo que cualquier información previa que posea sobre la probabilidad de los valores de los parámetros suele introducirse en el proceso de estimación y aumentar la calidad de información disponible y, por tanto, la precisión de las estimaciones.

Así como la estimación REML hace uso de la función de verosimilitud o funciones que describen la probabilidad de observar unos valores de la variable medida, dados los parámetros desconocidos; la inferencia bayesiana hace uso de la distribución a posteriori obtenida a partir de la función de verosimilitud, que describe la información contenida en los datos con respecto al parámetro de interés, y de la distribución a priori que se asigna a los parámetros.

La asignación de distribuciones a priori es una de los aspectos más discutidos al enfoque bayesiano debido a la dificultad de cuantificar la información previa perteneciente a un parámetro (Blasco 2001), por lo que en muchas de las aplicaciones bayesianas, en la estimación de parámetros genéticos, la información previa que se utiliza como a priori a menudo no está bien explicada. (Thompson, 2005).

2.6.1 Teorema de Bayes

El teorema de Bayes proporciona la expresión básica para calcular la distribución a posteriori de los parámetros, $\boldsymbol{\theta}$, sobre los que se quiere hacer la inferencia, dada la información observada, $\mathbf{y}$:

$$p(\boldsymbol{\theta}/\mathbf{y}) = \frac{f(\mathbf{y}/\boldsymbol{\theta})g(\boldsymbol{\theta})}{f(\mathbf{y})} \tag{47}$$

Donde:

$f(\mathbf{y}/\boldsymbol{\theta})$: es la verosimilitud asociada a los datos, dados los parámetros de interés.

$g(\boldsymbol{\theta})$: es la información a priori sobre dichos parámetros.

$f(\mathbf{y})$: es la información marginal de los datos para cualquier valor de los parámetros.

La estimación y otros aspectos de la inferencia sobre los parámetros de interés (estimación interválica y prueba de hipótesis) se hace a partir de la distribución a posteriori.

Una forma equivalente de (47), la cual omite $f(\mathbf{y})$ debido a que no depende de $\boldsymbol{\theta}$ por lo que puede ser considerado una constante, es la densidad a posteriori no normalizada:

$$p(\boldsymbol{\theta}/\mathbf{y}) \propto f(\mathbf{y}/\boldsymbol{\theta})g(\boldsymbol{\theta}) \tag{48}$$

Esta última ecuación resume la forma en la que la escuela bayesiana realiza la inferencia, donde, la distribución a priori de $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$, refleja el estado de incertidumbre sobre los posibles valores de $\boldsymbol{\theta}$ previos.

2.6.2 Información a priori

La información a priori es la información sobre los parámetros de interés antes de realizar las observaciones o el experimento y son definidos independientemente de los datos. El problema es que muchas veces esta información no está bien especificada (Blasco, 2001), pero cuando se tiene información a priori exacta no hay discusión en cuanto a la utilización de métodos bayesianos siendo integrados usando las reglas de probabilidad utilizando el teorema de Bayes.

2.6.3 Distribución posterior

Considerando el teorema de Bayes y la partición del vector de las cantidades sujetas a incertidumbre como $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]$, que representan a distintos aspectos del modelo de probabilidad, por ejemplo $\boldsymbol{\theta}_1$ puede ser el componente de localización y $\boldsymbol{\theta}_2$ el componente de dispersión. La densidad posterior conjunta de todos los desconocidos es:

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) \propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \tag{49}$$

Por tanto la densidad marginal posterior de cada parámetro (o conjunto de parámetros) son por definición:

$$p(\boldsymbol{\theta}_1 / \mathbf{y}) = p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) d\boldsymbol{\theta}_2 \tag{50}$$

$$p(\boldsymbol{\theta}_2 / \mathbf{y}) = p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) d\boldsymbol{\theta}_1 \tag{51}$$

La densidad posterior condicional puede ser identificada (conceptualmente) desde la distribución posterior conjunta como:

$$p(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2, \mathbf{y}) = \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y})}{p(\boldsymbol{\theta}_2 / \mathbf{y})} \tag{52}$$

Como se está interesado en la variación con respecto a $\boldsymbol{\theta}_1$, entonces:

$$\begin{aligned} p(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2, \mathbf{y}) &\propto p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) \\ &\propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 / \mathbf{y}) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &\propto L(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2, \mathbf{y}) p(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2) \end{aligned}$$

donde $L(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2, \mathbf{y})$ es la función verosímil con $\boldsymbol{\theta}_2$ tratado como una constante conocida, más que como un aspecto sujeto a incertidumbre. El procedimiento de desarrollo implica que una distribución posterior condicional puede (a menudo) ser identificado por inspección de la densidad posterior conjunta, pero reteniendo solo la parte que varía con el parámetro de interés, tratando a los remanentes como conocidos. Este método es útil para identificar la distribución posterior condicional en el contexto de métodos de cadenas de Markov Monte Carlo (MCMC).

2.6.4 Métodos de Monte Carlo y Cadenas de Markov (MCMC)

Como describe Gelman *et al* (2014) la estimación de las distribuciones posteriores han llevado a desarrollar varios métodos que permitan tomar observaciones muestrales desde

éstas distribuciones. El método de simular las cadenas de Markov (MCMC) es un método general basado en el muestreo de observaciones (valores) de θ de las distribuciones aproximadas y luego actualizar estos valores en el proceso iterativo, para mejorar la aproximación de la distribución objetivo $p(\theta/y)$. El muestreo es realizado secuencialmente, con la distribución de los valores muestreados que dependen de los últimos valores observados, formando así las muestras de la cadena.

Una cadena de Markov es una secuencia de variables aleatorias $(\theta_1, \theta_2, \ldots)$, para la cual cualquier valor de t, la distribución θ_t dado todos los valores de $\theta's$ dependen solo de los valores más recientes θ_{t-1}. La clave de éxito de este método no es la propiedad de la cadena de Markov descrita, sino que la aproximación a la distribución de interés va mejorando en cada paso de la simulación, es decir converge a la distribución objetivo.

2.6.5 Muestreo de Gibbs

Gelman *et al* (2014) también indican que el muestreador de Gibbs es un algoritmo particular de Cadenas de Markov muy útil en algunos problemas multidimensionales, el cual es definido en términos de subvectores de θ. Suponga que el vector parámetro θ ha sido dividido en d componentes o subvectores $\theta = (\theta_1, \theta_2, \ldots \theta_d)$. Cada iteración de los ciclos de muestreo de Gibbs a través de los subvectores de θ, proporcionan subconjuntos condicionales de los valores restantes. Por lo tanto, habría d pasos en la iteración t. A cada iteración t, los d subvectores de θ son obtenidos ordenadamente, cada θ_j^t es muestreado de la distribución condicional dado los otros componentes de θ, $p(\theta_j^t / \theta_{-j}^{t-1}, y)$, representa todos los componentes de θ excepto para θ_j, a sus valores recientes $\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$. Así cada subvector θ_j es actualizado con los valores más recientes de los otros componentes de θ, los cuales son valores de la iteración t para los componentes ya actualizados en la iteración $t-1$ de los otros componentes.

2.6.6 Estimación de densidad e inferencia bayesiana desde el muestreo de Gibbs

Wang *et al* (1994), presenta una breve descripción de cómo se obtiene distribuciones posteriores y parámetros de estas distribuciones, la cual se detalla a continuación basadas sobre este muestreador.

Suponga que cada x_i, $i = 1, 2, ..., m$ es una realización de las corridas del muestreador de Gibbs de variable x. Las m muestras dependientes son usadas para calcular características de la distribución a posterior $P(x)$ por integración de Monte-Carlo.

$$u = \int g(x) dP(x) \tag{53}$$

Puede ser aproximado por:

$$\hat{u} = \frac{1}{m} \sum_{i=1}^{m} g(x_i) \tag{54}$$

donde $g(x)$ puede ser una característica de $P(x)$ como la media o la varianza. Cuando $m \to \infty$, $\hat{u}$ converge casi ciertamente a u.

Otra forma de calcular características de $P(x)$ es estimando primero la densidad $p(x)$, y luego obtener estadísticas resúmenes desde la densidad estimada usando un procedimiento numérico unidimensional. Si y_i $(i = 1, 2, ..., m)$ es otra realización de las corridas del muestreador de Gibbs, un estimador de $p(x)$ es dado por el promedio de las m densidades condicionales $p(x / y_i)$.

$$\tilde{p}(x) = \frac{1}{m} \sum_{i=1}^{m} p(x / y_i) \tag{55}$$

Otra forma alternativa de estimar $p(x)$ es usar solo muestras de x_i, $i = 1, 2, ..., m$ con estimadores de densidad de Kernel, para luego obtener estadísticas resúmenes desde la densidad estimada usando procedimientos numéricos unidimensionales

Para realizar inferencia sobre funciones de parámetros originales, por ejemplo si se quiere hacer inferencia sobre la función $z = \frac{x}{y}$, la cantidad $z_i = \frac{x_i}{y_i}$, $i = 1, ..., m$ es considerada una muestra aleatoria dependiente de tamaño m de una distribución con densidad $p(z)$.

Las aproximaciones descritas para estimar $P(x)$ en el párrafo anterior pueden también ser usados para hacer inferencia sobre $p(z)$.

Otra alternativa para estimar $p(z)$ es usar técnicas estándares para transformar las densidades condicionales $p(x/y)$ o $p(y/x)$ a $p(z/y)$ o $p(z/x)$. Si se requiere una transformación desde x/y a z/y; el jacobiano de la transformación es $|y|$, por lo que la densidad condicional de z/y está dada por:

$$p(z/y) = |y|\, p(zy/y) \tag{56}$$

Un estimador de $p(z)$ se obtiene promediando m densidades condicionales de $p(z/y)$:

$$\tilde{p}(z) = \frac{1}{m}\sum_{i=1}^{m} |y_i|\, p(zy_i / y_i) \tag{57}$$

2.7 Estimación bayesiana del modelo lineal

2.7.1 El modelo estadístico

Wang *et al* (1994) resumen el modelo para datos de una variable respuesta fenotípica con distribución normal como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon} \tag{58}$$

donde $\mathbf{y}$ es un vector que corresponde a la variable observada de orden $n \times 1$, $\boldsymbol{\beta}$ es un vector de orden $p \times 1$, que corresponde a los efectos "fijos", desde una perspectiva frecuentista, mas no desde un enfoque bayesiano, $\mathbf{u}_i$ son vectores que corresponden a efectos "aleatorios" de orden $q_i \times 1$, $\mathbf{X}$ y $\mathbf{Z}_i$ son matrices de incidencia y $\boldsymbol{\varepsilon}$ es otro vector aleatorio que corresponde al componente del error del modelo desde un enfoque frecuentista. (Sorensen y Gianola 2002).

La distribución condicional que generan los datos es:

$$\mathbf{y}/\boldsymbol{\beta}, \mathbf{u}_1 ..., \mathbf{u}_c, \sigma_e^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i, \mathbf{I}\sigma_e^2) \tag{59}$$

2.7.2 Distribuciones a priori

Desde la perspectiva bayesiana asignar distribuciones a priori a todas las distribuciones desconocidas es necesario para completar la especificación bayesiana, en el modelo (58) implica asignar distribuciones a priori a $\mathbf{u}_i$ y $\boldsymbol{\beta}$. Usualmente una distribución a priori flat o uniforme se asigna a $\boldsymbol{\beta}$, lo que representaría falta de información previa sobre este vector, esto es:

$$p(\boldsymbol{\beta}) \propto \text{constant} \tag{60}$$

Además se asume que:

$$\mathbf{u}_i / \mathbf{G}_i, \sigma^2_{\mathbf{u}_i} \sim N(\mathbf{0}, \mathbf{G}_i \sigma^2_{\mathbf{u}_i}) \qquad i = 1, 2, ..., c \tag{61}$$

donde $\mathbf{G}_i$ es una matriz conocida (en un modelo animal al menos un $\mathbf{G}_i$, contiene información genealógica de los individuos) y $\sigma^2_{\mathbf{u}_i}$ es la varianza de la distribución priori de $\mathbf{u}_i$. Todos los $\mathbf{u}_i$ se asumen que tienen a prioris mutuamente independientes así como también independientes de $\boldsymbol{\beta}$.

Como a prioris para los componentes de varianza se asumen distribuciones independientes chi-cuadradas invertidas escaladas, por tanto:

$$p(\sigma^2_{\mathbf{u}_i} / \nu_{\mathbf{u}_i}, s^2_{\mathbf{u}_i}) \propto (\sigma^2_{\mathbf{u}_i})^{-\nu_{\mathbf{u}_i}/2-1} \exp(-\frac{1}{2} \nu_{\mathbf{u}_i} s^2_{\mathbf{u}_i} / \sigma^2_{\mathbf{u}_i}) \qquad i = 1, 2, ..., c \tag{62}$$

$$p(\sigma^2_e / \nu_e, s^2_e) \propto (\sigma^2_e)^{-\nu_e/2-1} \exp(-\frac{1}{2} \nu_e s^2_e / \sigma^2_e) \tag{63}$$

donde, ν_e, ν_{u_i} son parámetros que contienen información sobre el "grado de credibilidad" y $s^2_e, s^2_{u_i}$ pueden ser interpretados como los valores priori apropiadas de las varianzas.

2.7.3 Densidad posterior conjunta

La densidad posterior conjunta de $\mathbf{u}_i$, $\boldsymbol{\beta}$, $\sigma^2_{u_i}$ y σ^2_e $i = 1, 2, ..., c$ es el producto de las densidades priori asociadas y descritas anteriormente.

Haciendo:

$\boldsymbol{\theta}' = (\boldsymbol{\beta}', u_1', ..., u_c') = (\theta_1, \theta_2, ...\theta_N)$, $\boldsymbol{\theta}'_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_N)$

$\mathbf{v}' = (\sigma^2_{\mathbf{u}_1}, \sigma^2_{\mathbf{u}_2}, ..., \sigma^2_{\mathbf{u}_c})$, $\mathbf{v}'_{-i} = (\sigma^2_{\mathbf{u}_1}, \sigma^2_{\mathbf{u}_2}, ..., \sigma^2_{\mathbf{u}_{i-1}}, \sigma^2_{\mathbf{u}_{i+1}}, ..., \sigma^2_{\mathbf{u}_c})$ y

$\mathbf{v}' = (\nu_{\mathbf{u}_1}, \nu_{\mathbf{u}_2}, ..., \nu_{\mathbf{u}_c}, \nu_e)$, $\mathbf{s}' = (s^2_{\mathbf{u}_1}, s^2_{\mathbf{u}_2}, ..., s^2_{\mathbf{u}_c})$

$\mathbf{v}'$ y $\mathbf{s}'$ son los conjuntos de todas las varianzas a prioris y los grados de creencia respectivamente.

Como mostraron Macedo y Gianola; y Gianola *et al*, citado por Wang *et al* (1994), la distribución posterior conjunta tiene la forma de una distribución gamma-normal:

$$p(\boldsymbol{\theta}, \mathbf{v}, \sigma_e^2 / \mathbf{y}, \mathbf{s}, \nu) \propto$$
$$(\sigma_e^2)^{-(n+\nu_e+2)/2} \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i) + \nu_e s_e^2\right]\right\} \quad (64)$$
$$\times \prod_{i=1}^{c}\left[(\sigma^2_{\mathbf{u}_i})^{-(q_i+\nu_{\mathbf{u}_i}+2)/2} \exp\left\{-\frac{1}{2\sigma^2_{u_i}}\right\}\left[\mathbf{u}_i' \mathbf{G}_i^{-1} \mathbf{u}_i + \nu_{\mathbf{u}_i} s^2_{\mathbf{u}_i}\right]\right]$$

Las inferencias sobre cada uno de los parámetros desconocidos $\boldsymbol{\theta}, \mathbf{v}, \sigma_e^2$ son basadas en sus respectivas densidades marginales, las cuales son obtenidas por una integración sucesiva de la densidad conjunta (64) con respecto a cada uno de los parámetros de interés.

Es difícil llevar a cabo esta integración de forma analítica, por lo que es bastante usual utilizar el muestreo de Gibbs.

2.7.4 Densidades posteriores condicionales

Para implementar el muestreo de Gibbs es necesario determinar todas las densidades posteriores condicionales de los parámetros desconocidos. Cada una de estas densidades se obtiene considerando sólo un parámetro desconocido en (64) y tratando a los parámetros restantes como conocidos.

De este modo, la distribución posterior condicional de $\boldsymbol{\theta}$ dado los componentes de varianza y los datos es una normal multivariante con media $\hat{\boldsymbol{\theta}}$ y varianza $\hat{\mathbf{V}}$, es:

$$\boldsymbol{\theta} / \mathbf{y}, \mathbf{v}, \sigma_e^2 \sim N(\hat{\boldsymbol{\theta}}, \hat{\mathbf{V}}) \quad (65)$$

donde $\hat{\boldsymbol{\theta}} = \mathbf{W}^{-1}\mathbf{b}$ y $\hat{\mathbf{V}} = \mathbf{W}^{-1}\sigma_e^2$, tomando a las ecuaciones del modelo mixto dadas en (2) como $\mathbf{W}\hat{\boldsymbol{\theta}} = \mathbf{b}$, con $\mathbf{R} = \mathbf{I}$.

La distribución posterior condicional de σ_e^2 tiene la forma chi-cuadrada escalada invertida:

$$p(\sigma_e^2 / \mathbf{y}, \boldsymbol{\theta}, \mathbf{v}, \mathbf{s}, \nu) \propto$$
$$(\sigma_e^2)^{-(n+\nu_e+2)/2} \exp\left(-\frac{1}{2\sigma_e^2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i) + \nu_e s_e^2 \right] \right) \quad (66)$$

que puede ser presentada como:

$$\sigma_e^2 / \mathbf{y}, \boldsymbol{\theta}, \mathbf{v}, \mathbf{s}, \nu \sim \tilde{\nu}_e \tilde{s}_e^2 \chi_{\tilde{\nu}_e}^{-2}$$

con parámetros $\tilde{\nu}_e = n + \nu_e$, y $\tilde{s}_e^2 = \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{c} \mathbf{Z}_i \mathbf{u}_i) + \nu_e s_e^2 \right] / \tilde{\nu}_e$

La distribución posterior condicional de cada $\sigma_{\mathbf{u}_i}^2$ tiene también la forma chi-cuadrada escalada invertida:

$$p(\sigma_{\mathbf{u}_i}^2 / \mathbf{y}, \boldsymbol{\theta}, \mathbf{v}_{-i}, \sigma_e^2, \mathbf{s}, \nu) \propto (\sigma_{\mathbf{u}_i}^2)^{-(q_i+\nu_{\mathbf{u}_i}+2)/2} \exp\left(-\frac{1}{2\sigma_{u_i}^2}\left[\mathbf{u}_i' \mathbf{G}_i^{-1} \mathbf{u}_i + \nu_{\mathbf{u}_i} s_{\mathbf{u}_i}^2 \right] \right) \quad (67)$$

que puede ser presentada como:

$$\sigma_{\mathbf{u}_i}^2 / \mathbf{y}, \boldsymbol{\theta}, \mathbf{v}_{-i}, \sigma_{e_i}^2, \mathbf{s}, \nu \sim \tilde{\nu}_{\mathbf{u}_i} \tilde{s}_{\mathbf{u}_i}^2 \chi_{\tilde{\nu}_{\mathbf{u}_i}}^{-2} \qquad i = 1, ..., c$$

con parámetros $\tilde{\nu}_{\mathbf{u}_i} = q_i + \nu_{\mathbf{u}_i}$, y $\tilde{s}_{\mathbf{u}_i}^2 = \left[\mathbf{u}_i' \mathbf{G}_i^{-1} \mathbf{u}_i + \nu_{\mathbf{u}_i} s_{\mathbf{u}_i}^2 \right] / \tilde{\nu}_{\mathbf{u}_i}$

2.7.5 El Muestreo de Gibbs para obtener distribuciones marginales

Wang (1993) señala que el muestreo de Gibbs fue inicialmente usado en la estadística espacial en la reconstrucción de imágenes por Geman y Geman (1984). Aplicaciones de la inferencia bayesiana fueron descritas por Gelfand y Smith (1990) y Gelman *et al* (1990). Su utilidad como una herramienta estadística permite generar muestras de distribuciones complejas, siendo su desarrollo en algunos problemas incuestionables.

El propósito es generar muestras aleatorias desde la distribución posterior conjunta (64), a través de sucesivas muestras renovadas y tomadas desde el muestreador de Gibbs (65-67). Formalmente el trabajo del muestrador de Gibbs se describe a continuación:

1. Establecer valores iniciales para $\boldsymbol{\theta}, \mathbf{v}$ y σ_e^2.
2. Generar θ_i desde (35) y actualizar θ_i, $i = 1, ..., N$.
3. Generar σ_e^2 desde (36) y actualizar σ_e^2.
4. Generar $\sigma_{\mathbf{u}_i}^2$ desde (37) y actualizar $\sigma_{\mathbf{u}_i}^2$, $i = 1, ..., c$

5. Repetir pasos de (1) hasta (4) k -veces, donde k es la longitud de la cadena.

Cuando $k \to \infty$, se crea una cadena de Markov con una distribución de equilibrio que tiene a (64) como densidad. A este procedimiento suele llamarse algoritmo de una cadena simple.

Las iteraciones iniciales no son usualmente almacenadas como muestras puesto que se considera que la cadena no ha alcanzado la distribución de equilibrio; a este período se le conoce como "burn-in". Después de este período se almacena una observación cada d iteraciones, donde d es un número entero positivo, el número total de observaciones m, es el tamaño de la muestra almacenada. Estas m observaciones forman la distribución a posteriori con densidad dada en (64), siempre que el muestreador de Gibbs converja.

La i-ésima observación muestreada $\left\{\boldsymbol{\theta}_i, \mathbf{v}_i, \left(\sigma_e^2\right)_i\right\}$ $i = 1,...,m$ es un vector y cada elemento de este vector es una observación muestreada de la distribución marginal correspondiente.

2.7.6 El Error de Montecarlo

Ntzoufras (2009) señala que el error de Montecarlo (EMC) mide la variabilidad de cada estimación debido a la simulación y sugiere que el EMC debe ser bajo para calcular el parámetro de interés con mayor precisión. El EMC es proporcional a la inversa del tamaño de muestra generada en el proceso de simulación. Por tanto para un número suficientemente grande de iteraciones, el parámetro de interés será estimado con mayor precisión.

Existen dos formas populares de calcular el EMC, uno es conocido como el método de media de lotes (*batch mean*) y el otro como el método de estimador de ventana (*window estimator*). El método de Bach particiona los resultados de la muestra del proceso de simulación en k lotes, de manera que éste y el tamaño muestral de cada lote deben ser suficientemente grandes para estimar la varianza consistentemente y eliminar autocorrelaciones. Para obtener el EMC de cualquier estimador posterior de interés (media, moda, percentil, etc) se calcula el estimador en toda la muestra $\hat{U}$, y en cada lote $\hat{U}_b$, $b = 1,...,k$, para luego calcular $EMC(\hat{U}) = \sqrt{\frac{1}{k(k-1)} \sum_{b=1}^{k} (\hat{U} - \hat{U}_b)^2}$

El segundo método está basado sobre la expresión de la varianza en muestras autocorrelacionadas, esto es $EMC(\hat{U}) = SD(\hat{U})\sqrt{1+\sum_{k=1}^{\infty}\hat{\rho}_k(\hat{U})}$, donde $SD(\hat{U})$ es la desviación estándar del estimador en cada observación de la muestra del proceso de simulación y $\hat{\rho}_k(\hat{U})$ es la autocorrelación estimada del rezago.

III. MATERIALES Y MÉTODOS

3.1 Descripción de los datos

Los datos son registros de 3 397 lactaciones del primer al quinto parto de 1 359 vacas Holsteins, hijas de 38 toros en 57 rebaños. Todos los registros corresponden a vacas con al menos 100 días de leche. La información genealógica, pedigrí, de estas vacas comprende 5 generaciones con un total de 6 547 animales. Toda esta información ha sido descargada desde USDA (*United State Department of Agriculture*) http://www.aipl.arsusda.gov/, 2010 y están disponibles en el conjunto de datos *milk* y *pedCows* de la librería pedigreemm en R.

3.2 Codificación y descripción de las variables

Datos de producción *milk***:**

id: factor de identificación de la vaca o individuos (1 359 vacas o niveles)
lact: número de lactación o parto de la vaca (1, 2, 3, 4 o 5 partos o niveles)
herd: factor indicador del rebaño (57 rebaños o niveles)
dim: número de días en leche de la lactación (covariable)
milk: producción de leche estimada a 305 días en onzas

Datos de predigrí *pedCows***:**

sire: factor de identificación del progenitor macho (4 568 machos o niveles)
dam: factor identificación del progenitor hembra (3 848 hembras o niveles)
label: individuo o animal (6 547 animales o niveles)

3.3 Materiales

Computadora AMD A-series

Software R 3.01, usando las librerías que se detallan a continuación:

Librería	Referencia
Pedigreemm	Bates y Vásquez (2013)
MCMCglmm	Handfield (2016)
mcmcse	Flegal et al (2016)
lme4	Bates et al (2016)
arm	Gelman et al (2016)
ggplot2	Wickham and Chang (2016)
psych	Revelle (2016)
MASS	Ripley et al (2016)

3.4 Metodología aplicada

El procedimiento o plan de análisis realizado para la estimación se detalla a continuación:

1. Se realizó un breve análisis exploratorio univariado de los datos con la finalidad de observar y describir la distribución de variables y/o covariables que componen el modelo de forma independiente y un breve análisis exploratorio bivariado entre la variable respuesta y algunas variables predictoras con la finalidad de observar y describir la relación entre éstas.
2. A partir del modelo animal formulado se estimó el modelo y los componentes de varianza aplicando REML y se aplicó la prueba de razón de máxima verosimilitudes para seleccionar los componentes de varianza en el modelo.
3. Para realizar el análisis de residuales del modelo estimado se adaptó funciones en R, programadas por Nobre y Singer (2007), para incorporar la matriz de relaciones genéticas aditivas en el análisis y construir cada uno de los gráficos de diagnóstico descritos en el cuadro 1.
4. A partir del modelo animal formulado se estimó el modelo y los componentes de varianza utilizando el método bayesiano. El algoritmo MCMC se ejecutó con un

total de 100000 iteraciones, 10000 burn-in, almacenadas cada 10 iteraciones, como se describe en 2.7.5.

5. Se evaluó el comportamiento del algoritmo MCMC se observó la convergencia y autocorrelación de la cadena de las muestras.

3.5 Modelo animal formulado para estimar componentes de varianza

Siguiendo la formulación de Henderson del modelo animal dado en (1), Sorensen y Gianola (2002) y Mrode (2014) resumen matricialmente el modelo animal univariado a utilizar para datos de una variable fenotípica con distribución normal, como:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 u}_1 + \mathbf{Z_2 u}_2 + \boldsymbol{\varepsilon} \tag{68}$$

donde $\mathbf{y}$ es un vector que corresponde a la producción de leche estandarizada, $\boldsymbol{\beta}$ es un vector que corresponde a los efectos fijos del factor número de lactación o parto; y la covariable logaritmo del número de días en leche; el vector $\mathbf{u}_1$ corresponden al efecto genético aditivo del animal y $\mathbf{u}_2$ efecto aleatorio del rebaño, $\mathbf{X}$, $\mathbf{Z}_1$ y $\mathbf{Z}_2$ son matrices de incidencia relacionadas con $\boldsymbol{\beta}$, $\mathbf{u}_1$ y $\mathbf{u}_2$ respectivamente y ε es un vector aleatorio de residuales.

El modelo (68) escrito de forma individual para cada observación es:

$$y_{ijk} = \beta_0 + L_i + \beta_1 \log(\text{dim})_{ij} + c_j + h_k + e_{ijk} \tag{69}$$

donde:

y_{ijk} es la producción de leche estandarizada sobre el parto i, j ésima vaca, β_0 es la media general; L_i es el efecto fijo del número de lactación o parto (i = 1, 2, …, 5); dim_{ij} es el número de días en leche de la vaca j en la $\underline{i}$-ésima lactación (covariable); β_1 es el coeficiente de regresión de dim; c_j es el efecto aleatorio aditivo de la vaca j (j = 1, 2, …, 1359); h_k es el efecto aleatorio del rebaño k (k = 1, 2, …, 57); y e_{ijk} es el efecto aleatorio residual.

3.6 Modelo lineal mixto para estimar componentes de varianza por el método REML

Las distribuciones para los componentes aleatorios del modelo (69) tienen la forma siguiente:

$$\begin{pmatrix} c \\ h \\ e \end{pmatrix} \sim N \begin{bmatrix} \mathbf{A}_{1359}\sigma_c^2 & 0 & 0 \\ 0 & \mathbf{I}_{57}\sigma_h^2 & 0 \\ 0 & 0 & \mathbf{I}_{3397}\sigma_e^2 \end{bmatrix}$$

donde:

$c = \{c_j\}$ es el vector de efecto aditivo de la vaca (valor genético)

$h = \{h_k\}$ es el vector de efecto de rebaños

$e = \{e_{ijk}\}$ es el vector de residuales

σ_c^2, σ_h^2 y σ_e^2 son los componentes de varianzas de los efectos genéticos aditivos, entre rebaños, y varianza residual respectivamente.

A representa la matriz de relaciones genéticas aditivas entre vacas y es estimada con la información genealógica del pedigrí (*pedCows*)

La heredabilidad del efecto genético aditivo está dado por: $h_c^2 = \dfrac{\sigma_c^2}{\sigma_c^2 + \sigma_h^2 + \sigma_e^2}$

El componente ambiental que corresponde al efecto rebaño bajo este modelo está dado por:

$$h_h^2 = \frac{\sigma_h^2}{\sigma_c^2 + \sigma_h^2 + \sigma_e^2}$$

Para estimar los componentes de varianza del modelo se utilizó la librería Pedigreemm de R que es una extensión de la librería lme4, pero que permite ingresar la información genealógica de los individuos para la estimación del modelo mixto. En esta librería no se ha implementado funciones para el diagnóstico de residuales en función a lo presentado en 2.5 para modelos lineales mixtos.

3.6.1 Prueba de significancia de los componentes de varianza

Para probar significancia del componente genético aditivo, $\sigma_c^2 > 0$, se aplicó la prueba de razón de verosimilitudes descrita en 2.4.8, considerando:

Modelo 2: $y_{ijk} = \beta_0 + L_i + \beta_1 \log(\text{dim})_{ij} + c_j + h_k + e_{ijk}$

Modelo 1: $y_{ijk} = \beta_0 + L_i + \beta_1 \log(\text{dim})_{ij} + h_k + e_{ijk}$

Haciendo l_{R2} el log-verosímil del modelo 2 y l_{R1} log-verosímil del modelo 1, el estadístico de prueba $D = 2\log(l_{R2} / l_{R1}) = 2(\log l_{R2} - \log l_{R1})$ bajo la hipótesis nula se aproxima a una χ_1^2 .

Para probar significancia del componente rebaño, $\sigma_h^2 > 0$, se aplicó la prueba de razón de verosimilitudes descrita en 2.4.8, considerando:

Modelo 2: $y_{ijk} = \beta_0 + L_i + \beta_1 \log(\text{dim})_{ij} + c_j + h_k + e_{ijk}$

Modelo 1: $y_{ijk} = \beta_0 + L_i + \beta_1 \log(\text{dim})_{ij} + c_j + e_{ijk}$

Haciendo l_{R2} el log-verosímil del modelo 2 y l_{R1} log-verosímil del modelo 1, el estadístico de prueba $D = 2\log(l_{R2} / l_{R1}) = 2(\log l_{R2} - \log l_{R1})$ bajo la hipótesis nula se aproxima a una χ_1^2.

3.6.2 Análisis de residuales

Singer *et al* (2013) señalan que utilizar las herramientas de diagnóstico como las descritas en 2.5 puede no ser una tarea fácil, debido a que algunas funciones diseñadas para generar gráficos de diagnósticos no han sido implementados en la mayoría de los programas estadísticos. Aunque algunas de estas funciones pueden ser obtenidas de los autores Singer *et al* (2013), su uso en aplicaciones prácticas no siempre es directa.

Para el chequeo de supuestos y detección de observaciones extremas se editó las funciones en R programadas por estos autores. La metodología que se siguió fue descrita por Harville y Callanan citado por Vazquez *et al*. (2010), la cual consiste en post multiplicar a la matriz $\mathbf{Z}$ en (1) por la descomposición de Cholesky de la matriz de relaciones genéticas aditivas $\mathbf{A}$, es decir en (6):

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}} = \sigma^2 \mathbf{A}$$

Del modelo ajustado se extrajo información necesaria como $\mathbf{X}$, $\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, $\mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{Z}^T\sigma^2 + \sigma_{\mathbf{e}}^2\mathbf{I}$, para estimar los elementos $\hat{\boldsymbol{\xi}}_{ij}^*$, $\mathbf{V}_i^*$, $\hat{\mathbf{e}}_{ij}^*$, $\mathbf{M}_i$, $\mathbf{c}_k^T\hat{\mathbf{e}}_{ij}^*$ y construir los gráficos descritos en el cuadro 1. El código trabajado en R se muestra en el anexo.

3.7 Modelo lineal mixto para estimar componentes de varianza por el método bayesiano

La distribución condicional que generan los datos del modelo descrito en (68) es:

$$\mathbf{y}/\boldsymbol{\beta},\mathbf{u}_1,\mathbf{u}_2,\sigma_e^2 \sim N(\mathbf{X}\boldsymbol{\beta}+\sum_{i=1}^{2}\mathbf{Z}_i\mathbf{u}_i,\mathbf{I}\sigma_e^2)$$

Respecto a los supuestos de las distribuciones que son:

$\mathbf{u}_1/\mathbf{A} \sim N(\mathbf{0},\mathbf{A}\sigma_1^2)$, $\mathbf{u}_2 \sim N(\mathbf{0},\mathbf{I}\sigma_2^2)$ y $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{I}\sigma_\varepsilon^2)$, donde $\mathbf{A}$ es la matriz de covarianzas aditivas entre los individuos, además $\mathbf{u}_1$, $\mathbf{u}_2$ y $\boldsymbol{\varepsilon}$ son asumidas por ser independientes entre sí.

Respecto a los a prioris para $\boldsymbol{\beta}$ como es usual se asumió una distribución uniforme, es decir $p(\boldsymbol{\beta}) \propto$ constante, además de independencia entre $\boldsymbol{\beta}$, $\mathbf{u}_1$, $\mathbf{u}_2$ y $\boldsymbol{\varepsilon}$.

La densidad posterior conjunta de todos los parámetros desconocidos es proporcional a:

$$p(\boldsymbol{\beta},\mathbf{u}_1,\mathbf{u}_2,\sigma_1^2,\sigma_2^2,\sigma_e^2/\mathbf{y}) \propto$$
$$p(\boldsymbol{\beta})\,p(\mathbf{u}_1/\sigma_1^2)\,p(\sigma_1^2)\,p(\mathbf{u}_2/\sigma_2^2)\,p(\sigma_2^2)\,p(\sigma_e^2)\,p(\mathbf{y}/\boldsymbol{\beta},\mathbf{u}_1,\mathbf{u}_2,\sigma_e^2)$$

a partir de esta se deduce la distribución posterior completa de cada parámetro a estimar.

Para conseguir muestras de la distribución posterior conjunta se aplicó el muestreo de Gibss con una sola cadena de 100000 iteraciones y almacenadas cada 10 iteraciones, descartándose las 10000 primeras (burn-in). Para ello se utilizó la librería MCMCglmm implementado en el paquete R (Hadfield, 2010).

Para chequear el comportamiento del algoritmo MCMC se observó la convergencia a través de los gráficos de traza, que muestran la evolución de los valores muestreados a lo largo de las iteraciones y las autocorrelaciones de la cadena de las muestras. Asimismo se obtuvo el error de Montecarlo para cada estimación con la finalidad de medir la variabilidad de la estimación a través de la simulación.

IV. RESULTADOS Y DISCUSIÓN

4.1 Análisis exploratorio de datos

4.1.1 Análisis univariante

La distribución de la producción de leche a los 305 días aparentemente tiene un comportamiento no muy lejano a una distribución simétrica como se observa en la figura1. De acuerdo al Cuadro 1 presenta ligera asimetría negativa y algo leptocúrtica, además no tiene un comportamiento normal ($p<0.00$) que es lo esperado, esto indicaría la necesidad de modelar esta característica en función de otros componentes que influyen en su respuesta.

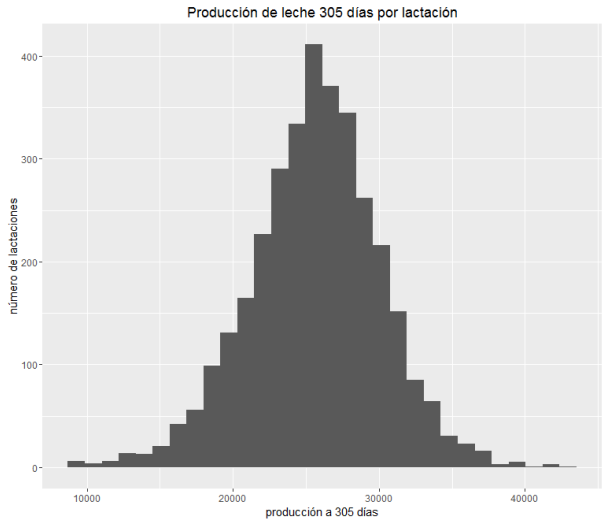

Figura 1 Distribución de la producción de leche a los 305 días

Cuadro 1 Estadísticas descriptivas para la producción de leche a los 305 días

media	sd	mediana	Min	max	skew	kurtosis
25632.14	4472.22	25799	8953	42614	-0.18	0.6

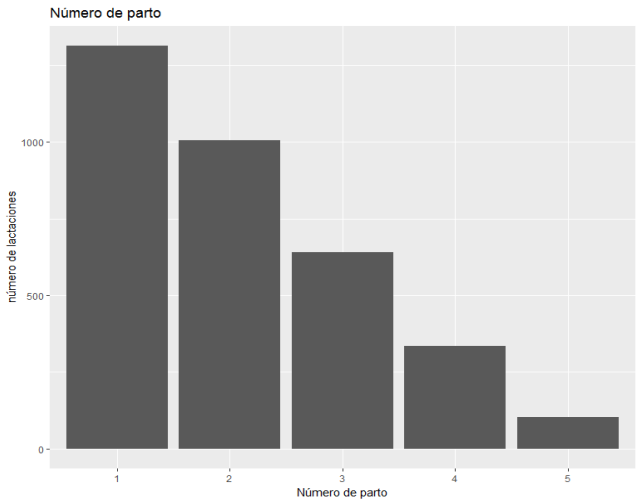

Figura 2 Distribución de las lactaciones observadas, según el número de parto

En la figura 2 se muestra que de todas las lactaciones registradas el 38.7 por ciento (1314 lactaciones) corresponden al primer parto, el 29.6 por ciento (1006 lactaciones) corresponden al segundo parto, el 18.8 por ciento (640 lactaciones) corresponden al tercer parto, el 9.8 por ciento corresponde al cuarto parto y el 3 por ciento corresponde al quinto parto.

La distribución del número de días en leche tiene una sesgo positivo, como se observa en la gráfica 3, con un coeficiente de asimetría de 1.16, siendo el número medio de días en leche por lactación de 347.4 días y más del 50% de las lactaciones fueron mayores a 325días .

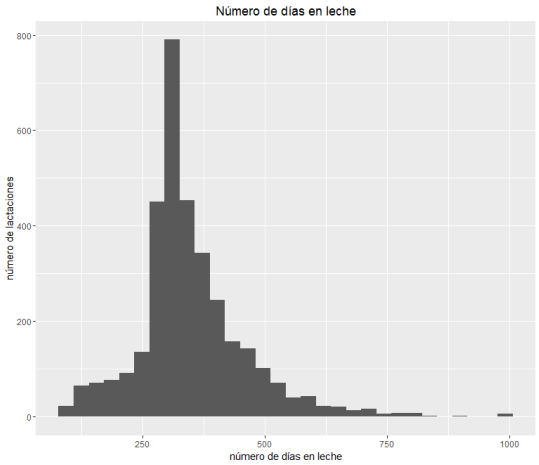

Figura 3 Distribución del número de días en leche

Cuadro 2 Estadísticas descriptivas para el número de días en leche

Media	sd	mediana	min	max	skew	kurtosis
347.42	111.12	325	100	999	1.16	3.67

4.1.2 Análisis Bivariante

La relación entre la producción de leche y el número de días en leche es directa pero no muy fuerte como se observa en la figura 4, el coeficiente de correlación de Pearson es 0.263, pero significativo (p=0.00), lo que justificaría tomar en cuenta esta covariable en el modelo que explica la producción de leche a los 305 días.

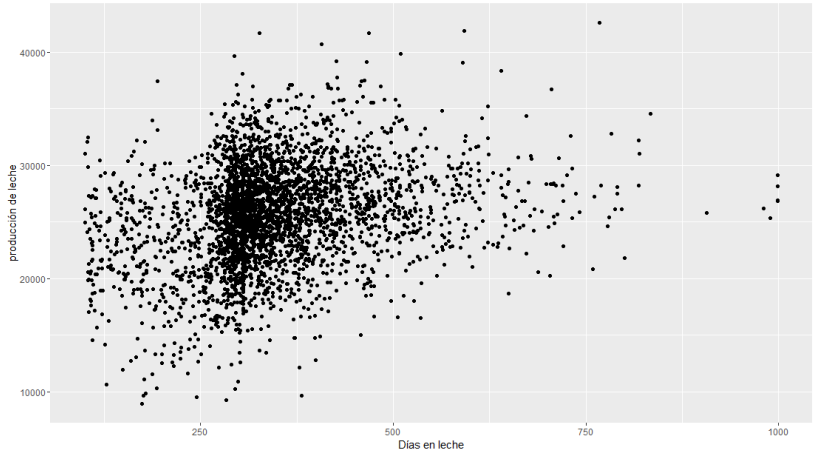

Figura 4 Dispersión entre días en leche y producción a los 305 días

La distribución de la producción para cada lactación es bastante similar en el 50% central de los datos como se muestra en la figura 5, por lo que no se observa un efecto claro del número de lactación o parto en la producción de leche. Se observa también valores atípicos especialmente en el primer parto o lactación que presenta ligeramente una mayor variabilidad que los otros partos, esto puede deberse a las diferentes respuestas de cada individuo.

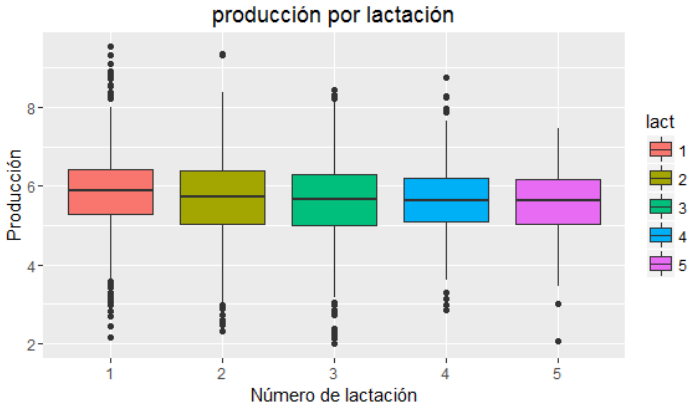

Figura 5 Producción de leche según el número de lactación

La producción presenta bastante dispersión entre los 57 rebaños observados, como se aprecia en la figura 6, por lo que justificaría tratarlo como un efecto aleatorio dentro del modelo, es decir asumir que proceden de una población de rebaños.

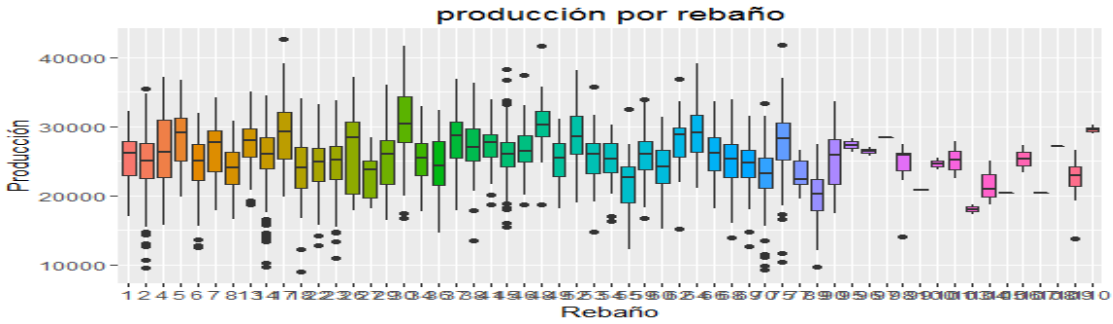

Figura 6 Producción de leche según el rebaño

4.2 Estimación del modelo utilizando REML

Fue estimado el modelo lineal mixto descrito en 3.5, con los componentes fijos y aleatorios detallados en las expresiones (68) y (69) utilizando la librería pedigreemm de R, que fue desarrollado por Vásquez et al (2010) como una extensión de la librería lme4 (Bates et al, 2015) que ajusta modelos lineales mixtos con efectos aleatorios correlacionados para respuesta de conteo, gaussiana y binaria.

Se aplicó el método de la razón de verosimilitudes (2.4.8) para seleccionar los componentes aleatorios que se incluyen en el modelo. Las hipótesis utilizadas están descritas en 3.6.1, cada prueba se realizó considerando que los componentes fijos ya están en el modelo. Los resultados se muestran en el Cuadro 3, y se observa que tanto el

componente genético del animal y de rebaño son significativos. Así también en el Cuadro 4 el modelo con menores estadísticos de ajustes BIC descrito en (45) y AIC descrito en (46) corresponden al modelo que incluye ambos componentes aleatorios.

Cuadro 3 Contribución de los componentes aleatorios del modelo

Modelo (contribución)	D	p-valor
Efecto genético aditivo (*c*)	230.3684	0.000
Efecto rebaño (*h*)	310.1868	0.000

D: estadístico de la prueba de razón de verosimilitudes

Cuadro 4 Estadísticos de ajustes de modelos lineales

Componentes aleatorios en el modelo	LogL	AIC	BIC
Solo efecto genético animal: *h*	-4306.931	8709.68	8629.861
Solo efecto rebaño: *c*	-4346.84	8629.861	8758.725
Ambos efectos aleatorios: *h+c*	-4191.746	8401.493	8456.669

LogL: máximo log-verosimil,
AIC: criterio de información de Akaike,
BIC: criterio de información bayesiano

La estimación de la heredabilidad (2.3.1) utilizando la metodología REML fue de 0.3177 como se muestra en el Cuadro 5, lo que indica una heredabilidad moderada de este carácter (entre 0.15 y 0.4 según Gutiérrez, JP (2010)), esto indica que el efecto genético aditivo, porcentaje de variación de la producción de leche que será trasmitida a su descendencia de la población estudiada es de 31.7% aproximadamente.

El porcentaje de la variabilidad de la producción de leche debido a la variabilidad de rebaños es de 19.62% menor que el de la variabilidad genética aditiva.

Cuadro 5 Estimados de los componentes de varianza y heredabilidad para la producción de leche

N	$\hat{\sigma}_c^2$	$\hat{\sigma}_h^2$	$\hat{\sigma}_e^2$	$\hat{h}^2$	$\hat{c}^2$	**AIC**
3397	0.3154	0.1955	0.4819	0.3177	0.19692	8401.493

σ_c^2, σ_h^2, σ_e^2, c^2 y h^2: varianza aditiva, varianza respecto al rebaño, varianza residual, heredabilidad aditiva y proporción de la varianza ambiental rebaño respectivamente

En el Cuadro 6 se presenta los coeficientes estimados de los efectos fijos del modelo, se observa que conforme el número de lactación aumenta el efecto en la producción media de leche (en relación con la primera lactación) va disminuyendo hasta la quinta lactación. El efecto del logaritmo de días en leche es positivo indicado un aumento en la producción cuando aumenta los días en leche.

Cuadro 6 Coeficientes de los componentes fijos del modelo

Componente	Coeficiente	Error estándar	t-valor
Intercepto	1.56446	0.27200	5.752
lact2	-0.14675	0.03002	-4.889
lact3	-0.27966	0.03580	-7.812
lact4	-0.29140	0.04698	-6.202
lact5	-0.35204	0.0778	-4.523
log(dim)	0.73970	0.04374	16.910

4.3 Diagnóstico de residuales

Se estimaron los tres tipos de residuales estandarizados y otros para luego obtener los gráficos con fines de diagnóstico como se especificó en el Cuadro 1, adaptando el código y las funciones en R proporcionadas por Singer *et al*. (2013) para este modelo en particular.

La figura 7 permite revisar la linealidad del efecto fijo, de la covariable días en leche en el modelo, en esta no se muestra un patrón por lo que indicaría que el logaritmo de días en leche tiene un efecto lineal en la producción de leche.

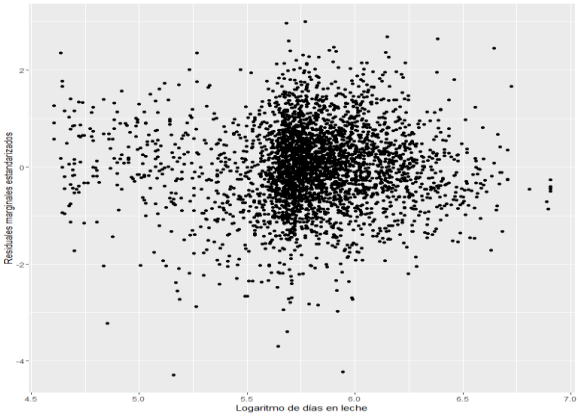

Figura 7 Residuales marginales estandarizados vs log (días en leche)

Otro gráfico que evalúa la linealidad de los efectos fijos se muestra en la figura 8, no se observa algún patrón definido por lo que no se descartaría la relación lineal con los efectos fijos, asimismo el histograma de la distribución de los residuales marginales estandarizados muestra un comportamiento simétrico.

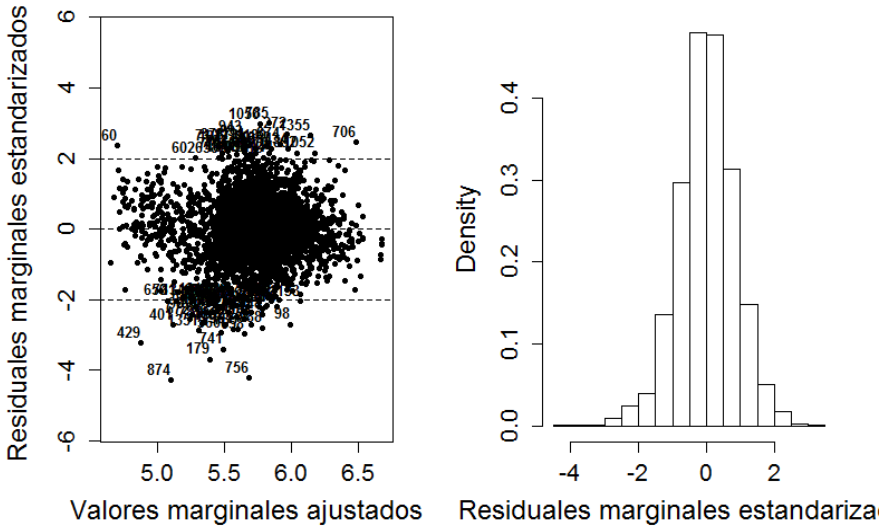

Figura 8 Residuales marginales estandarizados vs ajustados e histrograma de los residuales marginales

La figura 9, presenta los residuales marginales ajustados versus los índices de observación, a partir de este gráfico se encuentra que hay 95 observaciones atípicas de 45 animales. Se consideraron atípicas a aquellas observaciones con residuales, en términos absolutos, mayores que 2.

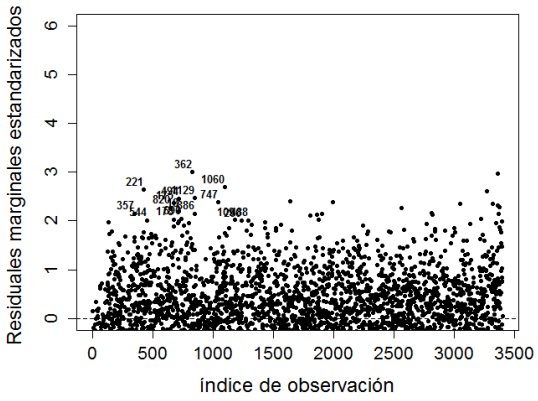

Figura 9 Residuales marginales estandarizados vs índices de observación

El gráfico de los residuales condicionales estandarizados versus las observaciones estimadas, figura 10, sugiere que no habría homocedasticidad de los errores condicicionales, puesto que no se observa un patrón aleatorio, así también el histograma nos sugiere un comportamiento simétrico de estos residuales.

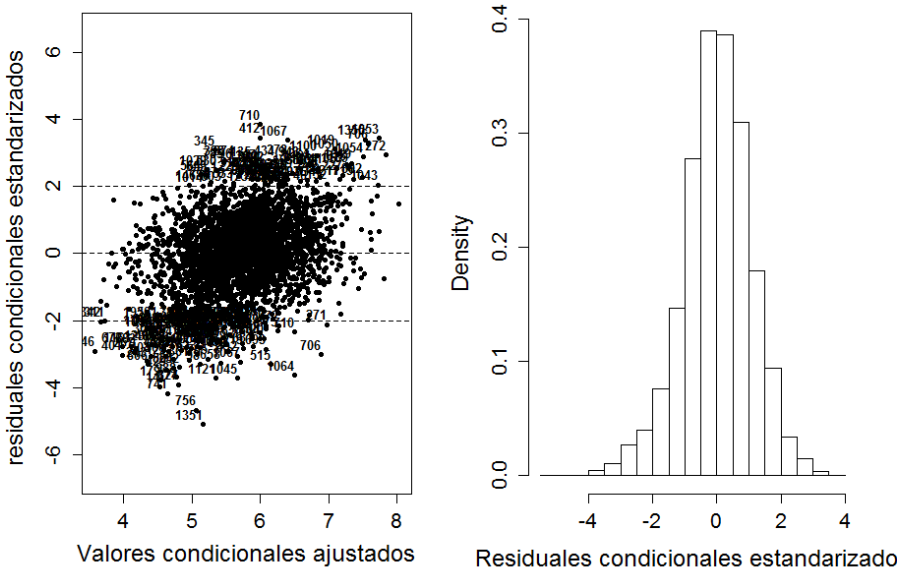

Figura 10 Residuales condicionales estandarizados vs ajustados e histrograma de los residuales condicionales

La figura 11, se muestra a los residuales condicionales ajustados versus los índices de observación, desde el cual se encuentra que hay 234 observaciones atípicas de 109 animales, las observaciones consideradas como atípicas son aquellas con residuales, en términos aboslutos, mayores que 2.

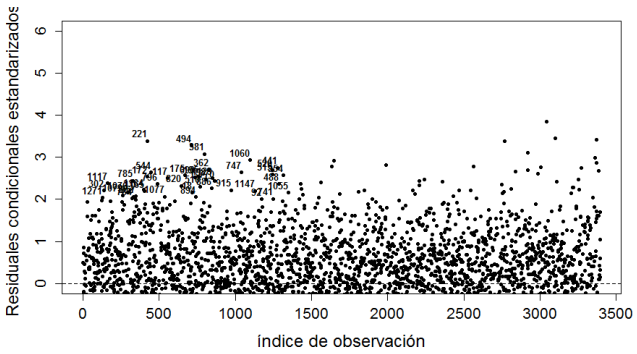

Figura 11 Residuales condicionales estandarizados vs índices de observación

En la figura 12 se observa que 215 animales pueden considerarse extremos. Se consideraron individuos extremos a aquellos cuya distancia de Mahalanobis fue mayor a dos veces la media de las distancias estimadas para cada animal.

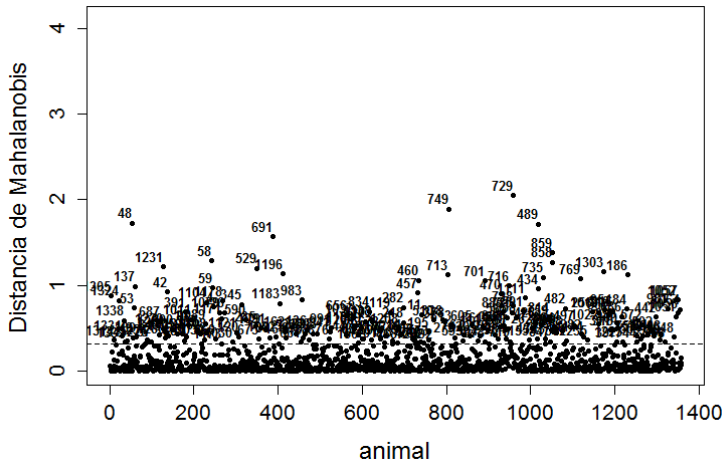

Figura 12 Distancia estandarizada de Mahalanobis vs índices de animal

En la figura 13 se observa que 7 rebaños pueden considerarse extremos. Se consideraron rebaños extremos a aquellos cuya distancia de Mahalanobis fue mayor a dos veces la media de las distancias estimadas para cada rebaño.

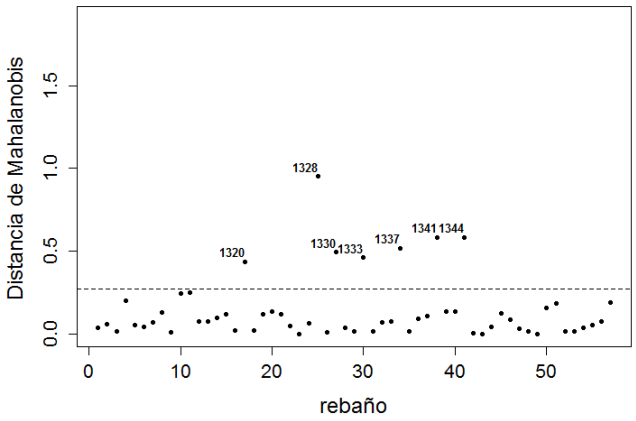

Figura 13 Distancia estandarizada de Mahalanobis vs índices de rebaño

El gráfico QQ-plot chi-cuadrado para la distancia de Mahalanobis para el efecto aleatorio de animal mostrado en la figura 14, muestra un comportamiento que ajusta a una distribución normal.

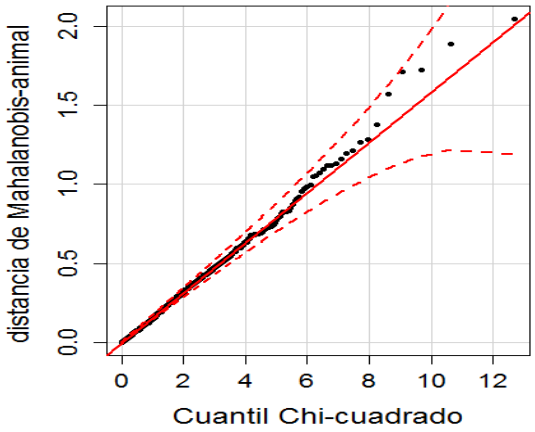

Figura 14 QQ plot chi-cuadrado para distancia estandarizada de Mahalanobis – animal

Sin embargo el gráfico QQ-plot chi-cuadrado para la distancia de Mahalanobis para el efecto aleatorio rebaño mostrado en la figura 15, muestra un comportamiento que podría no ajustar a una distribución normal.

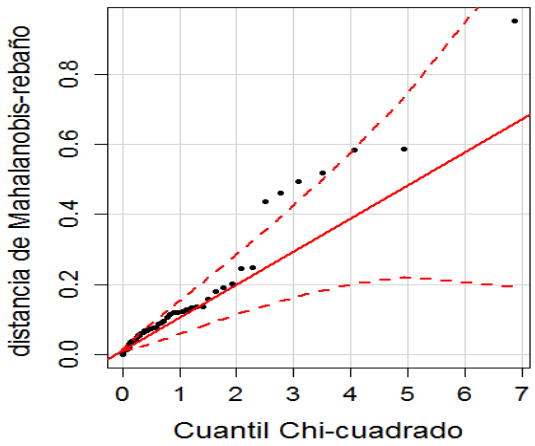

Figura 15 QQ plot chi-cuadrado para distancia estandarizada de Mahalanobis – rebaño

La figura 16 muestra las distancias estandarizadas de Lesaffre y Verbeke versus los animales, de este gráfico se tiene que 132 animales tiene distancias mayores a dos veces el valor de la distancia promedio estimada, por lo que para estos animales la estructura de varianzas y covarianza no sería muy adecuada .

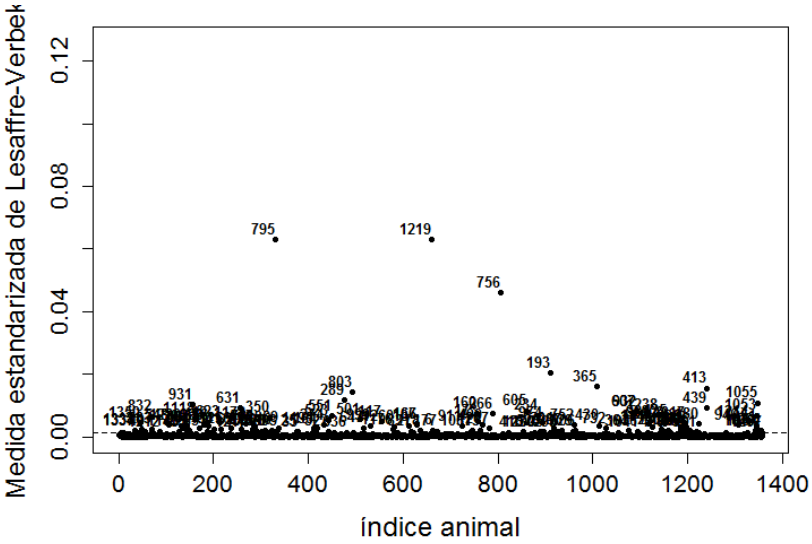

Figura 16 Medida estandarizada de Lesaffre-Verbeke vs animal

La figrua 17 muestra a través del gráfico QQ plot normal para los residuales estandarizados condicionales mínimos confundidos que éstos no tendrían un comportamiento normal, pese a que tienen una forma simétrica como muestra su respectivo histograma.

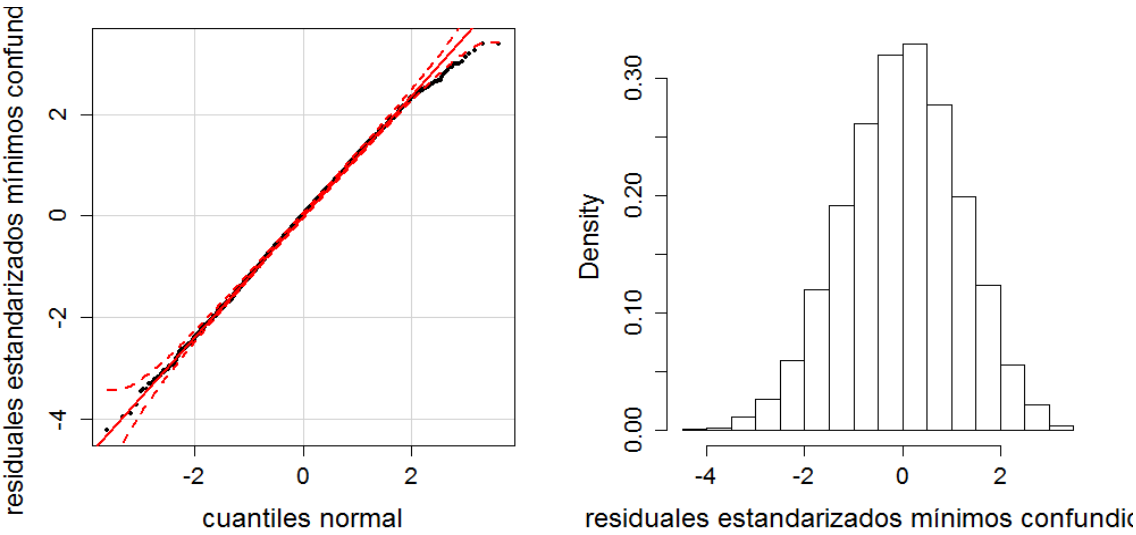

Figura 17 QQplot normal para los residuales estandarizados mínimos confundidos e histograma

4.4 Estimación del modelo utilizando inferencia bayesiana

Los estimados de las medidas de centralidad (media, mediana y moda) de las distribuciones posteriores para los componentes de variancia y la heredabilidad de ambas características fueron estimados utilizando la librería MCMCglmm de R, y son presentados en el Cuadro 7.

Cuadro 7 Estimados de los componentes de varianza y heredabilidad para la producción de leche

	σ_c^2	σ_h^2	σ_e^2	h^2	c^2
Media	0.3157	0.204	0.4828	0.3154	0.20162
Mediana	0.3154	0.19076	0.4824	0.3158	0.1979
Moda	0.3199	0.1876	0.482	0.3172	0.1925

σ_c^2, σ_h^2, σ_e^2, c^2 y h^2**:** varianza aditiva, varianza respecto al rebaño, varianza residual, heredabilidad aditiva y proporción de la varianza ambiental rebaño respectivamente

Los resultados muestran una heredabilidad similar con los presentados en el cuadro 5 vía REML, lo que indica una heredabilidad moderada de este carácter, asimismo el estimado del componente aleatorio debido al efecto rebaño fueron similares con ambas metodologías.

En la figura 18 se muestra las densidades marginales para los componentes del modelo, las cuales tiendes a ser simétricas (media, mediana y moda son bastante cercanas), y tienen un comportamiento prácticamente normal. Las trazas (convergencia) muestran un comportamiento aparentemente aleatorio para estos estimados de la distribución a posteriori.

La figura 18 permite verificar el comportamiento del algoritmo MCMC. En las gráficas de las trazas (evolución de las muestras a través de las iteraciones), lado izquierdo, no se observa ninguna tendencia en los nueve componentes.

Las correlaciones entre sucesivas muestras son bajas en casi todos los componentes del modelo, Cuadros 9 y 10, lo que podría indicar una fuerte convergencia de la cadena. Esto se refleja en los tamaños efectivos de la muestra no correlacionada que son altos para los componentes de interés, mayores a 1000 según lo mostrado en el Cuadro 8 que es lo mínimamente recomendado por Hadfield (2010). El componente que corresponde al error presenta una autocorrelación no muy baja, sin embargo, el error de Monte Carlo descrito en 2.7.6 es bajo, Cuadro 8, pero como el error de Monte Carlo está directamente relacionado con la inversa de la longitud de la cadena (o número de iteraciones del algoritmo descrito en 2.7.5), en definitiva este disminuirá cuando se aumente dicha longitud, en es este estudio no se corrió los modelos con mayor longitud de cadena, ni mayor número de cadenas, puesto que los resultados fueron similares a los obtenidos por REML y además las distribuciones posteriores de las características de interés mostraron ser bastante cercanas a la normal.

Cuadro 8 Tamaño efectivo muestral (TE) y error de Monte Carlo (EMC) de las distribuciones posteriores de la varianza genética y heredabilidad para los caracteres analizados

	σ_c^2	σ_h^2	σ_e^2	h^2	c^2
TE	2136	9000	4712	3125	9000
EMC	0.000576	0.00055	0.000244	0.000491	0.000432

σ_c^2, σ_h^2, σ_e^2, c^2 y h^2: varianza aditiva, varianza respecto al rebaño, varianza residual, heredabilidad aditiva y proporción de la varianza ambiental rebaño respectivamente

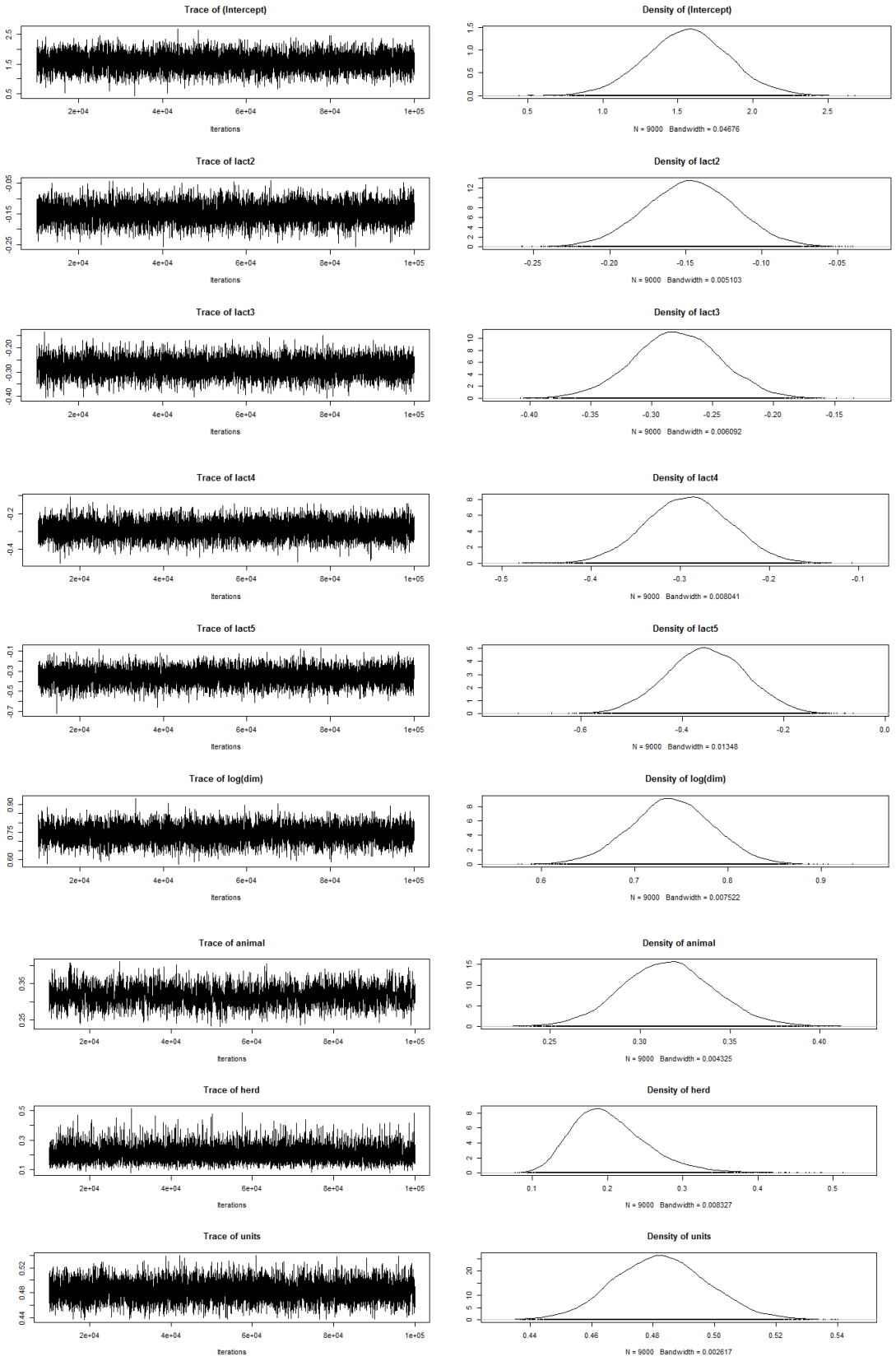

Figura 18 Evolución de los valores muestreados a lo largo de las iteraciones y las estimaciones de las funciones de densidades a posteriori para cada componente.

Cuadro 9 Autocorrelaciones de los componentes: número de lactación y logaritmo del número de días en leche

	(Intercept)	lact2	lact3	lact4	lact5	log(dim)
Lag 0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Lag 10	0.01198	-0.02787	-0.00425	-0.00056	-0.00991	0.00894
Lag 50	-0.00183	0.02923	0.00780	0.02572	-0.00552	0.00601
Lag 100	-0.00058	-0.01004	-0.01298	0.00434	-0.00859	-0.00202
Lag 500	-0.00336	-0.00314	-0.00982	0.00200	-0.00122	-0.01191

Cuadro 10 Autocorrelaciones de los componentes: animal, rebaño y error

	animal	herd	units
Lag 0	1.000000000	1.000000000	1.00000000
Lag 10	0.616322115	0.001095029	0.11517117
Lag 50	0.090100800	-0.013593962	0.03285455
Lag 100	0.022891093	-0.009120291	0.01713553
Lag 500	0.006656649	-0.004943959	0.01954783

En cuanto a la estimación de la heredabilidad del componente génetico aditivo animal en sentido amplio (Gutiérrez JP, 2010) son mostradas en el Cuadro 7 como h^2, y su densidad a posteriori y traza son mostrados en la figura 19, desde el cual no se observa problemas de convergencia y una distribución a posteriori simétrica. Además la heredabilidad de la producción está entre 0.264 y 0.365 al 95% de probabilidad.

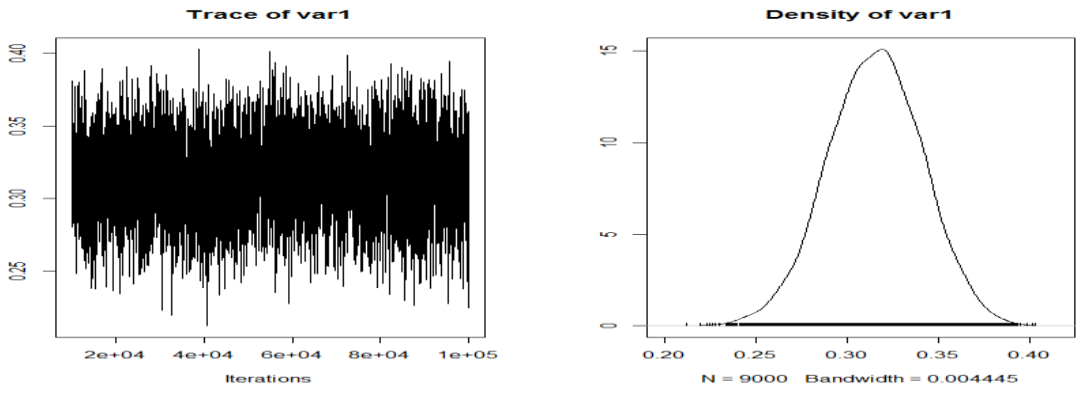

Figura 19 Evolución de los valores muestreados a lo largo de las iteraciones y las estimación de la función de densidad a posteriori para la heredabilidad

V. CONCLUSIONES

1. La estimación de los componentes de varianza fueron resultados similares con ambas metodologías, REML y bayesiano. El componente debido al efecto genético aditivo fue de 0.32 y el componente debido al efecto rebaño fue de 0.20.

2. En el diagnóstico del modelo vía REML, a partir de los gráficos de residuales, se observó linealidad de los efectos fijos del modelo, pero no se observó homocedasticidad de los residuales condicionales. Asimismo, se encontró que la estructura genética de parentesco, para las correlaciones entre individuos considerada en el modelo, no es adecuada para 132 animales evaluados. Además, se encontró hasta 234 observaciones, 215 animales y 7 rebaños con un comportamiento atípico.

3. En el diagnóstico del modelo vía REML, también se observó un comportamiento normal para el efecto aleatorio que corresponde al animal, pero no para el efecto aleatorio del rebaño, así como tampoco se observó normalidad para los errores condicionales. Debido a este análisis, las pruebas de hipótesis realizadas en el proceso de ajuste del modelo vía REML pierden validez, sin embargo se utilizó estos datos para ilustrar la metodología del análisis de residuales vía REML.

4. En el diagnóstico del modelo vía estimación bayesiana, no se encontró problemas de convergencia de la cadena Se obtuvieron errores de Montecarlo bajos y tamaños efectivos de muestra mayores a 1000 para cada componente del modelo.

5. La media estimada de la heredabilidad vía muestreo de Gibbs fue de 0.3154 y resultó similar a la estimada vía REML de 0.3177 para la producción de leche. Esto es debido a que los estimadores REML coinciden con la moda de la densidad

6. posterior marginal cuando se asume a priori constante para los efectos fijos y normalidad para los efectos aleatorios.

7. La magnitud de la heredabilidad estimada es moderada, lo que respalda la idea de que la variabilidad fenotípica de esta característica (producción de leche) está explicada en aproximadamente el 32% por la acción genética aditiva y el resto por otros factores.

VI. RECOMENDACIONES

- Considerar utilizar un modelo lineal mixto que incorpore heterogeneidad de los errores condicionales.
- Podría recomendarse escribir código en R que permita realizar el análisis de influencia del modelo lineal mixto incorporando estructura genealógica.
- Podría recomendarse utilizar procedimientos que tengan como respuesta no solo una distribución normal simétrica, sino otras distribuciones tanto para la metodología bayesiana como para REML.

VII REFERENCIAS BIBLIOGRAFÍCAS

Bates, D; Pinheiro J. 2000. Mixed-Effect Models in S and S-PLUS. New York. Springer-Verlag. 528 p.

Bates, D; DebRoy, S. 2004. Linear mixed models and penalized least squares. Journal of Multivariate Analysis 91 (1): 1-17.

Bates, D. 2010. lme4: Mixed-effects modeling with R. New York. Springer. 145 p. Consultado 15 febrero 2013. Disponible en http://lme4.r-forge.r-project.org/book/

Bates, D. 2014. Computational methods for mixed models (en línea). Department of Statistics. Consultado 24 marzo 2015. Disponible en https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf

Bates D; Machler M; Bolker M; Walker S. 2015. Fitting linear mixed-models using lme4. Journal of Statistical Software. 67 (1): 1-48. Consultado 15 febrero 2016. Disponible en https://www.jstatsoft.org/article/view/v067i01/0

Blasco, A. 2001. The Bayesian controversy in animal inbreeding. Journal of Animal Science. 79 (8): 2023–2046.

De Villemereuil, P. 2012. Estimation of a biological trait heritability using the animal model. How to use the MCMCglmm R package. (en línea, blog). Auckland, New Zealand. Consultado 24 de febrero 2014. Disponible en http://devillemereuil.legtux.org/wp-content/uploads/2012/12/tuto_en.pdf

Elzo, M; Vergara, O. 2012 Modelación aplicada a las ciencias genéticas: II. Evaluaciones Genéticas. Medellín, Colombia, Biogénesis. 134 p.

Gianola, D. 2000. Statistics in animal breeding. Journal of the American Statistical Association. 95 (449): 296-299

Gelman, A; Carlin, J; Stern, S; Dunson, D; Vehtari, A; Rubin, D. 2014. Bayesian Data Analysis. 3 ed. Chapman & Hall/CRC Press. Londres. 675 p.

Gilmour, AR; Gogel, BJ; Cullis, BR; Thompson, R. 2009. ASReml User Guide Release 3.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, Reino Unido. 398 p. Consultado 5 enero 2013. Disponible en http://galwey.genstat.co.uk/downloads/asreml/release3/UserGuide.pdf

Griffiths, AJF; Miller, JH; Suzuki DT, Lewontin, R y Gelbart, W. 2000. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Disponible en https://www.ncbi.nlm.nih.gov/books/NBK21766/

Gutierrez, J.P. 2010. Iniciación a la valoración genética animal. Metodología adaptada al EEES. Madrid, España. UCM Editorial Complutense. 368 p.

Hadfield, JD 2010. MCMC Methods for Multi–response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software, 33(2): 1-22. Consultado 13 julio 2014. Disponible en: https://www.jstatsoft.org/article/view/v033i02/v33i02.pdf

Hilden-Milden, J. 1995. Multilevel Diagnostics for Mixed and Hierarchical Linear Models. Tesis Ph.D. Los Angeles, Estados Unidos. University of California. 120 p.

Jiang, J. 2007. Linear and Generalized Linear Mixed Models and Their Applications. New York. Springer-Verlag. 257p.

León, M. 2004 Métodos de estimación de componentes de varianza en poblaciones. Una Reseña histórica. Revista Computadorizada de Producción Porcina 11 (1): 23-37. Consultado 5 enero 2013. Disponible en http://www.academia.edu/9249500/M%C3%89TODOS_DE_ESTIMACI%C3%93N_DE_

COMPONENTES_DE_VARIANZA_EN_POBLACIONES._UNA_RESE%C3%91A_HIST%C3%93RICA

Misztal I. 2008. Reliable computing in estimation of variance. Journal of Animal Breeding and Genetics 125 (6):363-370.

Montaldo, V; Barria N. 1998. Mejoramiento genético de animales. Ciencia al Día. 1 (2):1-19. Consultado 10 enero 2013. Disponible en http://www.ciencia.cl/CienciaAlDia/volumen1/numero2/articulos/cad-2-3.pdf

Mrode, R. 2014. Linear Model for the Prediction of Animal Breeding Value. 3 ed. Edinburgh, UK, CABI. 343 p.

Nobre, JS; Singer, JM. 2007. Residual Analysis for Linear Mixed Models. Biometrical Journal 49 (6): 863-875

Ntzoufras, I. 2009. Modeling using WinBUGS. New Yersey. Estados Unidos, John Willey & Sons. 520 p.

Patterson, H D; Thompson, R 1971. Recovery of inter-block information when block sizes are unequal. Biometrika. 58 (3): 545-554.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing (en línea, programa informático). Vienna, Austria. Consultado 20 enero 2013. Disponible en http://www.R-project.org/

Robinson, GK. 1991. That BLUP is a Good Thing: The Estimation of Random Effects. Statist. Sci. 6 (1): 15-32.

Searle, S; Casella, G; McCulloch C. 1992. Variance Components. New York. Estados Unidos. Wiley series in probability and statisticis. 501p.

Sorensen, D; Gianola, D. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. New York. Estados Unidos, Springer-Verlag. 740 p.

Singer, JM; Nobre JS; Rocha, F. Diagnostic and treatment for linear mixed models. Session CPS203 Proceedings of the ISI World Statistics Congress (59, 2013, Hong Kong, República popular China). Hong Kong, República popular China. 5486 p.

Thompson, R; Brotherstone, S; M.S. White, I. 2005. Estimation of quantitative genetic parameters. Philos Trans R Soc Lond B Biol Sci 360 (1459):1469-1477. Consultado 23 marzo 2013. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569516/

Vazquez, AI; Bates, D; Rosa GJ, Gianola, D; Weigel KA. 2010. Technical note: An R package for fitting generalized linear mixed models in animal breeding1. Journal of Animal Science. 88 (2): 497-504. Consultado 20 enero 2013. Disponible en https://www.animalsciencepublications.org/publications/jas/abstracts/88/2/497

Wang, D; Rutledge, J; Gianola, D. 1993. Marginal inferences about variance components in a mixed linear models using Gibbs sampling. Genetics Selection Evolution, BioMed Central, 1993, 25 (1), pp.41-62. Consultado 5 enero 2013. Disponible en https://hal.archives-ouvertes.fr/hal-00893980/document

Wang, D; Rutledge, J; Gianola, D. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genetics Selection Evolution, BioMed Central, 26 (2) : 91-115. Consultado 5 enero 2013. Disponible en https://hal.archives-ouvertes.fr/hal-00894021/document

VIII ANEXO

Script en R

```
library(pedigreemm)
library(lme4)
library(ggplot2)
library(HLMdiag)
data(milk)
data(pedCows)
milk$lact<-as.factor(milk$lact)

###ANÁLISIS DESCRIPTIVO UNIVARIADO Y BIVARIADO####
qplot(milk$milk, geom="histogram", xlab = "producción a 305 días",ylab="número de lactaciones", main="Producción de leche a 305 días por lactación")
library(psych)
describe(milk$milk)
qplot(milk$lact, geom="bar",xlab = "Lactación",ylab="número de lactaciones", main="Número de lactación")
qplot(milk$dim, geom="histogram", xlab = "número de días en leche",ylab="número de lactaciones", main="Número de días en leche")
qplot(milk$herd, geom="bar",xlab = "número de rebaño",ylab="número de lactaciones", main=" Distribución de frecuencias de las lactaciones, según el rebaño")

qplot(dim, milk, data=milk,xlab="Días en leche", ylab="producción de leche")
qplot(lact, milk, data=milk,geom=c("boxplot"),fill=lact, main="producción por lactación", xlab="Número de lactación", ylab="Producción")+theme(legend.position="none")
g2<-qplot(herd, milk, data=milk, geom=c("boxplot"),fill=herd,main="producción por rebaño",xlab="Rebaño", ylab="Producción")
g2+theme(legend.position="none")

####Construyendo el modelo inferencial clásico
milk <- within(milk, sdMilk <- milk / sd(milk))
fm1 <- pedigreemm(sdMilk ~ lact + log(dim) + (1|id) + (1|herd), na.action=na.omit,data = milk, pedigree = list(id = pedCows))


#####Función para obtener gráficos de probabilidad normal con bandas de confianza, obtenido de:
##http://stackoverflow.com/questions/4357031/qqnorm-and-qqline-in-ggplot2
```

```
gg_qq <- function(x, distribution = "norm", ..., line.estimate =
NULL, conf = 0.95,
                  labels = names(x)){
  q.function <- eval(parse(text = paste0("q", distribution)))
  d.function <- eval(parse(text = paste0("d", distribution)))
  x <- na.omit(x)
  ord <- order(x)
  n <- length(x)
  P <- ppoints(length(x))
  df <- data.frame(ord.x = x[ord], z = q.function(P, ...))

  if(is.null(line.estimate)){
    Q.x <- quantile(df$ord.x, c(0.25, 0.75))
    Q.z <- q.function(c(0.25, 0.75), ...)
    b <- diff(Q.x)/diff(Q.z)
    coef <- c(Q.x[1] - b * Q.z[1], b)
  } else {
    coef <- coef(line.estimate(ord.x ~ z))
  }

  zz <- qnorm(1 - (1 - conf)/2)
  SE <- (coef[2]/d.function(df$z)) * sqrt(P * (1 - P)/n)
  fit.value <- coef[1] + coef[2] * df$z
  df$upper <- fit.value + zz * SE
  df$lower <- fit.value - zz * SE

  if(!is.null(labels)){
    df$label <- ifelse(df$ord.x > df$upper | df$ord.x < df$lower,
labels[ord],"")
  }

  p <- ggplot(df, aes(x=z, y=ord.x)) +
    geom_point() +
    geom_abline(intercept = coef[1], slope = coef[2]) +
    geom_ribbon(aes(ymin = lower, ymax = upper), alpha=0.2)
  if(!is.null(labels)) p <- p + geom_text( aes(label = label))
  print(p)
  coef
}

### Obteniendo elementos del modelo
y<-as.vector(getME(fm1,"y"))
X<-as.matrix(getME(fm1,"X"))
Z<-as.matrix(getME(fm1,"Z"))
N<-length(y)                              # Número de observaciones
id<- as.numeric(getME(fm1,"flist")[[1]])
subject<-as.numeric(unique(id))
n<-length(as.numeric(names(table(id))))#Numero   de   unidades   o
sujetos(animales)
vecni<-(table(id))                          # vector con número de
observaciones por unidad
p<-ncol(X)                                  #Número de parámetros de
efectos fijos

md<-  (VarCorr(fm1))              # Estimación de los componentes
aleatorios (varianzas) en una matriz
```

```
if(length(md)==1)             # Para un solo componente aleatorio
{
  g<-as.matrix(md[[1]])
}
if(length(md)>1)                      #  Para  más  de  un  componente
aleatorio
{
  g<-as.matrix(md[[1]])
  for(i in 2:(length(md)))
  {
    g<- as.matrix(bdiag(g,as.matrix(md[[(i)]])))
  }
}
q<-dim(g)[1]                #número de componentes aleatorios

# Función para extraer la raíz cuadrada de una matrz
sqrt.matrix <- function(mat) {
  mat <- as.matrix(mat)  # new line of code
  singular_dec<-svd(mat,LINPACK=F)
  U<-singular_dec$u
  V<-singular_dec$v
  D<-diag(singular_dec$d)
  sqrtmatrix<-U%*%sqrt(D)%*%t(V)
  #  return(list(sqrt=sqrtmatrix))
}

### Calculando los residuales, estandarizados marginales
mr<-fm1@resp$y - lme4::getME(fm1, "X") %*% lme4::fixef(fm1)
sig0 <- lme4::getME(fm1, "sigma")
ZZt <- crossprod( lme4::getME(fm1, "A") )
m   <- nrow(ZZt)
R      <- Diagonal( n = m, x = sig0^2 )
V      <- R + ZZt
V.mr<-V-X%*%solve(t(X)%*%solve(V)%*%X)%*%t(X)

#  mrs<-mr/sqrt(diag(v.mr))
mrs<-matrix(0,N,1)
auxni=as.vector(vecni)
for (t in 1:n){
  li<- sum(vecni[1:t-1])+1
  ls<- sum(vecni[1:t])
  if(auxni[t]==1){
    auxr2 <- solve(sqrt(V.mr[li:ls,li:ls]))
    mrsi<-(auxr2)%*%mr[li:ls]
    mrs[li:ls,]<- mrsi
  }
  else
  {
    auxr2 <- solve(sqrt.matrix(V.mr[li:ls,li:ls]))
    mrsi<- auxr2%*%mr[li:ls]
    mrs[li:ls,]<- mrsi
  }
}


#shapiro.test(mrs)
```

```
#gg_qq(mrs,conf = 0.90)

##Diagnóstico para linealidad de los efectos fijos: con covariable
qplot(        log(milk$dim),mrs,ylab="Residuales        marginales
estandarizados",xlab="Logaritmo de días en leche")

##Diagnóstico para linealidad de los efectos fijos:con valores
ajustados
pred<-lme4::getME(fm1, "X") %*% lme4::fixef(fm1)
limit=2
par(mfrow=c(1,2), mar=c(14, 5, 1, 2))
plot(pred,    mrs,    xlab=expression(paste("Valores    marginales
ajustados")),
     ylab=expression(paste("Residuales                    marginales
estandarizados")),  pch=20,  cex=1.0,  cex.lab=1.5,  cex.axis=1.3,
ylim=c(-1.3*max(abs(range(mrs))),1.3*max(abs(range(mrs)))))
abline(h=0, lty=2)
abline(h=-limit, lty=2)
abline(h=limit, lty=2)
index=which(abs(mrs)>limit)
if(length(index)>0)
{
  text(pred[index],      mrs[index],      paste(id[index],sep="."),
adj=c(1,-.5), cex=.8, font=2)
}
hist(mrs,   freq=F,   main="",   xlab=expression(paste("Residuales
marginales estandarizados")), cex=0.9, cex.lab=1.5, cex.axis=1.3)


##Prsencia de observaciones outlying:
##plot(mrs,ylab="Residuales                                    marginales
estandarizados",xlab="índice de observación")
par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
plot(mrs,       ylab=expression(paste("Residuales       marginales
estandarizados")), xlab="índice de observación",
     pch=20,     ylim=c(0,2*max(mrs)),     cex=1.0,     cex.lab=1.5,
cex.axis=1.3)
abline(h=2*mean(mrs), lty=2)
index=which(abs(mrs)>2)
index1<-subject[index]
if(length(index)>0)
{
  text(index,mrs[index], index1, adj=c(1,-.5), cex=.8, font=2)
}


### Calculando los residuales condicionales, estandarizados (nivel
1)
rc<-resid(fm1)

Y <- fm1@resp$y
X <- lme4::getME(fm1, "X")
m <- length(Y)
# Construyendo V = Cov(Y)
sig0 <- lme4::getME(fm1, "sigma")
ZZt <- crossprod( lme4::getME(fm1, "A") )
```

```
R    <- Diagonal( n = m, x = sig0^2 )
V    <- R + ZZt
# Invirtiendo V
V.chol <- chol( V )
Vinv   <- chol2inv( V.chol)
# Calculando la Varianza de los residuales condicionales
XVXinv<- solve( t(X) %*% Vinv %*% X )
VinvX <- Vinv %*% X
V.rc<- R%*%(Vinv-VinvX %*% XVXinv %*% t(VinvX))%*%R

rcs<-matrix(0,N,1)

for (t in 1:n){
  li<- sum(vecni[1:t-1])+1
  ls<- sum(vecni[1:t])
  if(auxni[t]==1){
    auxr2 <- solve(sqrt(V.rc[li:ls,li:ls]))
    rcsi<-(auxr2)%*%rc[li:ls]
    rcs[li:ls,]<- rcsi
  }
  else
  {
    auxr2 <- solve(sqrt.matrix(V.rc[li:ls,li:ls]))
    rcsi<- auxr2%*%rc[li:ls]
    rcs[li:ls,]<- rcsi
  }
}

#shapiro.test(rcs)
#gg_qq(rcs,conf = 0.90)

##Presencia de observaciones outlying
##plot(rcs,ylab="Residuales                              condicionales
estandarizados",xlab="Índice de observación")
par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
plot(rcs,      ylab=expression(paste("Residuales      condicionales
estandarizados")), xlab="índice de observación",
     pch=20,     ylim=c(0,2*max(rcs)),     cex=1.0,     cex.lab=1.5,
cex.axis=1.3)
abline(h=2*mean(rcs), lty=2)
index=which(abs(rcs)>2)
index1<-subject[index]
if(length(index)>0)
{
  text(index,rcs[index], index1, adj=c(1,-.5), cex=.8, font=2)
}


##Homocedasticidad de los errores condicionales
##qplot(predict(fm1),rcs,ylab="Residuales            condicionales
estandarizados",xlab="Ajustados")
predi=predict(fm1)
limit=2
par(mfrow=c(1,2))
plot(predi,   rcs,   xlab=expression(paste("Valores   condicionales
ajustados")),
```

```
     cex=1.0,              cex.lab=1.5,              cex.axis=1.3,
ylab=expression(paste("residuales condicionales estandarizados")),
pch=20,                                                  ylim=c(-
1.3*max(abs(range(rcs))),1.3*max(abs(range(rcs)))))
abline(h=0, lty=2)
abline(h=-limit, lty=2)
abline(h=limit, lty=2)
index=which(abs(rcs)>limit)
if(length(index)>0)
{
  text(predi[index],     rcs[index],     paste(id[index],sep="."),
adj=c(1,-.5), cex=.8, font=2)
}
hist(rcs,   freq=F,   main="",   xlab=expression(paste("Residuales
condicionales     estandarizados")),     cex=0.9,     cex.lab=1.5,
cex.axis=1.3)


### Calculando los efectos para los dos efectos aleatorios
estandarizados (nivel 2)
re <- lme4::ranef(fm1)
library(arm)
se.re <- se.ranef(fm1)
resid<-re$id/se.re$id ##residual para individuo estandarizado
resherd<-re$herd/se.re$herd   ##residual   para   rebaño   (herd)
estandarizado

#shapiro.test(resid$`(Intercept)`)
#gg_qq(resid$`(Intercept)`,conf = 0.90)

#shapiro.test(resherd$`(Intercept)`)
#gg_qq(resherd$`(Intercept)`,conf = 0.90)

#distancia de Mahalanobis
aux=t(Z)%*%(Vinv-VinvX %*% XVXinv %*% t(VinvX))%*%Z
library(MASS)

##Distancia de Mahalanobis estandarizada para efecto individuo
auxId<-aux[1:1359,1:1359]
qm<-0
l<-1
dm<-matrix(0,n,1)
reId<-as.matrix(re[[1]],n,1)

for(j in 1:n)
{
    gbi<-auxId[j,j]
    rei<-reId[(l*j-qm):(l*j)]
    dmi<- t(rei)%*%ginv(gbi)%*%rei
    dm[j]<-dmi
  }
dmId=dm

##Presencia de sujetos (animal o individuos) outlying
#qplot(subject,dmId,ylab="distancia     de     Mahalanobis     para
animal",xlab="animal")
```

```
par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
plot(dmId,  ylab=expression(paste("Distancia  de  Mahalanobis")),
xlab="animal",
     pch=20,    ylim=c(0,2*max(dmId)),    cex=1.0,    cex.lab=1.5,
cex.axis=1.3)
abline(h=2*mean(dmId), lty=2)
index=which(dmId>2*mean(dmId))
index1<-subject[index]
if(length(index)>0)
{
  text(index,dmId[index], index1, adj=c(1,-.5), cex=.8, font=2)
}


##Distancia de Mahalanobis para efecto rebaño
R<- as.numeric(getME(fm1,"flist")[[2]])
rebanho<-as.numeric(unique(R))

auxR<-aux[1360:1416,1360:1416]
qm<-0
l<-1
dm<-matrix(0,57,1)
reR<-as.matrix(re[[2]],57,1)

for(j in 1:57)
{

  gbi<-auxId[j,j]
  rei<-reR[(l*j-qm):(l*j)]
  dmi<- t(rei)%*%ginv(gbi)%*%rei
  dm[j]<-dmi
}
dmR=dm
##Prsencia de sujetos (animal o individuos) outlying
#qplot(rebanho,dmR,ylab="distancia     de     Mahalanobis     para
rebaño",xlab="rebaño")
par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
plot(dmR,  ylab=expression(paste("Distancia  de  Mahalanobis")),
xlab="rebaño",
     pch=20,    ylim=c(0,2*max(dmR)),    cex=1.0,    cex.lab=1.5,
cex.axis=1.3)
abline(h=2*mean(dmR), lty=2)
index=which(dmR>2*mean(dmR))
index1<-subject[index]
if(length(index)>0)
{
  text(index,dmR[index], index1, adj=c(1,-.5), cex=.8, font=2)
}

## QQplot2

qqPlot2<-function(x,           distribution="norm",           ...,
ylab=deparse(substitute(x)),
                  xlab=paste(distribution,           "quantiles"),
main=NULL, las=par("las"),
                  envelope=.95,
```

```
                  col=palette()[1], col.lines=palette()[2], lwd=2,
pch=1, cex=par("cex"),
                  cex.lab=par("cex.lab"),cex.axis=par("cex.axis"),
                  line=c("quartiles", "robust", "none"),
                  labels = if(!is.null(names(x))) names(x) else
seq(along=x),
                  id.method = "y",
                  id.n = if(id.method[1]=="identify") Inf else 0,
                  id.cex=1, id.col=palette()[1], grid=TRUE)
{
  line <- match.arg(line)
  good <- !is.na(x)
  ord <- order(x[good])
  ord.x <- x[good][ord]
  ord.lab <- labels[good][ord]
  q.function <- eval(parse(text=paste("q", distribution, sep="")))
  d.function <- eval(parse(text=paste("d", distribution, sep="")))
  n <- length(ord.x)
  P <- ppoints(n)
  z <- q.function(P, ...)
  plot(z, ord.x, type="n", xlab=xlab, ylab=ylab, main=main,
las=las,cex.lab=cex.lab, cex.axis=cex.axis)
  if(grid){
    grid(lty=1, equilogs=FALSE)
    box()}
  points(z, ord.x, col=col, pch=pch, cex=cex)
  if (line == "quartiles" || line == "none"){
    Q.x <- quantile(ord.x, c(.25,.75))
    Q.z <- q.function(c(.25,.75), ...)
    b <- (Q.x[2] - Q.x[1])/(Q.z[2] - Q.z[1])
    a <- Q.x[1] - b*Q.z[1]
    abline(a, b, col=col.lines, lwd=lwd)
  }
  if (line=="robust") {
    coef <- coef(rlm(ord.x ~ z))
    a <- coef[1]
    b <- coef[2]
    abline(a, b)
  }
  conf <- if (envelope == FALSE) .95 else envelope
  zz <- qnorm(1 - (1 - conf)/2)
  SE <- (b/d.function(z, ...))*sqrt(P*(1 - P)/n)
  fit.value <- a + b*z
  upper <- fit.value + zz*SE
  lower <- fit.value - zz*SE
  if (envelope != FALSE) {
    lines(z, upper, lty=2, lwd=lwd, col=col.lines)
    lines(z, lower, lty=2, lwd=lwd, col=col.lines)
  }
  labels(z, ord.x, labels=ord.lab,
              id.method = id.method, id.n = id.n, id.cex=id.cex,
id.col=id.col)
}

## qqplot chi-cuadrado para la distancia de Mahalanobis para
rebaño- diagnóstico de normalidad
```

```
#qqplot(qchisq(ppoints(500), df = 1), dmId,main = expression("Q-Q
para la distancia de Mhalanobis-animal"),xlab="cuantil chi-
cuadrado gl=1 ",ylab="distancia de Mahalanobis para animal")
#qqline(dmId, distribution = function(p) qchisq(p, df = 1),prob =
c(0.1, 0.6), col = 2)

par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
quant.chisq<-qqPlot2(dmId, distribution='chisq', df=1, pch=20,
cex=1.0,cex.lab=1.5,cex.axis=1.3,
                     ylab=expression(paste("distancia de
Mahalanobis-rebaño")),xlab="Cuantil Chi-cuadrado")



## qqplot chi-cuadrado para la distancia de Mahalanobis para
sujeto (animal)
#qqplot(qchisq(ppoints(500), df = 1), dmR,main = expression("Q-Q
plot para la distancia de Mhalanobis-rebaño"),xlab="cuantil chi-
cuadrado gl=1 ",ylab="distancia de Mahalanobis para rebaño")
#qqline(dmR, distribution = function(p) qchisq(p, df = 1), prob =
c(0.1, 0.6), col = 2)

par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
quant.chisq<-qqPlot2(dmR, distribution='chisq', df=1, pch=20,
cex=1.0,cex.lab=1.5,cex.axis=1.3,
                     ylab=expression(paste("distancia de
Mahalanobis-animal")),xlab="Cuantil Chi-cuadrado")

##Obteniendo los residuales mínimos confundidos estandarizados
Q<-Vinv-VinvX %*% XVXinv %*% t(VinvX)
R    <- Diagonal( n = m, x = sig0^2 )
auxqn<-eigen(( sig0*sig0 *Q), symmetric = T, only.values = FALSE,
EISPACK = F)
lt<-t(sqrt(solve(diag((auxqn$values[1:(N-p)]))))                %*%
t(auxqn$vectors[1:(N-p),1:(N-p)])  %*%  sqrt(solve(R[1:(N-p),1:(N-
p)])) )
var.resmc<- lt %*% V.rc[1:(N-p),1:(N-p)] %*% t(lt)
resmcp<- (lt %*% rc[1:(N-p)] )/sqrt(diag(var.resmc))

##qqplot de normalidad para diagnosticar normalidad de los errores
condicionales
##qqnorm(resmcp)
##qqline(resmcp)
resmcp<-as.numeric(resmcp@x)
par(mfrow=c(1,2),mar=c(14, 5, 1, 2))
qqPlot2(resmcp,    ylab="residuales    estandarizados    mínimos
confundidos",   xlab="cuantiles   normal",   pch=20,   cex=0.75,
cex.lab=1.5,cex.axis=1.3)
hist(resmcp,  freq=F,  xlab="residuales  estandarizados  mínimos
confundidos", main="", cex=1.0, cex.lab=1.5, cex.axis=1.3, pch=20)


# Medida estandarizada de Lesaffre-Verbeke
lesverb<- rep(0,n)
auxni=as.vector(vecni)
for (t in 1:n){
```

```
  li<- sum(vecni[1:t-1])+1
  ls<- sum(vecni[1:t])
  if(vecni[t]==1){
    auxr2 <- solve(sqrt(V.mr[li:ls,li:ls]))
    Ri<-(auxr2)%*%mr[li:ls]
    auxt<- diag(vecni[t])-Ri%*%t(Ri)
    lesverb[t]<- sum(diag(auxt%*%t(auxt)))
  }
  else
  {
    auxr2 <- solve(sqrt.matrix(V.mr[li:ls,li:ls]))
    Ri<- auxr2%*%mr[li:ls]
    auxt<- diag(vecni[t])-Ri%*%t(Ri)
    lesverb[t]<- sum(diag(auxt%*%t(auxt)))
  }
}
lesverbp<- lesverb/sum(lesverb)

##Plot para diagnóstico de la matriz de covarianza dentre de las
unidades (animales)
##plot(lesverbp,ylab="medida de Lesaffre y Verbeke",xlab="Índice
de observación")
par(mfrow=c(1,1),mar=c(14, 5, 1, 2))
plot(lesverbp,ylab=expression(paste("Medida estandarizada de
Lesaffre-Verbeke")),
     xlab="índice animal", cex=1.0, cex.lab=1.5, cex.axis=1.3,
pch=20, ylim=c(0,2*max(abs(range(lesverbp)))))
abline(h=2*mean(lesverbp),lty=2)
index=which(lesverbp>2*mean(lesverbp))
index1<-subject[index]
if(length(index)>0)
{
  text(index, lesverbp[index], index1, adj=c(1,-.5), cex=.8,
font=2)
}

##Probando significancia de los efectos aleatorios:
###Efecto individuo
fm2 <- pedigreemm(sdMilk ~ lact + log(dim) +(1|herd),
na.action=na.omit,data = milk)
1-pchisq(2*(logLik(fm1)-logLik(fm2)),1)
###Efecto rebaño
fm3 <- pedigreemm(sdMilk ~ lact + log(dim) + (1|id),
na.action=na.omit,data = milk, pedigree = list(id = pedCows))
1-pchisq(2*(logLik(fm1)-logLik(fm3)),1)

##Medidas de ajustes:
logLik(fm1)
logLik(fm2)
logLik(fm3)
AIC(fm1)
AIC(fm2)
AIC(fm3)
BIC(fm1)
BIC(fm2)
BIC(fm3)
```

```
########################################PARTE
BAYESIANA#########################################

####Construyendo el modelo bayesiano
##Ilustrativo con librería MCMCglmm para peso al nacimiento

##Para trabajar con GeneticsPed se debe descargar desde
bioconductor con:

animal<-as.numeric(pedCows@label)
sire<-pedCows@sire
dam<-pedCows@dam
ped<-cbind(animal,sire,dam)

milk=edit(milk) ##colocar animal en vez de id

prior=list(R=list(V=1,nu=0.002),G=list(G1=list(V=1,nu=0.002),G2=li
st(V=1,nu=0.002)))
model <- MCMCglmm(sdMilk ~ lact + log(dim),random= ~ animal +
herd, family = "gaussian",prior = prior, pedigree = ped, data =
milk, nitt = 100000,burnin = 10000, thin = 10)


library(mcmcse)
plot(model$Sol)
plot(model$VCV)
autocorr.diag(model$Sol)
autocorr.diag(model$VCV)
effectiveSize(model$Sol)
effectiveSize(model$VCV)
heidel.diag(model$VCV)
heidel.diag(model$Sol)
summary(model)
HPDinterval(model$VCV[, "animal"])
HPDinterval(model$VCV[, "herd"])
HPDinterval(model$VCV[, "units"])

herit <- model$VCV[, "animal"]/(model$VCV[, "animal"] +
model$VCV[, "units"]+model$VCV[, "herd"])

effectiveSize(herit)
mean(herit)
HPDinterval(herit)
plot(herit)

rebanho <- model$VCV[, "herd"]/(model$VCV[, "animal"] +
model$VCV[, "units"]+model$VCV[, "herd"])
effectiveSize(rebanho)
mean(rebanho)
HPDinterval(rebanho)

#ess(model$VCV[, "CA"], imse = TRUE, verbose = FALSE)
#ees(herit, size = "sqroot", g = NULL, method = "bm", warn =
FALSE)

mcse(herit, size = "sqroot", g = NULL, method = "bm", warn = TRUE)
```

```
mcse(rebanho, size = "sqroot", g = NULL, method = "bm", warn =
TRUE)
mcse(model$VCV[, "units"], size = "sqroot", g = NULL, method =
"bm", warn = TRUE)
mcse(model$VCV[, "herd"], size = "sqroot", g = NULL, method =
"bm", warn = TRUE)
mcse(model$VCV[, "animal"], size = "sqroot", g = NULL, method =
"bm", warn = TRUE)

estimate_mode <- function(x){
  d <- density(x)
  d$x[which.max(d$y)]}

estimate_mode(herit)
estimate_mode(rebanho)
estimate_mode(model$VCV[, "units"])
estimate_mode(model$VCV[, "herd"])
estimate_mode(model$VCV[, "animal"])
```