Selección y estimación de parámetros genéticos en bovinos lecheros

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires, Área Ciencias Agropecuarias

Juan David Corrales Alvarez
Zootecnista, Facultad de Ciencias Agrarias, Universidad de Antioquia
- 2007 -

Lugar de trabajo: Facultad de Agronomía, Universidad de Buenos Aires

Escuela para Graduados Ing. Agr. Alberto Soriano
Facultad de Agronomía – Universidad de Buenos Aires

COMITÉ CONSEJERO

Director de tesis
Rodolfo Juan Carlos Cantet
Ingeniero Agrónomo., Universidad de Buenos Aires, Argentina.
MSc. Montana State University, Estados Unidos de América.
MSc. University of Illinois, Estados Unidos de América.
Ph.D., University of Illinois, Estados Unidos de América.

Consejero de Estudios
Sebastián Munilla Leguizamón
Ingeniero Agrónomo., Universidad de Buenos Aires, Argentina.
Doctor., Universidad de Buenos Aires, Argentina.

JURADO DE TESIS

Director de tesis
Rodolfo Juan Carlos Cantet
Ingeniero Agrónomo., Universidad de Buenos Aires, Argentina.
MSc. Montana State University, Estados Unidos de América.
MSc. University of Illinois, Estados Unidos de América.
Ph.D., University of Illinois, Estados Unidos de América.

JURADO
Manuel Baselga Izquierdo
Ingeniero Agrónomo (Universidad Politecnica de Valencia)
Ph.D (Universidad Politecnica de Valencia)

JURADO
Ignacio Aguilar Garcia
Ingeniero Agrónomo (Universidad de la República, Uruguay)
MSc (University of Georgia, Estados Unidos de América)
PhD (University of Georgia, Estados Unidos de América)

JURADO
Sergio Jorge Bramardi
Ingeniero Agrónomo (Universidad Nacional del Comahue)
Magiser Scientiae en Biometría (Universidad de Buenos Aires)
Doctor Programa de Estadística (Universidad Politécnica de Valencia)

Fecha de defensa de la tesis: 13 de Abril de 2016

A mis padres, hermanos y sobrinos que son mi gran motivación personal. A todos los del grupo MGA de la Universidad de Buenos Aires que siempre me apoyaron.

AGRADECIMIENTOS

Esta tesis fue posible gracias a todas aquellas personas e instituciones con las cuales tuve la oportunidad de compartir y aprender. En primer lugar, quiero agradecer al Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS), entidad que me brindo el apoyo financiero para poder realizar el doctorado por fuera de mi país de origen (Colombia).

Agradezco a mi director de tesis Dr. Rodolfo Juan Carlos Cantet (Fito), que me orientó, corrigió, incentivó y sobre todo me dio su ejemplo en toda la formación del doctorado.

A mi consejero de estudios, Dr. Sebastián Munilla Leguizamón por su tiempo, apoyo, valiosísimos aportes, consejos y su amistad.

Al Departamento de Producción Animal de la Facultad de Agronomía de la Universidad de Buenos Aires, por la disponibilidad de sus instalaciones y el tratamiento personal proporcionado.

A los miembros de la Escuela para Graduados Alberto Soriano.

A Susana y Amelia, excelentes mujeres que en cada curso siempre estuvieron para motivarme en el proceso de formación.

A los docentes de la Cátedra de Mejoramiento Animal. A Anita Birchmeier, por ser una madre de la cual recibí consejos y un apoyo incondicional. A todos, incluyendo a Laura y Valeria por su apoyo, cariño y respeto.

A mis parceros del grupo de Mejoramiento Genético Animal: Carolina Garcia, Natalia Forneris, Yeni Bernal, Andres Rogberg, Jose Luis Gualdron y Sebastian Munilla.

A los colegas de las cátedras de Fisiología animal, Bovinos de carne y Nutrición y alimentación animal, por los agradables momentos vividos durante este tiempo.

A todos mis amigos en Buenos Aires en especial a Andrea, Martin, Luciana, Santiago, Doctora Magusa, Tatiana, Juan Felipe, Richard, Sergio y todos aquellos que hicieron de mi estadía en Argentina una excelente experiencia.

Por último, a mis padres, hermanos y sobrinos que son mi gran motivación personal para realizar todos estos esfuerzos.

DECLARACIÓN

Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.

Juan David Corrales Alvarez.

PUBLICACIONES DERIVADAS DE LA TESIS

- Corrales, JD.; Munilla, S.; y Cantet, RJC.; 2015. Polynomial order selection in random regression models via penalizing adaptively the likelihood. *J*ournal of *Animal Breeding and Genetics*. 132(4): 281-288.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

Página

ÍNDICE DE FIGURAS

ABREVIATURAS

AIC	Criterio de Información Akaike
BIC	Criterio de Información Bayesiana
DS	diferencial de selección
e.g.	Acrónimo del latín *exempli gratia* (‘dado como ejemplo’). Se utiliza para indicar ‘véase, por ejemplo…’.
ESS	Tamaño efectivo de muestra.
h^2	Heredabilidad en sentido estricto.
HPD	Intervalo de alta densidad posterior
i	intensidad de selección
i.e.	Acrónimo del latín *id est* (‘esto es’). Se utiliza para indicar ‘es decir, …’.
IA	Inseminación artificial
IEP	Intervalo entre partos
LEG	Polinomio de Legendre
LnL	Logaritmo de la Verosimilitud
LRT	Test de coeficiente de verosimilitud
MCMC	Cadenas de Markov y simulación de Monte Carlo. Refiere a un conjunto de métodos de simulación por Monte Carlo basados en la teoría de las cadenas de Markov.
MM	Modelo mixto
MRA	Modelo de regresión aleatoria
MSEP	Error cuadrático medio de predicción
pág.	Abreviatura de ‘página’.
PAL	Penalización Adaptada a la verosimilitud
PP	Prueba de progenie
REML	Máxima verosimilitud restringida.
$\mathbf{r_{ti}}$	Confiabilidad
SG	Selección genómica
SN	Normal Asimétrica
TM	modelo verdadero
UT	modelo desconocido
VCG	Valores de cría genómicos
wMSEP	Error cuadrático medio de predicción ponderado

Título: Selección y estimación de parámetros genéticos en bovinos lecheros

RESUMEN

La estimación de parámetros genéticos en bovinos lecheros requiere ajustar modelos estadísticos para datos longitudinales (modelos de regresión aleatoria y/o medidas repetidas). Es esencial en regresión aleatoria determinar el orden correcto de los polinomios que describen los coeficientes aleatorios en el tiempo. Tradicionalmente, esta tarea se desarrolló empleando criterios de selección como AIC o BIC, que no siempre logran dilucidar claramente el modelo apropiado. Esta tesis introduce el criterio PAL para la selección del orden del polinomio en modelos de regresión aleatoria, empleando una penalización adaptativa de la función de verosimilitud. Comparativamente, PAL presentó un desempeño superior a AIC y BIC, y su aplicación a datos de producción produjo un modelo parsimonioso, con buena bondad de ajuste y habilidad de predicción. Se abordó además el problema de estimar parámetros genéticos bajo un modelo de medidas repetidas para caracteres que muestran distribución asimétrica. Para evitar desvíos del supuesto de normalidad de los errores, se presentó un modelo alternativo basado en asumir una distribución Normal asimétrica. Se implementó un enfoque bayesiano con muestreo de Gibbs en datos de intervalos entre partos. Incluyendo un parámetro adicional, se obtuvieron estimaciones más precisas de los parámetros genéticos. Estos desarrollos metodológicos estuvieron motivados por los desafíos que enfrenta actualmente el programa de mejoramiento genético de la raza Holstein en Colombia. Entre ellos, figura también el impacto de introducir toros extranjeros provenientes del programa de selección genómica norteamericana. En este marco, se estimó el progreso genético esperado en la población comercial colombiana. Los cálculos reflejaron que utilizar toros genómicos jóvenes se traducirá en una respuesta genética acelerada y en una disminución de la diferencia genética entre el hato comercial y el núcleo del cual provienen.

Palabras claves: bovinos lecheros, selección del modelo, polinomios de legendre, distribución normal asimétrica, estructura de covarianza, selección genómica.

Title: Selection and estimation of genetic parameters in dairy cattle

ABSTRACT

Estimation of genetic parameters in dairy cattle requires fitting statistical models for longitudinal data (e.g., random regression or repeated measures models). Random regressions are typically used for performance traits that change over time. Prior to estimate genetic parameters, researcher should choose the order of the polynomial that better fits the data. Traditionally, this task has been carried out through either AIC or BIC, which are selection criteria. However, they do not always infer the model that is most appropriate. In the current thesis we introduced the PAL, an alternative criterion for selecting an appropriate polynomial order, which is based on penalizing adaptively the likelihood. Comparatively, PAL outperformed AIC and BIC in a simulation, and its application to milk production data produced a model that fitted data best in terms of parsimony and predictive ability. In addition, we addressed the problem of estimating genetic parameters under a model of repeated measures for traits displaying asymmetric distribution. To avoid deviations from the Normal errors assumption, a model based on an asymmetric distribution was presented. Using a Bayesian approach and Gibbs sampling, we fitted the model to calving interval data. By means of introducing a single additional parameter, more accurate estimates of the genetic parameters were obtained. All these methodological developments were motivated by the challenges currently faced by the Holstein breeding program in Colombia. Among them, the impact of the introduction of foreign genomic young bulls showed up as noteworthy. Therefore, we calculated the expected genetic progress produce by disseminating genomically selected bulls into the Colombian commercial population. Our calculations showed that the use of young bulls would result in an accelerated genetic response and a decrease in the genetic lag between the Colombian commercial herd and the foreign nucleus.

Key words: dairy cattle, model selection, legendre polynomials, skew normal distribution, covariance structure, genomic selection.

Capítulo 1

Introducción general

1.1 Introducción

El objetivo del mejoramiento genético animal está basado en la determinación de los mejores individuos para una o varias características de importancia en la producción, aplicando principios de genética cuantitativa que permitan seleccionar los padres de los individuos de la siguiente generación. Los hijos de estos animales son en promedio superiores a esta generación y por lo tanto subirán la media de la población. Para determinar la superioridad genética es necesario la utilización de diferentes metodologías que permitan la obtención de parámetros genéticos y ambientales para las características de importancia en los sistemas de producción animal.

En bovinos lecheros los parámetros genéticos, especialmente para producción, calidad, características de tipo y de fertilidad se obtienen a través de la estimación de componentes de (co)varianza utilizando diferentes metodologías basadas en la generalización del modelo animal presentado por Henderson (1982). En particular, las características para las cuales se obtienen medidas repetidas en el tiempo a partir de registros diarios de producción tales como: producción de leche, grasa y proteína, son evaluadas a partir del modelo de regresión aleatoria (Meyer 1998). El modelo estándar de regresión aleatoria (Jamrozik y Schaeffer 1997) requiere de la estimación de componentes de varianzas y covarianzas asociados a los coeficientes de funciones aleatorias que describen la desviación respecto a la trayectoria de la curva de lactancia, los cuales pueden ser estimados usando REML (Meyer 1998) o análisis bayesiano a través de muestreo de Gibbs (Jamrozik y Schaeffer 1997; Rekaya et al. 1999). Los modelos de regresión aleatoria utilizan polinomios (splines, polinomios de Legendre, etc) que describen la trayectoria de la curva fenotípica, genética y ambiental permanente de características de producción medidas en el tiempo. Polinomios de Legendre (LEG) de diferente orden han sido comúnmente utilizados para modelar la estructura de covarianza entre los coeficientes de regresión aleatorios debido a sus beneficios por ser ortogonales y la posibilidad de estimar componentes de varianza genética a lo largo de toda la lactancia (Kirkpatrick et al. 1990; Pool y Meuwissen 2000; Schaeffer 2004). Por lo tanto, uno de los objetivos iniciales al utilizar un modelo de regresión aleatoria es la selección del orden más correcto de los polinomios de Legendre, para esto los criterios más comunes para la selección del orden del polinomio de Legendre son el AIC ó criterio de información Akaike (Akaike 1974) y el BIC ó criterio de información bayesiana (Schwarz 1978). La selección de modelos basados en estos dos criterios, se basan principalmente en la selección del modelo que presenta menor valor para AIC ó para BIC. Sin embargo, la utilización de AIC favorece el modelo con mayor número de parámetros (Hurvich y Tsai 1989; McQuarrie et al. 1997) y la utilización del BIC tiene un desempeño pobre cuando el "modelo verdadero" no está entre los modelos candidatos (Burnham et al. 2011), definiendo como "modelo verdadero" aquel que presenta la mejor aproximación al proceso generador de los datos (Burnham y Anderson, 2002). Por lo tanto, es necesario la introducción de un nuevo criterio que permita obtener los componentes de varianza genéticos bajo un modelo parsimonioso y confiable a través de la selección del orden de los LEG más correcto para posteriormente estimar los componentes de (co)varianza del modelo.

Uno de los supuestos que se requieren para la utilización de algunas metodologías de estimación de componentes de varianza es la normalidad de los datos. Sin embargo, la existencia de normalidad en algunos casos es difícil de comprobar y su uso es cuestionado

considerablemente por la falta de robustez frente a las desviaciones de la distribución cuando los datos muestran asimetría con colas alargadas hacia un sentido de la distribución (asimetría positiva a la derecha o asimetría negativa a la izquierda, Figura 1.1), y por lo tanto no puede proporcionar una estimación precisa de la variación entre los individuos (Lachos et al. 2009). Por lo anterior, es importante utilizar un modelo estadístico con una considerable flexibilidad en los supuestos de distribución de los efectos aleatorios. La distribución Skew-normal (**SN**) o normal asimétrica consiste en una distribución de probabilidad que generaliza la distribución normal y permite un valor de asimetría diferente de cero. Varona et al. (2009) presentaron la distribución SN para modelar el efecto residual en el contexto de un modelo animal.

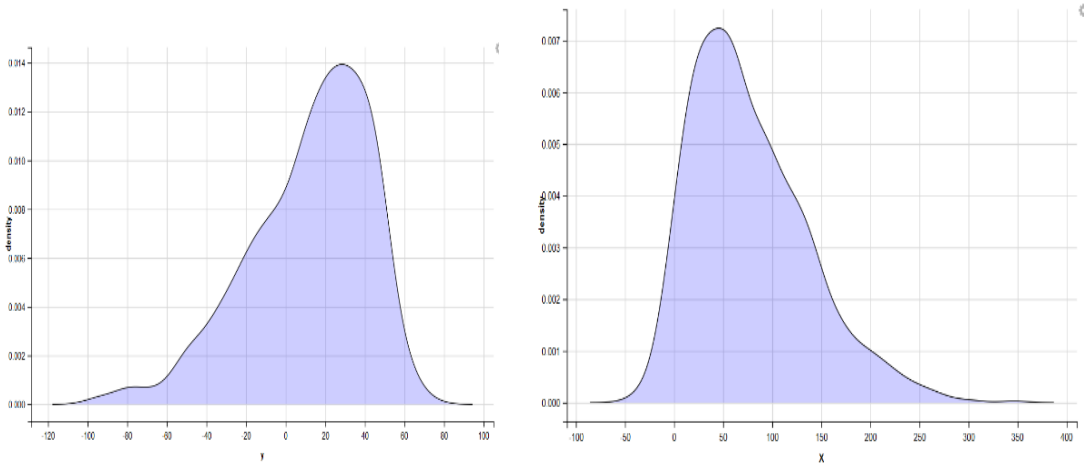

Figura 1.1. Gráficos de distribución asimétrica: asimétrica hacia la izquierda (variable y) y hacia la derecha (variable x).

Algunos caracteres relacionados con la eficiencia reproductiva en ganado lechero se han encontrado que tienen cierto nivel de asimetría, por ejemplo, el inicio la actividad luteal postparto (Darwash et al. 1997), el desempeño reproductivo del hato (Plaizier et al. 1998), los intervalos, parto-primer servicio, primer servicio-concepción y parto-concepción (Tiezzi et al. 2011) y el intervalo parto-primera inseminación (Weigel y Rekaya 2000). La eficiencia reproductiva es importante en el ámbito productivo y en los programas de mejoramiento genético, debido a que una alta eficiencia contribuye a reducir el intervalo entre generaciones traduciéndose en un aumento del progreso genético. Para alcanzar un progreso genético óptimo es necesario conocer la verdadera situación de los hatos en cuanto a sus condiciones genéticas; a través de los registros genealógicos, desempeño productivo y reproductivo. De esta manera, se podrán determinar los parámetros genéticos necesarios para evaluar el progreso genético anual y conocer las habilidades predichas de transmisión que permitan orientar los apareamientos.

En la última década debido a la disponibilidad de paneles densos de marcadores moleculares, la ganadería bovina lechera a nivel mundial ha ido incorporando nueva información genética para la identificación de los individuos superiores a través de la

evaluación genómica. Diferentes estudios han probado la superioridad de la selección genómica comparada con la tradicional prueba de progenie, principalmente por la disminución en los intervalos generacionales y el incremento de la intensidad de selección lo cual permite aprovechar el potencial genético de los individuos a una temprana edad (Schaeffer 2006). En Colombia la raza Holstein utiliza aproximadamente el 90% de material genético importado (pajillas) por lo que la determinación del progreso debido a la selección de toros genómicos es una herramienta importante para la confianza del productor debido principalmente a que se espera que permita incrementar el progreso genético de la población al disminuir el retraso generacional entre la población local y la evaluada. Adicionalmente, teniendo resultados positivos se podría proponer la realización de evaluaciones genómicas locales o la incorporación a una evaluación genómica internacional.

En virtud de lo expuesto anteriormente, los objetivos generales de la tesis son los siguientes.

1. Introducir un método de selección del orden del polinomio en modelos de regresión aleatoria; 2. Aplicar el modelo Normal asimétrico a datos de intervalos entre partos en bovinos lecheros. 3. Evaluar el progreso genético a partir de la implementación de la selección genómica en la población Holstein de Colombia.

El documento está organizado en cuatro capítulos, incluyendo este capítulo introductorio. En el Capítulo 2 se introduce el criterio de selección basado en la penalización adaptada a la verosimilitud (PAL) para la selección del orden del polinomio para los efectos genéticos aditivos y ambientales permanentes en un modelo de regresión aleatoria, el desempeño del procedimiento es evaluado a través de una simulación y su utilización es ilustrada a través de una base de datos para producción de leche al primer parto en bovinos Holstein de Colombia.

En el capítulo 3 se presenta el modelo lineal mixto desde una perspectiva bayesiana con distribución skew-normal o normal asimétrica para modelar caracteres asimétricos en ganado lechero y será ilustrado a través de una aplicación para datos de intervalo entre partos de vacas de la raza Holstein de Colombia.

Por último, en el capítulo 4 se presenta una simulación determinística para valorar el progreso genético de la población Holstein de Colombia a partir de la introducción de material genético de toros evaluados por prueba genómica siguiendo el método descrito por Hill (1974) y utilizando datos reales de la población Holstein de Colombia.

Los algoritmos para el modelo skew-normal fueron programados en fortran 77. Dada su extensión, los códigos completos no están incluidos en el documento. Sin embargo, pueden ser solicitados al autor.

Capítulo 2

Selección del orden del polinomio en modelos de regresión aleatoria a través de una penalización adaptativa de la verosimilitud

2.1. Introducción

El modelo de regresión aleatoria (MRA) es ampliamente utilizado en los programas de evaluación genética en ganado lechero para estimar componentes de varianza en caracteres de producción que se expresan a lo largo del tiempo. En un MRA, la curva de lactancia se modela a través de una trayectoria promedio más un desvío individual, asociado a los efectos genéticos aditivos y ambientales permanentes, que se define a partir de un conjunto de coeficientes de regresión aleatorios. Para modelar la estructura de covarianza entre estos coeficientes de regresión se utilizan generalmente polinomios ortogonales de Legendre. En este contexto, la confiabilidad de la predicción de los efectos genéticos y ambientales permanentes en el MRA depende de definir apropiadamente el orden de los LEG. Para producción de leche se han utilizado típicamente LEG de igual orden (entre 3 y 5) para modelar los efectos genéticos aditivos y ambientales permanentes (Pool y Meuwissen 2000; Strabel et al. 2005; Herrera et al. 2013). Sin embargo, no es necesario ajustar modelos con el mismo orden del LEG para ambos efectos (Pool y Meuwissen 2000; Liu et al. 2006; Bignardi et al. 2009). Por ejemplo, Liu et al. (2006) encontraron que el mejor modelo consistió en un polinomio de orden 5 para los efectos genéticos aditivos y de orden 7 para los efectos ambientales permanentes.

En lo que resta del capítulo, cuando se hable de selección de modelos se lo hará sólo mencionando "selección", para evitar reiterar el uso de las palabras modelo y modelos. El criterio de información de Akaike (AIC; Akaike 1974) y el criterio de información bayesiano (BIC; Schwarz 1978) se utilizan frecuentemente para seleccionar un modelo parsimonioso y con un orden del polinomio que ajuste mejor la trayectoria de la curva de lactancia (Bignardi et al. 2009; Pereira et al. 2013). Tanto AIC como BIC fueron desarrollados para minimizar la distancia de Kullback-Leibler entre el modelo operativo y el ideal (el "modelo verdadero" o "true model") como base fundamental para la selección entre modelos alternativos (Burnham y Anderson 2004). En estudios longitudinales, la utilización de AIC como criterio de selección tiende a favorecer el modelo con un mayor número de parámetros cuando el tamaño muestral es lo suficientemente grande (Hurvich y Tsai 1989; McQuarrie et al. 1997). Por otro lado, BIC tiene un pobre desempeño cuando el modelo verdadero no se encuentra entre los modelos candidatos (Burnham et al. 2011). En la práctica, el modelo verdadero es raramente conocido (casi que exclusivamente en simulaciones) y, en consecuencia, no está claro cuál criterio debe utilizarse. Para superar este problema, Stoica y Babu (2013) presentaron un nuevo criterio de selección, empleando una penalización adaptativa de la verosimilitud (en inglés, "Penalizing Adaptively the Likelihood", referida por su acrónimo, PAL). En términos simples, un criterio adaptativo es aquel que utiliza información de los pasos previos para mejorar su desempeño selectivo. El PAL permite seleccionar el orden del modelo cuando el "mejor" modelo es desconocido (Entiéndase como mejor modelo aquel que mejor ajusta los datos).

Este capítulo se encuentra organizado del siguiente modo. En primer lugar, se introduce el PAL como criterio de selección del orden del polinomio para los efectos genéticos aditivos y ambientales permanentes en un modelo de regresión aleatoria. Luego se evalúa el desempeño del procedimiento mediante cuatro escenarios distintos en un experimento de simulación. Finalmente, se ilustra la implementación del PAL ajustando datos de producción de leche respecto del día del control en vacas Holstein de Colombia.

2.2. Materiales y métodos

2.2.1. Modelo

En notación matricial, la ecuación de un modelo de regresión aleatoria se representa como

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{u} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{e} \qquad (1)$$

donde $\boldsymbol{y}$ es el vector $N \times 1$ de observaciones, $\boldsymbol{\beta}$ es el vector $p \times 1$ de efectos fijos y $\boldsymbol{u}$ $\left(N_{\mathrm{A}} m_1 \times 1\right)$ y $\boldsymbol{\gamma}$ $\left(N_{\mathrm{D}} m_2 \times 1\right)$ son los vectores de los coeficientes de regresión aleatoria asociados con los efectos genéticos aditivos y ambientales permanentes, respectivamente. En esta notación N_{A} es el número de animales en el archivo de pedigrí, N_{D} es el número de vacas con registro, y m_1 y m_2 corresponden al orden del polinomio de Legendre para la correspondiente función. Finalmente, $\boldsymbol{e}$ es el vector de errores. Asimismo, $\boldsymbol{X}$, $\boldsymbol{Z}$ y $\boldsymbol{W}$ son las matrices de incidencia para los efectos fijos y los coeficientes aleatorios genéticos aditivos y ambientales permanentes, respectivamente (ver Schaeffer 2004, para más detalles). Las matrices de varianzas-covarianzas genética aditiva, ambiental permanente y del error son iguales a

$$\mathrm{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\gamma}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & 0 & 0\\ 0 & \boldsymbol{P_e} & 0\\ 0 & 0 & \boldsymbol{R}\end{bmatrix}$$

La varianza de los valores aditivos es $\boldsymbol{G} = \boldsymbol{A} \otimes \boldsymbol{K}_{\mathrm{A}}$, siendo $\boldsymbol{A}$ $\left(N_{\mathrm{A}} \times N_{\mathrm{A}}\right)$ la matriz de relaciones aditivas entre los animales y $\boldsymbol{K}_{\mathrm{A}}$ una matriz cuadrada que contiene las covarianzas entre los coeficientes de regresión aleatoria para los efectos genéticos aditivos. El símbolo $\otimes$ representa el operador producto Kronecker (Searle 1982). Adicionalmente, $\boldsymbol{P_e} = \boldsymbol{I}_{N_{\mathrm{D}}} \otimes \boldsymbol{K}_{\mathrm{Pe}}$, donde $\boldsymbol{I}_{N_{\mathrm{D}}}$ es una matriz identidad de orden N_{D} y $\boldsymbol{K}_{\mathrm{Pe}}$ una matriz cuadrada con las covarianzas entre los coeficientes de regresión aleatoria para los efectos ambientales permanentes. Finalmente, los errores se estructurarán en clases agrupando aquellos con una varianza residual cercana ± 0.3, de esta forma, $\boldsymbol{R} = \mathrm{Diag}\left\{\sigma^2_{e_k}\right\}$ es una matriz diagonal con el mismo componente de varianza $\sigma^2_{e_k}$ dentro de la *k*-ésima clase de errores.

Asumiendo que el vector de datos $\boldsymbol{y}$ sigue una distribución normal multivariada, la expresión igual a menos dos veces el logaritmo de la máxima verosimilitud residual o restringida $\left(-2\ln L\right)$ es igual a

$$-2\ln L = const + N_A \ln\left|\boldsymbol{K}_{\mathrm{A}}\right| + m_1 \ln\left|\boldsymbol{A}\right| + N_{\mathrm{D}} \ln\left|\boldsymbol{K}_{\mathrm{Pe}}\right| + \ln\left|\boldsymbol{C}\right| + \ln\left|\boldsymbol{R}\right| + \boldsymbol{y}'\boldsymbol{P}\,\boldsymbol{y} \qquad (2)$$

En (2), $\boldsymbol{C}$ es la matriz de coeficientes de las ecuaciones del modelo mixto, $\ln\left|\cdot\right|$ denota el logaritmo natural del determinante de la correspondiente matriz $(\cdot)$ y $\boldsymbol{y}'\boldsymbol{P}\,\boldsymbol{y}$ es la suma de cuadrados del error del modelo (Meyer y Hill 1997). Sea $\boldsymbol{V}$ la matriz de varianzas-covarianzas del vector de observaciones $\boldsymbol{y}$, entonces se define $\boldsymbol{P}$ como

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{V}^{-1}\boldsymbol{X}'$$

La función (2) es la piedra angular de los dos criterios que se utilizan más frecuentemente para seleccionar el orden de los LEG en un modelo de regresión aleatoria: el criterio de AIC y BIC. Más precisamente, estos dos métodos muestran una formulación similar al momento de minimizar la verosimilitud penalizada (Burnham y Anderson 2002), la que se puede resumir en la siguiente expresión

$$\min_{p}\left[-2\ln L + p\ \omega\right] \qquad (3)$$

En (3), el escalar *p* es el número de parámetros aleatorios del modelo, mientras que ω es un coeficiente de penalización. Si la verosimilitud restringida (2) es empleada en (3), el coeficiente de penalización para AIC es $\omega_{\mathrm{AIC}} = 2a_n^*$ (Müller *et al.* 2013), donde $a_n^* = (N-b)/(N-b-p-1)$, con *N* igual al número de observaciones y b igual el número de parámetros a ser estimados para los efectos fijos. A su vez, para BIC es $\omega_{\mathrm{BIC}} = \ln(N)$. Estos dos métodos difieren en la manera de seleccionar el 'mejor' modelo entre los candidatos que se están evaluando (Yang, 2005): mientras AIC clasifica los modelos basados en un criterio eficiente (*i.e.*, el mejor modelo es aquel que minimiza asintóticamente la varianza del error; Casella y Berger 2002, pp. 470-473), BIC es conocido por ser un criterio consistente (*i.e.*, cuando el tamaño de la muestra tiende a infinito, la probabilidad de que BIC seleccione el mejor modelo se aproxima a uno; Casella y Berger 2002, pp. 467-470). Si el verdadero modelo está entre los que se encuentran bajo consideración, BIC debería ser el mejor criterio a elegir. Sin embargo, si éste no es el caso, AIC sería preferible (Burnham y Anderson 2004). Aun así, AIC es criticado en el análisis de datos longitudinales porque tiende a elegir los modelos de alto orden cuando el tamaño de la muestra crece ilimitadamente (Shibata 1981; McQuarrie et al. 1997). Además, Yang (2005) demostró teóricamente que ni AIC ni BIC son simultáneamente óptimos en términos de consistencia y eficiencia.

Con el objeto de resumir estas ideas, considere un conjunto de modelos anidados $M_1 \subset M_2 \subset \ldots \subset M_{\tilde{n}}$, donde el subíndice indica el número de modelos y M_{n_0} un modelo verdadero con p_{n_0} parámetros. Por modelo "verdadero" $\left(M_{n_0}\right)$ se entiende al modelo con el menor número de parámetros posible que se halla cercano (en el sentido de la distancia

de Kullback-Leibler) al proceso que genera las observaciones (Davidson y Mackinnon 2004). En las ciencias biológicas es incierta la existencia de un modelo verdadero dentro del conjunto de modelos en consideración, por lo tanto, no está claro cuál criterio debería ser utilizado. Para solucionar este problema, Stoica y Babu (2013) presentaron una alternativa basada en una penalización adaptativa de la verosimilitud (criterio PAL). Note que el término negativo del logaritmo de la verosimilitud (– 2LnL) en (3) disminuye cuando *p* aumenta, mientras que el término de penalización se incrementa. Stoica y Babu (2013) explicaron en forma intuitiva cómo construir un criterio para obtener un término de penalización ideal:

a) Cuando el número de parámetros es menor que en el modelo "verdadero", i.e. $p_i < p_{n0}$, el criterio debe efectuar una penalización pequeña, que a su vez pierda peso cuando *p* aumenta.
b) En cambio, si el número de parámetros es mayor que los del modelo "verdadero", i.e. $p_i > p_{n0}$, la penalización debería aumentar con el incremento de *p*.

Para desarrollar un criterio que considere estos dos principios simultáneamente, Stoica y Babu (2013) propusieron el siguiente coeficiente de penalización

$$\omega_{\mathrm{i,\,PAL}} = \ln\left(p_{\tilde{n}}\right)\left(\frac{\ln\left(r_i+1\right)}{\ln\left(\rho_i+1\right)}\right) \tag{4}$$

El escalar $p_{\tilde{n}}$ corresponde al número de parámetros del modelo de mayor orden dentro del conjunto considerado (por ejemplo, si el conjunto de modelos es $M_1 \subset M_2 \subset \ldots \subset M_{10}$, ñ=10 y contiene 48 parámetros aleatorios entonces $p_{\tilde{n}} = 48$). Asimismo, las expresiones

$$r_i = 2\ln L_{i-1} - 2\ln L_1$$

$$\rho_i = 2\ln L_{\tilde{n}} - 2\ln L_{i-1}$$

representan a los cocientes de verosimilitud generalizados entre el modelo M_{i-1} y el modelo reducido M_1 o el modelo completo $M_{\tilde{n}}$, respectivamente. Donde, *i*=1, …, ñ y corresponde al número del modelo que se está evaluando (e.g., en 4 modelos, M_1 con p_1=8 , M_2 con p_2=12, M_3 con p_3=18 y $M_{\tilde{n}}$ con $p_{\tilde{n}}$=26; como se puede ver en el ejemplo de cálculo del PAL). En este trabajo, asumiremos que M_1 es un "modelo de referencia" y $r_2 = 0$. Como resultado, el criterio PAL para seleccionar el orden del modelo se define por

$$\mathrm{PAL}_i = -2\,\ln L_i + p_i\,\ln\left(p_{\tilde{n}}\right)\frac{\ln\left(r_i+1\right)}{\ln\left(\rho_i+1\right)} \tag{5}$$

De acuerdo a este criterio, el mejor modelo es aquel que presenta un menor valor de PAL. En el apartado 2.2.2 se incluye un pequeño ejemplo demostrativo sobre cómo calcular el PAL. Cuando el modelo verdadero está dentro de los candidatos, PAL selecciona el

mismo orden del modelo que BIC; en caso contrario, PAL favorece un modelo similar a AIC (ver Stoica y Babu 2013, para más detalles). Dadas estas propiedades, el criterio PAL aparece como un método atractivo cuando se desea seleccionar el orden del LEG tanto para los efectos genéticos aditivos como para los ambientales permanentes, dentro de un modelo de regresión aleatoria.

A modo de representación gráfica de estos argumentos, en la Figura 2.1 se presenta el desempeño de la penalización de cada uno de los tres criterios (AIC, BIC y PAL) bajo un modelo simulado con 33 parámetros aleatorios. En esta figura, se puede observar el cambio de pendiente del criterio PAL cuando el "mejor" modelo es encontrado.

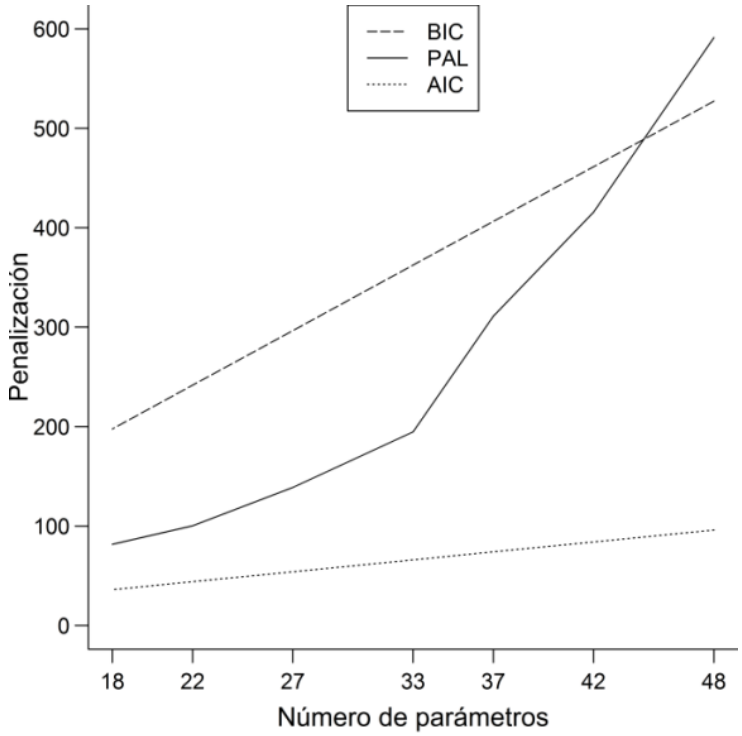

Figura 2.1. Penalización utilizando diferentes criterios de selección: penalización adaptativa de la verosimilitud (PAL), criterio de información Bayesiana (BIC) y el criterio de información de Akaike.

2.2.2. Ejemplo de cálculo del criterio PAL

Considere los siguientes datos simulados para ilustrar el cálculo del criterio PAL. El verdadero modelo es el LEG33. Sin embargo, los resultados de otros tres modelos están incluidos también en el cuadro 2.1. La tercera columna corresponde al logaritmo de la verosimilitud, que se obtuvo a partir de la ecuación (2).

Cuadro 2.1. Ejemplo de cálculo del criterio PAL.

	LEGm_1m_2	Número de parámetros	lnL	ω_{PAL}	PAL
1)	LEG11	8	–97.864		
2)	LEG22	12	–93.755	0	187.510
3)	LEG33	18	–92.247	3,6656	184.560
4)	LEG44	26	–92.226	8,0792	184.662

En este ejemplo $\tilde{n}$ = 4 y $p_{\tilde{n}}$ = 26 (modelo LEG44). Luego, para obtener $\omega_{\text{PAL-LP33}}$, primero se calcularon los valores r_{LEG33} y ρ_{LEG33} empleando la expresión (5):

$$r_{\text{LEG33}} = 2\ln L_{\text{LEG22}} - 2\ln L_{\text{LEG11}} = 2\times(-93.755) - 2\times(-97.864) = 8.218;$$
$$\rho_{\text{LP33}} = 2\ln L_{\text{LEG44}} - 2\ln L_{\text{LEG22}} = 2\times(-92.226) - 2\times(-93.755) = 3.016$$

Entonces

$$\omega_{\text{PAL-LEG33}} = \ln(26)\frac{\ln(8.218+1)}{\ln(3.016+1)} = 3,6656$$

Como resultado, el valor del criterio PAL para la selección del orden del modelo se obtiene de la ecuación (6) como

$$\text{PAL(LEG33)} = -2\ln L_{\text{LEG33}} + 18\ \omega_{\text{PAL-LEG33}}$$

Reemplazando:

$$\text{PAL}(\text{LEG33}) = -2\times(-92.247) + 18\times 3.6656 = 184.560$$

El cálculo de PAL para los modelos LEG44 y LEG22 se desarrolló de manera similar.

2.2.3. Experimento de simulación

Los datos para la simulación fueron creados teniendo en cuenta la estructura y la genealogía de los registros de la raza Holstein en Colombia (Cuadro 2.2). En el modelo se incluyeron los efectos fijos de hato-día de control, edad de la vaca (como regresión lineal y

cuadrática) y la trayectoria fenotípica. En la siguiente descripción, LEGm_1m_2 se refiere al orden del polinomio de Legendre para los efectos genéticos aditivos (m_1) y ambientales permanentes (m_2), respectivamente. Se consideraron cuatro escenarios de acuerdo a la fracción de registros en el conjunto de datos, los que fueron simulados a partir de un modelo verdadero: 1) **TM**: conjunto de datos en el que todos los registros (100%) fueron simulados a partir de alguno de los siguientes modelos: LEG33, LEG44, LEG55 y LEG66; 2) **NS**: 95% de los registros fue generado a partir del modelo verdadero y el 5% restante fue muestreado al azar de acuerdo a uno de los otros tres modelos simulados. Por ejemplo, mientras se consideraba el modelo LEG55 como el verdadero, un 5% de registros fue generado aleatoriamente a partir de los modelos LEG33, LEG44 o LEG66. 3) **NR**: 95% de los registros fue generado a partir del modelo verdadero y el 5% restante fue muestreado aleatoriamente de la base de datos de la Holstein; y 4) **UT**: si bien cada registro fue muestreado a partir de uno de los siguientes cuatro modelos: LEG35, LEG45, LEG55 y LEG65, ninguno de ellos era representativo de la población en su conjunto. Nótese que en este caso los cuatro modelos se obtuvieron variando el orden del polinomio de los efectos genéticos aditivos y manteniendo constante el de los efectos ambientales permanentes. Luego, el "mejor" modelo fue elegido usando el error cuadrático medio de predicción (MSEP) para producción de leche en 305 días:

$$MSEP = \sum_{i=1}^{l} \sum_{j=6}^{305} \frac{\left(\hat{a}_{ij} - a_{ij} \right)^2}{l * 300},$$

donde l representa al número de vacas, 300 al número de valores de cría en el trayecto de la lactancia, a_{ij} corresponde al *j* ésimo verdadero valor de cría simulado de la *i* ésima vaca y $\hat{a}_{ij}$ es el $BLUP\,(a_{ij})$ del *j* ésimo día de la *i* ésima vaca utilizando el orden del polinomio seleccionado. En cada réplica, los verdaderos valores de cría se muestrearon en igual proporción de cada uno de los cuatro modelos evaluados. Por su parte, los valores de cría predichos se obtuvieron para cada uno de los cuatro modelos de ajuste. El modelo con el mínimo valor de MSEP fue considerado mejor. Para cada escenario, se simularon 100 réplicas en total y se analizaron utilizando los criterios AIC, BIC y PAL.

2.2.4. Análisis de la base de datos de producción de leche

Se recolectaron 60.513 registros diarios de producción de leche, tomados en el día de control lechero sobre un total de 6.675 vacas Holstein de primera lactancia, entre los años 1989 y 2008, en 164 hatos pertenecientes a la Asociación Holstein de Colombia. Los registros de control lechero se tomaron en un período de 5 a 305 días de lactancia en vacas con una edad al primer parto de entre 19 y 48 meses. Sólo se incluyeron en el análisis registros de producción de leche de entre 5,1 y 48,4 kg. El número mínimo de vacas para formar un grupo contemporáneo fue de 7 individuos. El archivo de pedigrí contenía 17.062 animales. El cuadro 2.2 provee una descripción de los datos utilizados en este estudio.

Cuadro 2.2. Descripción de la base de datos.

Item	Valor
Número de registros (tomados al día del control lechero)	60.513
Número de vacas con registro	6.675
Número de animales en el pedigrí	17.062
Número de grupos contemporáneos	4.211
Media de producción de leche (kg)	18,80 ± 5,95
Media de edad al primer parto (meses)	31,67 ± 4,61

Valores posteriores a ± corresponden a las desviaciones estándar de cada carácter.

Al igual que anteriormente, LEGm_1m_2 representa un modelo de regresión aleatoria basado en polinomios de Legendre en el que los dos subíndices indican el orden del polinomio para los efectos genéticos aditivos (primero) y los ambientales permanentes (último). Se consideraron los siguientes órdenes para los polinomios: m_1 = 3, … , 6 y m_2 = 3, … , 6. Todos los modelos posibles que surgen de la combinación de los órdenes de los polinomios, más el modelo simple que ajusta únicamente el intercepto para los efectos genéticos aditivos y ambientales permanentes (LEG11), fueron evaluados con AIC, BIC, y PAL (un total de 17 modelos). Además, se incluyeron efectos fijos de hato-día de control, edad de la vaca (con términos lineal y cuadrático) y la trayectoria fenotípica. Todos los modelos tuvieron seis intervalos o clases para la varianza residual (6-35, 36-95, 96-125, 126-215, 216-245, 246-305 días en lactancia) y la trayectoria fenotípica se modeló con un LEG de orden 5. Los valores de −2lnL, AIC, BIC, el test del cociente de verosimilitud (LRT) y las estimaciones de los componentes de varianza y covarianza para cada uno de los 17 modelos de regresión aleatoria se obtuvieron a partir del programa computacional Wombat (Meyer 2007). Este programa provee estimaciones REML de los componentes de varianza utilizando el algoritmo "average information" (Gilmour et al. 1995). El modelo reducido (M_1, modelo con menor orden dentro del análisis), necesario para computar el criterio PAL, fue LEG11, mientras que el de mayor orden fue LEG66 ($\tilde{n}$ = 48). Adicionalmente, se determinó la habilidad predictiva de cada uno de los modelos mediante la suma de cuadrados del error de predicción ponderada (wMSEP), tal como lo describen Odegård et al. (2003). El procedimiento consiste en separar dos subconjuntos de datos generados por exclusión de observaciones a partir del conjunto de datos inicial. Posteriormente, se calcula el MSEP en cada subconjunto y, finalmente, la wMSEP se computa como el promedio de ambos MSEP.

La heredabilidad estimada en el día de control *t* fue calculada utilizando la siguiente fórmula (Van Der Werf et al. 1998; Jakobsen et al. 2002) :

$$\hat{h}_t^2 = \frac{\hat{\sigma}^2_{a(t)}}{\hat{\sigma}^2_{a(t)} + \hat{\sigma}^2_{pe(t)} + \hat{\sigma}^2_{e(t)}}$$

donde $\hat{\sigma}^2_{a(t)}$, $\hat{\sigma}^2_{pe(t)}$ y $\hat{\sigma}^2_{e(t)}$ representan las varianzas genética aditiva, ambiental permanente y residual en el día *t*, respectivamente.

2.3. Resultados

Se calculó la probabilidad de seleccionar el modelo con el orden correcto del polinomio a través de los criterios AIC, BIC y PAL en los diferentes escenarios simulados, a modo de resumen de los resultados obtenidos en el experimento de simulación (Figura 2.2). Cuando el verdadero modelo estaba entre los candidatos, tanto PAL como BIC seleccionaron con probabilidad igual a uno el orden correcto del LEG para los efectos genéticos aditivos y ambientales permanentes. En cambio, no siempre el criterio AIC seleccionó el orden correcto. Cuando se introdujo 5% de ruido aleatorio en los datos, AIC tendió a sobreestimar el orden correcto. Por el contrario, incluso cuando el orden del modelo era desconocido (mejor modelo seleccionado a través del menor MSEP), PAL seleccionó el mejor modelo con mayor probabilidad que AIC. En este último escenario, BIC no seleccionó el mejor modelo en ninguna réplica.

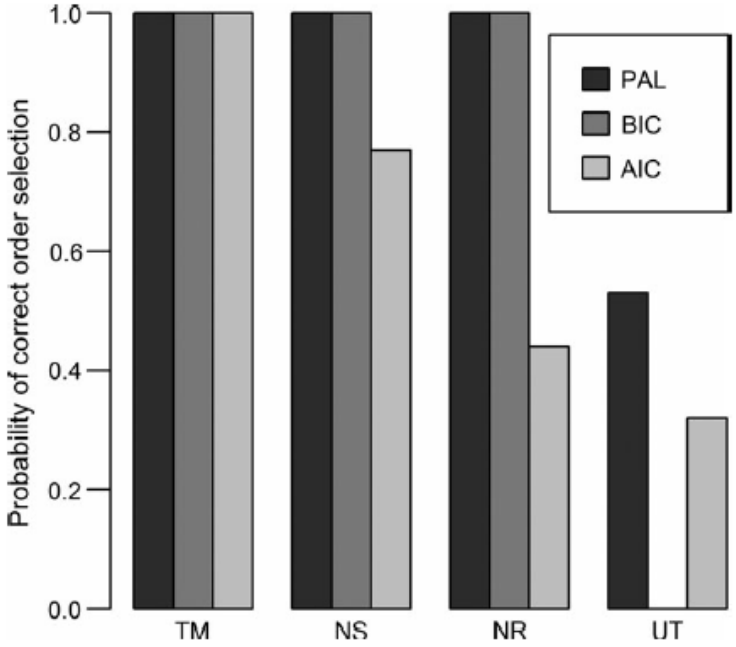

Figura 2.2. Probabilidad de seleccionar el orden correcto del polinomio de Legendre en cuatro escenarios simulados: criterio de penalización adaptativa de la verosimilitud (PAL), criterio de información bayesiana (BIC) y el criterio de información de Akaike (AIC).

En el Cuadro 2.3 se presentan los valores de los diferentes criterios de selección (–2lnL, PAL, AIC, BIC, LRT) y el wMSEP para cada uno de los modelos de regresión aleatoria con diferentes órdenes del LEG ajustados a los datos de producción lechera de la población Holstein de Colombia. El menor valor de AIC correspondió al modelo LEG66, seguido por el LEG56. Estos dos fueron los modelos con el mayor número de parámetros en el conjunto considerado en este estudio. En cambio, los mejores modelos de acuerdo al criterio PAL fueron similares a los seleccionados por BIC: el mejor modelo fue LEG36, seguido por LEG46. El orden del polinomio de Legendre para los efectos genéticos aditivos fue menor al que se encontró para los efectos ambientales permanentes. De acuerdo al test del cociente de verosimilitud (LRT), el modelo con más parámetros (LEG66) fue el mejor. Sin embargo, las diferencias en la habilidad predictiva (wMSEP) de los modelos de orden 6 para los efectos ambientales permanentes fueron pequeñas y, por lo tanto, los modelos LEG66 y LEG36 fueron clasificados aproximadamente igual.

Cuadro 2.3. Criterios de selección de modelos con diferentes órdenes de polinomios de Legendre (LEG) para los efectos genéticos aditivos (m_1) y ambientales permanentes (m_2).

	Número de	Criterios de Selección[1]					
LEGm_1m_2	**parámetros**	**-2lnL**	**AIC**	**BIC**	**PAL**	**LRT***	**wMSEP**
1) LEG66	48	180,814	**180,910**	181,342	181,406	**(1-2) 18**	**1.5898**
2) LEG56	42	180,832	180,916	181,294	181,248	(2-3) 20	1.5905
3) LEG46	37	180,852	180,926	181,259	181,163	(3-4) 36	1.5987
4) LEG36	33	180,888	180,954	**181,250**	**181,083**	(4-8) 380	1.5992
5) LEG65	42	180,936	181,020	181,397	181,189	(5-6) 282	1.6608
6) LEG55	36	181,218	181,290	181,614	181,433	(6-7) 22	1.7988
7) LEG45	31	181,240	181,302	181,580	181,423	(7-8) 28	1.8000
8) LEG35	27	181,268	181,322	181,565	181,407	(8-12) 618	1.7964
9) LEG64	37	181,122	181,196	181,528	181,332	(9-10) 268	1.7442
10) LEG54	31	181,390	181,452	181,730	181,550	(10-11) 465	1.8729
11) LEG44	26	181,855	181,907	182,141	181,989	(11-12) 31	2.0265
12) LEG34	22	181,886	181,930	182,128	181,986	(12-16) 1297	2.0269
13) LEG63	33	181,287	181,353	181,649	181,465	(13-14) 305	1.8483
14) LEG53	27	181,592	181,646	181,889	181,728	(14-15) 451	1.9905
15) LEG43	22	182,043	182,087	182,285	182,143	(15-16) 1140	2.1263
16) LEG33	18	183,183	183,219	183,381	183,183	(16-17) 9310	2.4090
17) LEG11	3	192,493					5.6990

[1]REML log-Likelihood (-2lnL), AIC= Criterio de información de Akaike, BIC= Criterio de información bayesiana, PAL= Penalización adaptativa de la verosimilitud, LRT= Cociente de verosimilitud entre los modelos (i.e. 1-2 significa la comparación entre el modelo 1 y el modelo 2). *P<0.01 and wMSEP= MSEP ponderado de dos muestras independientes.

Como fuera mencionado, entonces, tanto el criterio PAL como BIC seleccionaron el modelo LEG36 (con un polinomio de Legendre de orden 3 para el efecto genético aditivo y 6 para el efecto ambiental permanente, respectivamente) como el más apropiado. Para este modelo, las varianzas residuales estimadas fueron iguales a 5,11, 3,57, 3,35, 3,06, 2,72, y 2,63, para los 6 rangos comprendidos entre los días 6-35, 36-95, 96-125, 126-215, 216-245 y 246-305 de producción de leche, respectivamente. La Figura 2.3 muestra las varianzas genética aditiva y ambiental permanente para producción de leche del modelo seleccionado, definidas sobre la trayectoria de producción entre los días 5 y 305. La menor varianza genética aditiva fue 3,11 kg^2 en el día 305 de lactación, mientras que el mayor valor fue 8,10 kg^2 en el día 102. Los valores máximo y mínimo de las varianzas ambientales permanentes correspondieron a 8,53 kg^2 al día 17 y 11,27 kg^2 al día 305 de lactancia. Las heredabilidades estimadas (h^2) sobre la trayectoria de producción de leche variaron entre 0,18 para el día 305 y 0,39 al día 142 de lactancia.

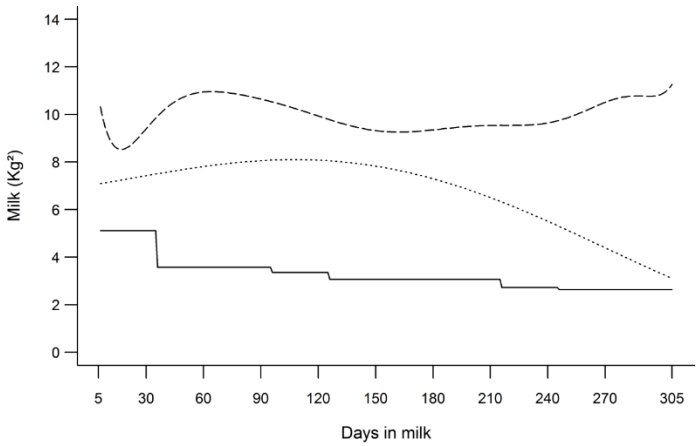

Figura 2.3. Varianza ambiental permanente (línea traceada), varianza genética aditiva (línea punteada) y varianza residual (línea solida) para producción de leche diaria en la primera lactancia.

2.4. Discusión

En este estudio se presentó el criterio PAL como un método novedoso para seleccionar el orden del polinomio de Legendre en modelos de regresión aleatoria, y se comparó su desempeño contra otros métodos estándar (i.e., AIC y BIC) a través de un experimento de simulación. Los tres métodos seleccionaron el modelo correcto cuando el 100% de los datos fue simulado a partir de un único modelo. Sin embargo, cuando se introdujo algo de ruido en las simulaciones, PAL y BIC seleccionaron el orden correcto, mientras que AIC tendió a sobreestimar el orden del modelo. Estos resultados son consistentes con estudios previos que muestran que AIC tiende a sobre-ajustar parámetros con datos longitudinales (Schwarz 1978; Shibata 1981; Hurvich y Tsai 1989; Stoica y Babu 2013). Por el contrario, cuando los datos fueron simulados bajo un modelo en el cual se desconocía el orden correcto, PAL se desempeñó mejor que AIC y que BIC en términos de seleccionar el modelo con el menor MSEP. Que AIC sea más efectivo que BIC en este contexto puede explicarse por el hecho de que el número de modelos no crece rápidamente en dimensión y, en consecuencia, el MSEP del modelo seleccionado por AIC se aproxima asintóticamente al mínimo valor del conjunto de modelos candidatos (Shibata 1981; Yang 2005).

Si bien el uso de los polinomios ortogonales de Legendre en un modelo de regresión aleatoria permite modelar más flexiblemente la curva de lactancia, un orden elevado del LEG es frecuentemente imposible de implementar en bases de datos grandes o en poblaciones numerosas. Esto se debe a que, por un lado, se requeriría un enorme poder de cómputo y, por otro, a que existe la posibilidad de obtener correlaciones negativas entre controles distantes (Pool y Meuwissen 2000; Jamrozik et al. 2001; Bohmanova et al. 2008). Por esta razón, los conjuntos de datos simulados consideraron un polinomio de Legendre con un orden máximo de 6 tanto para los efectos genéticos aditivos como para los ambientales permanentes. Por otro lado, se consideró que un polinomio de orden menor a 3 no ajusta bien las desviaciones de una curva de lactancia típica. Por lo tanto, los modelos que se evaluaron en este trabajo estaban en el rango LEG33 a LEG66. Todos los modelos en este intervalo fueron ajustados a los datos.

En una implementación práctica de modelos de regresión aleatoria con LEG, la primera tarea que el analista debe desarrollar es la de definir el orden apropiado del polinomio (tanto para los efectos genéticos aditivos como para los ambientales permanentes). En muchas ocasiones, sin embargo, es difícil establecer el orden correcto, principalmente porque los criterios estadísticos no están claramente definidos (Bignardi *et al.* 2009). Por ejemplo, López-Romero y Carabaño (2003), Liu et al. (2006) y Bignardi et al. (2009) utilizaron tanto AIC como BIC como criterios para elegir el orden del polinomio en modelos de regresión aleatoria. Sin embargo, en todos estos trabajos ambos criterios difirieron respecto al modelo a seleccionar. Los resultados del experimento de simulación del presente trabajo sugieren que PAL es un buen criterio para elegir el orden de los polinomios de Legendre, en cualquier implementación de un modelo de regresión aleatoria para datos de producción de leche. El criterio PAL proporciona una regla para seleccionar entre diferentes órdenes de los polinomios de Legendre y, en particular, cuando los resultados producidos por AIC difieren de aquellos producidos por BIC. Stoica y Babu (2013) señalaron que hasta el momento no existe una prueba teórica de la superioridad del PAL sobre AIC y/o BIC. Sin embargo, este criterio puede ser empleado con la idea de que su aplicación conjuga el uso de AIC y BIC, y se ajusta a un criterio consistente con el marco de inferencia de los datos y los modelos a comparar.

Como se deduce de nuestra implementación sobre la base de datos de producción de leche, PAL permite seleccionar órdenes diferentes para los efectos genéticos aditivos y los ambientales permanentes. En general, el orden apropiado del LEG para los efectos ambientales permanentes tiende a ser mayor que para los efectos aditivos (Pool y Meuwissen 2000; López-Romero y Carabano 2003; Carabaño et al. 2007). Por ejemplo, Liu et al. (2006) utilizaron el logaritmo de la verosimilitud y un índice que consideraba diferentes criterios de información para seleccionar el orden del polinomio, y concluyeron que LEG57 fue el modelo más apropiado. Por su parte, Bignardi et al. (2009) seleccionaron un orden de LEG7.12 mediante los criterios AIC y BIC, mientras López-Romero y Carabaño (2003) seleccionaron un modelo con LEG de orden 2 - 3 para los efectos genéticos aditivos y 5 - 6 para los efectos ambientales permanentes. En nuestra implementación, los parámetros genéticos estimados del modelo seleccionado por PAL fueron de similar magnitud a aquellos reportados en investigaciones previas por Jakobsen et al. (2002), López-Romero y Carabaño (2003), y López-Romero et al. (2003). Las varianzas residuales al comienzo de la lactancia fueron mayores que en los otros intervalos. Resultados similares fueron reportados por López-Romero et al. (2003) cuando evaluaron la heterogeneidad de varianzas en un modelo de regresión aleatoria. En este trabajo también se obtuvieron estimaciones de las varianzas de los efectos ambientales permanentes mayores que aquellas de los efectos genéticos aditivos. Una explicación posible radica en el hecho que, bajo las condiciones típicas de pastoreo en Colombia, la producción de leche es altamente influenciada por efectos ambientales.

En conclusión, los resultados sugieren que cuando se desea seleccionar el orden del LEG en un modelo de regresión aleatoria, PAL se presenta como un criterio muy útil. Dos consideraciones prácticas son importantes al implementar PAL. En primer lugar, que es necesario una estructura anidada de los modelos a comparar. Aun así, en el contexto de evaluar el orden del polinomio de Legendre en un modelo de regresión aleatoria, esta estructura anidada surge naturalmente por la adición de órdenes a los polinomios. En

segundo lugar, es necesario considerar el modelo reducido dentro del conjunto a comparar para poder computar así el cociente de verosimilitud generalizada en la fórmula del PAL.

Capítulo 3

Modelo lineal mixto con distribución skew-normal para caracteres asimétricos con aplicación a intervalo entre partos en bovinos lecheros

3.1. Introducción

El modelo lineal mixto (MM) es ampliamente utilizado para estimar los valores de cría de los animales (Henderson 1984). La primera etapa para su implementación involucra estimar parámetros genéticos y ambientales, para luego obtener las predicciones de los valores de cría a través de la resolución del sistema de ecuaciones del MM. Los denominados métodos bayesianos constituyen una aproximación general para abordar el problema de la estimación de parámetros genéticos de dispersión, simultáneamente con los valores de cría en un escenario probabilístico (Sorensen y Gianola 2002). La aproximación bayesiana fue sugerida como un método inferencial para la solución de muchos problemas en el mejoramiento genético animal (Gianola y Fernando 1986). Como cualquier otro enfoque en este campo de estudio, la inferencia bajo el MM bayesiano se basa en el uso de aproximaciones normales para los efectos aleatorios (Gelman et al. 2009). En particular, el MM bayesiano modela la distribución típicamente normal de los caracteres cuantitativos como una mezcla de distribuciones Normales independientes, asociadas con los efectos genéticos y ambientales.

Sin embargo, ciertos caracteres cuantitativos importantes en la producción animal no siguen una distribución normal y muestran, en cambio, una densidad asimétrica con colas alargadas hacia una u otra cola de la distribución (asimetría positiva hacia a la derecha o negativa hacia a la izquierda). En bovinos lecheros, por ejemplo, el inicio de la actividad luteal postparto (Darwash et al. 1997), la edad al primer parto (Heinrichs y Vazquez-Anon 1993), los intervalos parto-primer servicio, primer servicio-concepción y parto-concepción (Tiezzi et al. 2011), la probabilidad de infecciones intramamarias (Rodriguez-Zas et al. 1998), el conteo de células somáticas (Rönnegård et al. 2013), el color de la leche (Winkelman et al. 1999) y la producción de leche (Jamrozik et al. 2004) son ejemplos de caracteres que muestran cierto grado de asimetría. En todos estos casos, la existencia de un mínimo pero no de un máximo biológico y/o cuando las vacas reciben un tratamiento preferencial, podrían explicar la existencia de asimetría en la distribución de los valores fenotípicos (Kuhn et al. 1994; Plaizier et al. 1998). A los efectos de abordar esta problemática, ha surgido un considerable interés por flexibilizar el supuesto de normalidad de los efectos ambientales (i.e., el error del modelo). En poblaciones lecheras es particularmente relevante desarrollar modelos estadísticos con una distribución robusta de los errores.

Una posible alternativa es modelar los errores con distribuciones diferentes a la normal, – aunque en términos generales – el enfoque desafíe la teoría subyacente a los métodos estándares de inferencia. Para evitar este problema, Sahu et al. (2003) desarrollaron una metodología bayesiana de estimación sobre la base de una familia de distribuciones asimétricas, entre cuyos miembros figura la distribución Normal asimétrica (SN, por ‘Skew-normal’ en inglés). Bajo un enfoque bayesiano jerárquico, esta metodología regresa los errores del modelo en un coeficiente que pondera el grado de asimetría. Como consecuencia, el término de error puro de esta regresión bayesiana sigue una distribución normal y se ajusta así a los supuestos de la teoría clásica. Varona et al. (2008) adaptaron el trabajo de Sahu et al. (2003) al análisis de características cuantitativas en producción animal, definiendo la asimetría como una medida de sensibilidad (positiva o

negativa) de las influencias ambientales sobre el fenotipo, e ilustraron el procedimiento en datos de tamaño de camada en cerdos. El objetivo de este capítulo es aplicar el modelo Normal asimétrico a datos de intervalos entre partos en bovinos lecheros, y comparar su desempeño contra el modelo de errores Normales frecuentemente utilizado.

3.2. Materiales y métodos

3.2.1. Modelo

Considere un modelo lineal para un único carácter con medidas repetidas en observaciones fenotípicas con una distribución asimétrica de los errores:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{a} + \mathbf{W}\mathbf{p} + \boldsymbol{\varepsilon} \quad (1)$$

donde

$\boldsymbol{y}$: es el vector ($n \times 1$) de observaciones fenotípicas.

$\mathbf{X}$: matriz ($n \times p$) de incidencia de los efectos fijos.

$\boldsymbol{\beta}$: vector ($p \times 1$) de los efectos fijos sistemáticos y ambientales.

$\mathbf{Z}$: matriz de incidencia ($n \times q$) para los efectos genéticos aditivos.

$\boldsymbol{a}$: es el vector ($q \times 1$) de valores de cría.

$\mathbf{W}$: matriz de incidencia ($n \times n$) para el vector de efectos ambientales permanentes. En esta matriz se considera una estructura de los efectos permanentes dentro de vaca lo cual explica el orden de la matriz ($n \times n$) igual al número de datos.

$\mathbf{p}$: es el vector ($n \times 1$) de efectos ambientales permanentes.

$\boldsymbol{\varepsilon}$: vector ($n \times 1$) de errores aleatorios.

Sin perder generalidad, asúmase una parametrización de rango completo del vector $\boldsymbol{\beta}$, tal que $\mathrm{r}[\mathbf{X}] = p$. El vector de valores de cría $\boldsymbol{a}$ es estocásticamente independiente de los otros efectos aleatorios en el modelo y se distribuye normalmente con esperanza cero y matriz de (co)varianza igual a $\mathbf{A}\sigma_A^2$. La matriz $\mathbf{A}$ contiene las relaciones aditivas y σ_A^2 corresponde a la varianza aditiva. Asimismo, el vector **p** sigue una distribución Normal con media cero y matriz de (co)varianza igual a $\mathbf{U} = \mathbf{P}\sigma_p^2$, donde σ_p^2 es la varianza ambiental permanente y $\mathbf{P}$ es una matriz positiva definida que modela la estructura de covarianza entre las medidas repetidas en el tiempo. Finalmente, se considera que el vector $\boldsymbol{\varepsilon}$ sigue una distribución Normal asimétrica (Azzalini y Dalla Valle 1996).

De acuerdo a la derivación presentada por Sahu et al. (2003), la Normal asimétrica es una distribución de probabilidad continua definida mediante la transformación

$$\boldsymbol{\varepsilon} = \delta \mathbf{z} + \boldsymbol{e} \tag{2}$$

donde δ es un parámetro de asimetría, $\boldsymbol{z}$ un vector aleatorio cuyos elementos son estrictamente positivos que sigue una distribución normal truncada en cero de la normal tipificada y $\boldsymbol{e}$ es otro vector aleatorio, pero con distribución Normal. La media y varianza de la Normal asimétrica son iguales a (Sahu et al. 2003):

$$E(\boldsymbol{\varepsilon}) = \mathbf{I}\sqrt{\frac{2}{\pi}}\delta \qquad \text{y} \qquad Var(\boldsymbol{\varepsilon}) \equiv \mathbf{I}\sigma_{\varepsilon}^2 = \mathbf{I}\left(\sigma_e^2 + \left(1 - \frac{2}{\pi}\right)\delta^2\right) \tag{3}$$

Nótese que si $\delta = 0$, entonces $\boldsymbol{\varepsilon} = \boldsymbol{e}$ y consecuentemente $\boldsymbol{\varepsilon}$ sigue una distribución normal. En cambio, si el valor del parámetro de asimetría es positivo (i.e. $\delta > 0$), se obtiene una distribución asimétrica positiva (cola larga hacia la derecha). Por el contrario, para valores de $\delta < 0$ se obtiene una distribución asimétrica negativa (cola larga hacia la izquierda). De acuerdo con esta definición, y asumiendo que los errores siguen una distribución Normal asimétrica, es posible escribir el modelo (1) de manera ligeramente distinta:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{a} + \mathbf{W}\boldsymbol{p} + \delta\boldsymbol{z} + \boldsymbol{e} \tag{4}$$

3.2.2. Estimación bayesiana de los componentes de (co)varianza

Considérese a continuación la implementación de un análisis bayesiano jerárquico del modelo (4) para estimar los componentes de (co)varianza (Sahu et al. 2003; Varona et al. 2008). En la primera etapa del análisis es necesario especificar la distribución condicional conjunta de las observaciones. De acuerdo a los supuestos del modelo definidos en la sección precedente, esta distribución corresponde a un proceso normal multivariado:

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{z}, \delta, \sigma_e^2 \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{a} + \mathbf{W}\boldsymbol{p} + \delta\,\boldsymbol{z}, \mathbf{I}\sigma_e^2\right)$$

Explícitamente,

$$p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{z}, \delta, \sigma_e^2\right) \propto \exp\left[-\frac{1}{2\sigma_e^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{a} - \mathbf{W}\boldsymbol{p} - \delta\,\boldsymbol{z})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{a} - \mathbf{W}\boldsymbol{p} - \delta\,\boldsymbol{z})\right] \tag{5}$$

En la segunda etapa, es necesario asignar distribuciones a priori a todas las incógnitas del modelo.

3.2.2.1 Distribuciones a priori

Para el vector de efectos fijos se asumirá un proceso normal multivariado, con el objetivo de obtener distribuciones posteriores propias (Hober y Casella 1996). En particular, se establecerá que , donde $\mathbf{K}$ es una matriz diagonal con elementos sumamente grandes $\left(k_i > 1\times 10^8\right)$ de modo de asegurar un estado de incertidumbre importante en relación con los valores de los parámetros en $\boldsymbol{\beta}$. Por su parte, los efectos genéticos aditivos $\boldsymbol{a}$ y ambientales permanentes $\boldsymbol{p}$ se distribuyen a priori como $\boldsymbol{\beta}|\mathbf{K} \sim NMV(\mathbf{0},\mathbf{K})$

$$\boldsymbol{a}\left|\mathbf{A},\sigma_a^2\right. \sim NMV\left(\mathbf{0},\mathbf{A}\sigma_a^2\right)$$

$$\boldsymbol{p}\left|\mathbf{P},\sigma_p^2\right. \sim NMV\left(\mathbf{0},\mathbf{P}\sigma_p^2\right)$$

Finalmente, de acuerdo a la definición de la distribución normal asimétrica multivariada propuesta por Sahu et al. (2003) para los errores, se asumirá que

$$\boldsymbol{z} \sim N(\mathbf{0},\mathbf{I})\ \ I(z_i > 0) \tag{6}$$

Donde la función $I(z_i > 0)$ indica que los valores de z_i deben ser estrictamente positivos.

$$\delta \sim N\left(0,\Gamma\right) \tag{7}$$

En el segundo nivel de la jerarquía, es también necesario asignar distribuciones a priori a los componentes de varianza. Se asumirá que las varianzas σ_a^2, σ_p^2 y σ_e^2 siguen a priori distribuciones Chi-cuadrado invertidas. Explícitamente,

$$p\left(\sigma_a^2\right) \propto \left(\sigma_a^2\right)^{-\left(\frac{\nu_A}{2}+1\right)} \exp\left\{-\frac{\nu_A\, s_A^2}{2\,\sigma_a^2}\right\} \tag{8}$$

$$p\left(\sigma_p^2\right) \propto \left(\sigma_p^2\right)^{-\left(\frac{\nu_P}{2}+1\right)} \exp\left\{-\frac{\nu_P\, s_P^2}{2\,\sigma_p^2}\right\} \tag{9}$$

$$p\left(\sigma_e^2\right) \propto \left(\sigma_e^2\right)^{-\left(\frac{\nu_E}{2}+1\right)} \exp\left\{-\frac{\nu_E\, s_E^2}{2\,\sigma_e^2}\right\} \tag{10}$$

En esta notación, s_A^2, s_P^2 y s_E^2 representan valores "razonables" para la varianzas aditiva, ambiental permanente y del error "puro", respectivamente, mientras que ν_A, ν_P y ν_E indican los grados de credibilidad asignados a estos valores. Todos estos términos constituyen los denominados "hiperparámetros" y son especificados por el analista para describir el conocimiento a priori o su opinión experta respecto a la distribución de los componentes de varianza. Para los análisis realizados se utilizaron los siguientes valores para los hiperparametros $s_A^2 = 1.158$ y $s_E^2 = 5.613$ para el modelo unicarácter y para el modelo multicaracter $s_A^2 = 191.38$ y $s_E^2 = 6347.56$ $s_{\mathrm{p}}^2 = 787.33$ $\nu_A = 10$, $\nu_P = 10$ y $\nu_E = 10$.

3.2.2.2. Distribución condicional conjunta

Asúmase ahora que $\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{z}, \delta$ y σ_e^2 son mutuamente independientes a priori. De acuerdo al teorema de Bayes, la densidad posterior conjunta será entonces proporcional al producto de la función de verosimilitud y del producto de cada densidad a priori que se definió en la sección precedente. Simbólicamente

$$\begin{aligned} p\left(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{z}, \delta, \sigma_e^2 \mid \boldsymbol{y}\right) \propto\; & p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{z}, \delta, \sigma_e^2\right) p\left(\boldsymbol{\beta} \mid \boldsymbol{K}\right) p\left(\boldsymbol{a} \mid \mathbf{A}\sigma_a^2\right) p\left(\boldsymbol{p} \mid \mathbf{P}\sigma_p^2\right) \times \\ & \times p\left(\sigma_a^2 \mid \nu_A, s_A^2\right) p\left(\sigma_p^2 \mid \nu_P, s_P^2\right) p\left(\delta \mid \Gamma\right) p\left(\boldsymbol{z}\right) p\left(\sigma_e^2 \mid \nu_E, s_E^2\right) \end{aligned} \tag{11}$$

A partir de la expresión (11) es posible identificar la distribución condicional posterior de cualquier parámetro de interés, manteniendo el resto de ellos constante. En el próximo apartado se derivarán las distribuciones condicionales de todos los parámetros del modelo.

3.2.2.3. Distribuciones condicionales posteriores

Defínase, en primer lugar, al vector de parámetros de "posición" como $\boldsymbol{\theta}' = \left[\boldsymbol{\beta}' \quad \boldsymbol{a}' \quad \boldsymbol{p}'\right]$, y al vector de componentes de varianza como $\Omega' = \left[\sigma_a^2 \quad \sigma_p^2 \quad \sigma_e^2\right]$.

Considérese luego el siguiente sistema de ecuaciones lineales:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X}+\mathbf{I}\frac{\sigma_e^2}{k_i} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W}+\mathbf{P}^{-1}\frac{\sigma_e^2}{\sigma_p^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'(\boldsymbol{y}-\delta\boldsymbol{z}) \\ \mathbf{Z}'(\boldsymbol{y}-\delta\boldsymbol{z}) \\ \mathbf{W}'(\boldsymbol{y}-\delta\boldsymbol{z}) \end{bmatrix} \tag{12}$$

Este sistema puede escribirse sucintamente como $\mathbf{C}\boldsymbol{\theta}=\boldsymbol{r}$, donde $\mathbf{C}$ representa a la matriz de coeficientes y $\boldsymbol{r}$ al vector de "elementos de la derecha" de las ecuaciones (*right hand sides*). Las soluciones al sistema de ecuaciones (12) pueden escribirse como $\hat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{r}$. Por las propiedades de la distribución Normal multivariada, la distribución condicional de $\boldsymbol{\theta}$ será Normal multivariada (*cf.* Sorensen y Gianola, Cap. 13.2.1), con parámetros definidos por:

$$\boldsymbol{\theta}\,|\,\boldsymbol{\Omega},\boldsymbol{z},\delta,\boldsymbol{y}-\delta\boldsymbol{z} \sim N_{p+q+n}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{p}} \end{bmatrix}, \begin{bmatrix} \mathbf{X}'\mathbf{X}+\mathbf{I}\frac{\sigma_e^2}{k_i} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W}+\mathbf{P}^{-1}\frac{\sigma_e^2}{\sigma_p^2} \end{bmatrix}^{-1} \right) \tag{13}$$

k_i en estas matrices es el mismo para todos los efectos fijos.

Por su parte, la distribución condicional posterior de los componentes de varianza corresponde a las siguientes distribuciones Chi-cuadrado escaladas invertidas:

$$p\left(\sigma_a^2\,|\,\boldsymbol{\theta},\sigma_p^2,\sigma_e^2,\boldsymbol{z},\delta,\boldsymbol{y}\right) \propto \left(\sigma_a^2\right)^{-\left(\frac{q+\nu_A}{2}+1\right)} \exp\left\{-\frac{\boldsymbol{a}'\mathbf{A}^{-1}\boldsymbol{a}+\nu_A\, s_A^2}{2\,\sigma_a^2}\right\} \tag{14}$$

$$p\left(\sigma_p^2 \middle| \boldsymbol{\theta}, \sigma_a^2, \sigma_e^2, \boldsymbol{z}, \delta, \boldsymbol{y}\right) \propto \left(\sigma_p^2\right)^{-\left(\frac{n+\nu_P}{2}+1\right)} \exp\left\{-\frac{\boldsymbol{p}'\mathbf{P}^{-1}\boldsymbol{p} + \nu_P\, s_P^2}{2\,\sigma_p^2}\right\} \tag{15}$$

$$p\left(\sigma_e^2 \middle| \boldsymbol{\theta}, \sigma_a^2, \sigma_p^2, \boldsymbol{z}, \delta, \boldsymbol{y}\right) \propto \left(\sigma_e^2\right)^{-\left(\frac{n+\nu_E}{2}+1\right)} \exp\left\{-\frac{\boldsymbol{e}'\boldsymbol{e} + \nu_E\, s_E^2}{2\,\sigma_e^2}\right\} \tag{16}$$

Finalmente, y siguiendo los desarrollos de Sahu et al. (2003), defíniremos

$$\boldsymbol{w} = \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{a} - \mathbf{W}\boldsymbol{p}$$

Luego, las distribuciones condicionales posteriores de $\boldsymbol{z}$ y δ son las siguientes

$$p\left(\boldsymbol{z} \middle| \boldsymbol{\theta}, \Omega, \delta, \boldsymbol{y}\right) \sim N_n\left[\frac{\delta}{\sigma_e^2 + \delta^2}\boldsymbol{w}, \frac{\sigma_e^2}{\sigma_e^2 + \delta^2}\right] I\ (\boldsymbol{z} > 0) \tag{17}$$

$$p\left(\delta \mid \boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{z}, \boldsymbol{y}\right) \sim N\left[\left(\frac{1}{\Gamma} + \frac{\boldsymbol{z}'\boldsymbol{z}}{\sigma_e^2}\right)^{-1} \frac{1}{\sigma_e^2}\, \boldsymbol{z}'\boldsymbol{w}, \left(\frac{1}{\Gamma} + \frac{\boldsymbol{z}'\boldsymbol{z}}{\sigma_e^2}\right)^{-1}\right] \tag{18}$$

En el caso particular en que $\Gamma = 1$, y definiendo $\varphi = \sigma_e^2 + \boldsymbol{z}'\boldsymbol{z}$, la distribución condicional posterior de δ es igual a:

$$p\left(\delta \mid \boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{z}, \boldsymbol{y}\right) \sim N\left[\frac{1}{\varphi}\boldsymbol{z}'\boldsymbol{w}, \frac{\sigma_e^2}{\varphi}\right]$$

3.2.2.4. Distribuciones marginales posteriores: muestreo de Gibbs

La última etapa del análisis implica obtener las distribuciones marginales posteriores de todas las incógnitas del modelo, en particular las de aquellos parámetros de interés: los componentes de varianza. Dado que todas las distribuciones condicionales posteriores son

conjugadas, es posible implementar el algoritmo del muestreo de Gibbs cuya implementación requiere muestrear secuencialmente de todas las distribuciones condicionales posteriores. Una vez que el algoritmo convergió, el muestreo secuencial de las distribuciones condicionales resultará en un muestreo de las densidades marginales de cada parámetro. El algoritmo del muestreo de Gibbs se realizará entonces siguiendo los pasos a continuación:

1. Construir las MME (12), sumar k_i^{-1} al elemento diagonal de cada efecto fijo y, finalmente, resolver el sistema de ecuaciones
2. Muestrear β, $\boldsymbol{a}$ y $\boldsymbol{p}$ a partir de la expresión (13)
3. Calcular $\boldsymbol{w} = \boldsymbol{y} - \mathbf{X}\beta - \mathbf{Z}\boldsymbol{a} - \mathbf{W}\boldsymbol{p}$
4. Calcular las formas cuadráticas para las varianzas
5. Muestrear $\boldsymbol{z}$ de la densidad (17)
6. Muestrear la varianza del error "puro" (16)
7. Muestrear la varianza aditiva de la densidad (14)
8. Muestrear la varianza de los efectos permanentes de la densidad (15)
9. Muestrear δ de la densidad (18)
10. Regresar al paso 1 y repetir los pasos 1-9 tantas veces como requiera el muestreo para alcanzar convergencia.

En cada muestreo, los valores resultantes de los parámetros de covarianza son almacenados para obtener luego estadísticos descriptivos de la distribución posterior marginal de cada uno de los parámetros de interés. El procedimiento completo involucra un período de calentamiento o "*burn-in*", que depende de los valores asignados inicialmente y del número de ciclos adecuados para asegurar la convergencia del algoritmo.

3.2.3. Implementación del análisis en datos de intervalo entre partos en bovinos lecheros

En esta sección se describe la implementación del análisis bayesiano jerárquico con el objeto de estimar los parámetros genéticos para el carácter intervalo entre partos (IEP: diferencia en días entre dos partos consecutivos) en bovinos lecheros. Los datos provienen de registros de vacas Holstein pertenecientes a la Asociación Holstein de Colombia, nacidas entre los años 1988 a 2008 y distribuidas en 148 hatos lecheros. Se ajustaron dos modelos: uno uni-carácter para el primer intervalo entre partos (**IEP1**), y otro modelo con medidas repetidas de intervalo entre los cuatro primeros partos (**IEPR**). En la Figura 3.1 se esquematizan los caracteres evaluados.

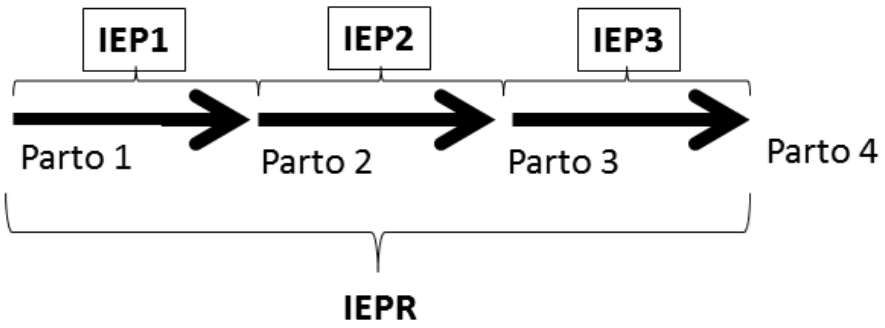

Figura 3.1. Descripción grafica del carácter intervalo entre partos

Para el análisis con medidas repetidas se consideró que las vacas tuvieran los tres intervalos entre partos (IEP1, IEP2 e IEP3). Se estableció un valor mínimo de 10 vacas para asegurar un tamaño mínimo razonable en cada grupo de contemporáneas (Hato-Año de parto). El Cuadro 3.1 presenta una descripción de los datos utilizados en el estudio.

Cuadro 3.1. Estadísticos descriptivos de los datos utilizados para IEPR.

Ítem	Valor*
Número de intervalos entre partos	27.528
Número de vacas con registro	9.176
Número de animales en el pedigrí	38.619
Número de grupos contemporáneos	1.257
Media de edad al primer parto (meses)	$31{,}49 \pm 4{,}31$
Media de intervalo entre partos (días)	434 ± 105
Media de producción de leche ajustada a los 305 días (kg)	$6.760 \pm 1{,}874$

*Valores ± corresponden a las desviaciones estándar de cada carácter.

El modelo $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{a} + \boldsymbol{\varepsilon}$ fue ajustado para el IEP1, mientras que para IEPR se

ajustó un modelo de medidas repetidas con el efecto ambiental permanente $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{a} + \mathbf{W}\boldsymbol{p} + \boldsymbol{\varepsilon}$. Para ambos modelos, los efectos fijos consistieron en edad al parto y hato-año de parto. Los errores fueron modelados mediante una distribución Normal asimétrica de acuerdo a la descomposición $\boldsymbol{\varepsilon} = \delta \boldsymbol{z} + \mathbf{e}$. A modo de comparación, se ajustó también un modelo con distribución normal para lo cual se fijó $\delta = 0$ y $\boldsymbol{\varepsilon} = \mathbf{e}$.

La estructura de covarianzas entre las medidas repetidas de una misma vaca fue la $\mathbf{P}$ propuesta por Cantet et al. (2005), la cual considera una caída lineal de la correlación con el tiempo. La estructura considera el hecho que medidas cercanas en el tiempo tienen mayor correlación que medidas más alejadas. Así, por ejemplo, las medidas realizadas entre el parto 1 y el 2 están más correlacionadas que entre el parto 1 y el 3. Específicamente, **P** es una matriz diagonal en bloques, donde cada bloque corresponde a las covarianzas entre medidas repetidas de una misma vaca. Los elementos diagonales de un bloque cualquiera son iguales a 1 y por fuera de la diagonal equivalen a $\left(1 - \frac{t_j - t_i}{t_{nx}}\right)$ para dos tiempos cualquiera *i* y *j*, donde $t_{nx} \geq t_j > t_i \geq t_1$ y *nx* es el número de medidas repetidas. Nótese que, bajo esta estructura, la correlación entre los efectos se incrementa linealmente cuanto más cercanas sean las medidas (Cantet et al. 2005). En nuestro caso y para cualquier vaca, el bloque de **P** correspondiente es

$$\mathbf{P}_{vaca} = \begin{bmatrix} 1 & 1 - \frac{t_2 - t_1}{t_3} & 1 - \frac{t_3 - t_1}{t_3} \\ 1 - \frac{t_2 - t_1}{t_3} & 1 & 1 - \frac{t_3 - t_2}{t_3} \\ 1 - \frac{t_3 - t_1}{t_3} & 1 - \frac{t_3 - t_2}{t_3} & 1 \end{bmatrix} = \mathbf{1}\mathbf{1}' - \Psi$$

El vector **1** es de orden 3 × 1 y todos sus elementos son iguales a 1. La matriz **Ψ** posee elementos iguales a $\frac{t_j - t_i}{t_{nx}}$, cuando *j* > *i* o a 0 cuando *j* = *i*. Por ejemplo, la vaca identificada con el código 2009 tuvo registró 4 partos, los cuales sucedieron en las siguientes edades.

Cuadro 3.2. Datos para el ejemplo de cálculo de la matriz de covarianzas.

Número de parto	Tiempo	Edad en días	IEP
1	t_0	990	
2	t_1	1440	450
3	t_2	1950	510
4	t_{nx}	2419	469

El bloque de la matriz **P** correspondiente a la estructura de covarianza para las medidas repetidas de esta vaca es el siguiente:

$$\mathbf{P}_{2009} = \begin{bmatrix} 1 & 1-\frac{1950-1440}{2419} & 1-\frac{2419-1440}{2419} \\ 1-\frac{1950-1440}{2419} & 1 & 1-\frac{2419-1950}{2419} \\ 1-\frac{2419-1440}{2419} & 1-\frac{2419-1950}{2419} & 1 \end{bmatrix}$$

Con lo cual tenemos que

$$\mathbf{P}_{2009} = \begin{bmatrix} 1 & 0.789 & 0.595 \\ 0.789 & 1 & 0.806 \\ 0.595 & 0.806 & 1 \end{bmatrix}$$

La matriz **P** tiene la gran ventaja de la facilidad de cómputo de su inversa, la cual puede calcularse a partir de la siguiente expresión:

$$\mathbf{P}_{vaca}^{-1} = \left(\mathbf{11}' - \boldsymbol{\Psi}\right)^{-1}$$

Luego, la inversa se puede computar sencillamente utilizando la fórmula para la inversa de una suma de matrices (Harville 1997).

$$\mathbf{P}^{-1} = \left(-t_{nx}\right)\left[\boldsymbol{\Psi}^{-1} + \boldsymbol{\Psi}^{-1}\mathbf{1}\left(\frac{1}{1-\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}}\right)\mathbf{1}'\boldsymbol{\Psi}^{-1}\right]$$

Para estimar los componentes de (co)varianza se utilizó el algoritmo del muestreo de Gibbs descrito en la sección precedente, mediante un programa escrito en lenguaje Fortran 77. La implementación involucró 200.000 ciclos de muestreo de los cuales se descartaron los primeros 50.000 como período de calentamiento. Se calcularon varios estadísticos posteriores, tales como la media, la mediana, el desvío estándar y el intervalo de alta densidad del 95% (95%HPD) con el programa BOA (Smith 2007), realizad en un entorno *R* (*R* Core Team, 2015). La convergencia de la cadena MCMC fue evaluada mediante el estadístico diagnóstico de Geweke (Z-score, Geweke 1992) y el tamaño efectivo de la muestra (ESS, Kass et al. 1998) calculado como:

$$ESS = \frac{M}{1 + 2\sum_{i=1}^{k}\rho(i)}$$

donde ρ(*i*) representa la autocorrelación calculada para el *i*-ésimo retraso ("lag") $i = 1, \ldots, k=100$, siendo *M* es el número total de muestras una vez descartadas aquellas empleadas en el "calentamiento". Finalmente, se utilizó el criterio DIC (Spiegelhalter et al. 2002) de modo de poder comparar los modelos con errores normales *vs* normales asimétricos.

3.3. Resultados

En la Figura 3.2 se presentan las distribuciones fenotípicas de los caracteres *IEP1* e *IEPR* y los valores de asimetría correspondientes. Ambos caracteres muestran un marcado patrón de asimetría positiva. El mayor valor se presentó para el carácter que incluía los intervalos de los cuatro primeros partos. Estos patrones de asimetría son característicos para este carácter, dado que se conoce el mínimo biológico (duración de la gestación más los

días abiertos), pero se desconoce el valor máximo debido a que las vacas pueden permanecer en el hato por diferentes razones.

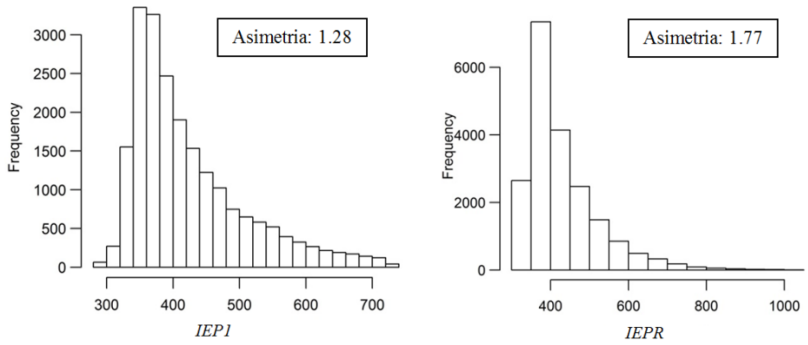

Figura 3.2. Distribución de los valores fenotípicos de los caracteres IEP1 e IEPR.

El método de comparación DIC favoreció siempre al modelo con errores Normales asimétricos respecto del modelo con errores Normales (Cuadro 3.3). Los valores de DIC estimados para IER1 y IEPR bajo el modelo SN fueron 117.832 y 152.683, respectivamente. En cambio, los valores DIC para los modelos con distribución Normal de los errores fueron 117.860 y 152.700 para IEP1 e IEPR, respectivamente. Diferencias superiores a 7 entre los dos modelos comparados son consideradas importantes (Spiegelhalter et al. 2002) y favorecieron a los modelos SN.

Cuadro 3.3. Comparación de modelos de ajuste para IEP1 e IEPR mediante el DIC.

	Modelo		
Carácter	**Normal**	**SN**	**Diferencia**
IEP1	117860	117832	28*
IEPR	152700	152683	17*

* Diferencias mayores a 7 unidades son consideradas importantes (Spiegelhalter et al. 2002).

En el cuadro 3.4 se presentan los estadísticos descriptivos de la distribución posterior de los componentes de varianza (σ_a^2 , σ_p^2 y σ_e^2), del parámetro de asimetría (δ) y de las heredabilidades (h^2), para los caracteres IEP1 e IEPR bajo el modelo SN. Las medias y medianas fueron similares para todos los parámetros, evidenciando que las distribuciones marginales posteriores son simétricas. Las medias marginales posteriores de h^2_{IEP1} y h^2_{IEPR} fueron iguales a 0.135 y 0.045, respectivamente, y el error estándar fue igual a 0.012 para h^2_{IEP1} y 0.008 para h^2_{IEPR}. La media marginal posterior del parámetro de asimetría para el carácter IEP1 fue δ_{IEP1} = 0,71 con un error estándar de 0.67. Por su parte, para el carácter IEPR el valor correspondiente de asimetría fue δ_{IEPR} = 0.62, con un error estándar de 0.56. De acuerdo a los resultados del análisis bayesiano, la probabilidad que el valor del parámetro de asimetría sea mayor a cero fue 0.86 para δ_{IEP1} y 0.87 para δ_{IEPR}. Las distribuciones marginales posteriores de la varianza del error y del parámetro de asimetría se muestran en la Figura 3.3.

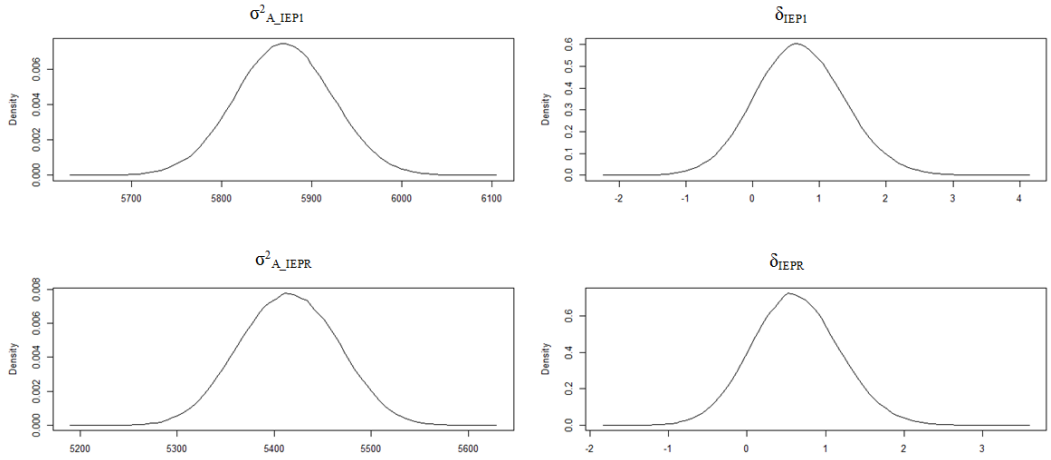

Figura 3.3. Distribuciones marginales posteriores de la heredabilidad y el parámetro de asimetría para los caracteres IEP1 e IEPR.

Cuadro 3.4. Estadísticos descriptivos de las distribuciones marginales posteriores de la varianza genética aditiva (σ^2_a), la varianza ambiental permanente (σ^2_p), la varianza del error (σ^2_e), el parámetro de asimetría (δ) y la heredabilidad (h^2) para los caracteres IEP1 e IEPR

	Media	**Mediana**	**SD**	**HDP 2.5%**	**HDP 97.5%**	**ESS**
IEP1						
σ^2_a	912,94	912,14	89,17	738,71	1088,05	503
σ^2_e	5869,15	5868,98	53,21	5766,32	5974,55	1490
δ	0,71	0,70	0,67	–0,62	2,01	68714
h^2	0,14	0,14	0,01	0,11	0,16	504
IEPR						
σ^2_a	271,82	269,64	46,91	181,35	366,51	153
σ^2_p	385,18	385,33	74,11	239,28	526,99	105
σ^2_e	5414,66	5414,51	49,36	5318,64	5509,87	110
δ	0,62	0,61	0,56	–0,44	1,74	72370
h^2	0,05	0,04	0,01	0,03	0,06	146

SD= Desviación estándar; **95% HDP=** Intervalo de alta densidad posterior**; ESS**= Tamaño efectivo de muestra.

Por último, es importante destacar que el análisis bayesiano culminó en convergencia, hecho que se evidencia en la Figura 3.4 que muestra las cadenas de Markov. Estas fueron evaluadas mediante los promedios acumulados por iteración para h^2_{IEP1}, h^2_{IEPR}, δ_{IEP1} y δ_{IEPR}. Todas las cadenas muestran un comportamiento apropiado y convergen rápidamente, excepto por aquella correspondiente a h^2_{IEPR}. En este caso, la alta autocorrelación entre muestreos, reflejada en el reducido tamaño efectivo de muestras (ESS, Cuadro 3.4), hizo lenta la convergencia. El comportamiento justifica el elevado número de muestreos descartados como período de calentamiento. En el otro extremo, los valores de autocorrelación para los parámetros de asimetría indicaron un mezclado adecuado de las cadenas: para un *lag* de 100 ciclos, los valores fueron 0.004 para δ_{IEP1} y 0.0016 para δ_{IEPR}.

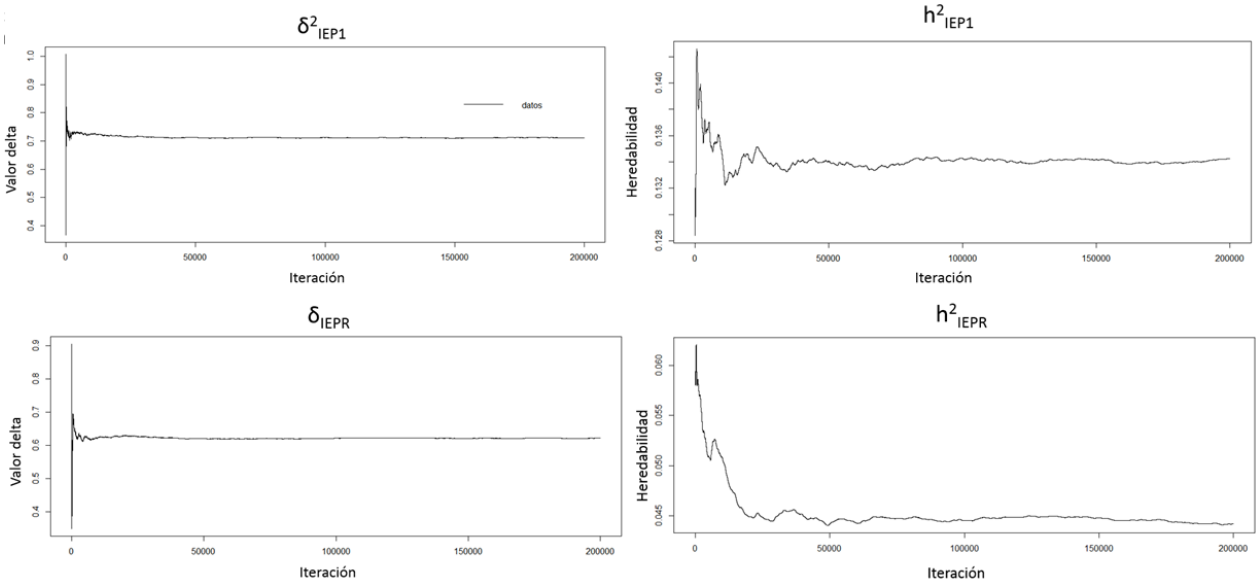

Figura 3.4. Promedios acumulados por iteración para de la heredabilidad y el parámetro de asimetría para los caracteres IEP1 e IEPR.

3.4. Discusión

En este capítulo se ajustó un modelo con errores asimétricos a observaciones del intervalo entre partos de vacas Holstein en Colombia, y se estimaron los parámetros genéticos utilizando el algoritmo del muestreo de Gibbs bajo un enfoque bayesiano. La metodología se implementó para dos caracteres: primer intervalo entre partos (IEP1) e intervalo entre los cuatro primeros partos (IEPR); en el segundo caso, se incluyó un efecto aleatorio para modelar las medidas repetidas dentro de un mismo individuo. La media posterior del parámetro de asimetría fue positiva para ambos caracteres y consistente con la distribución fenotípica observada de los intervalos. El método de comparación de modelos DIC favoreció el ajuste del modelo Normal asimétrico respecto del modelo con errores Normales que es frecuentemente empleado en la literatura para ajustar el carácter (Tiezzi et al. 2011).

La metodología empleada se basa en un miembro de la familia de distribuciones asimétricas presentadas por Sahu et al. (2003): la distribución Normal asimétrica multivariada. Los citados autores demostraron que modelos con efectos aleatorios que sigan la densidad SN pueden ajustarse sin complicaciones mediante métodos de cadenas de Markov y simulación Montecarlo (MCMC), como el muestreo de Gibbs. En particular, el esquema de muestreo empleado fue similar al presentado por Varona et al. (2008, Modelo 2), excepto para modelar la estructura de covarianza para medidas repetidas de un mismo individuo. Mientras que Varona et al. (2008, Modelo 3) desarrollaron un modelo jerárquico para el parámetro de asimetría, hecho que requiere definir un parámetro de asimetría para cada individuo, en la presente investigación se empleó una estructura de covarianzas más parsimoniosa, que modela una caída lineal de las correlaciones con el tiempo, que no requiere estimación de sus elementos no-diagonales, con elementos diagonales iguales uno y por lo tanto dependiente de un único parámetro de dispersión (Cantet et al. 2005).

En el contexto de la estimación de parámetros genéticos para caracteres cuantitativos, la implementación de un análisis bayesiano como el aquí desarrollado es una alternativa parsimoniosa para ajustar datos de campo que muestran una distribución asimétrica. Existen en la literatura ejemplos de aplicación de la metodología para registros del tamaño de la camada en cerdos, depresión endogámica y expresión génica (Casellas et al. 2008; Varona et al. 2008; Oliveira y Bueno Filo 2010). Contrariamente, en lechería no se registra utilización alguna del enfoque, si bien existen varios caracteres muestran cierto grado de asimetría (e.g., Tiezzi et al. 2011; Rönnegård et al. 2013). La asimetría en estos caracteres puede atribuirse, entre otras causas, a la existencia de un tratamiento preferencial en vacas de alta producción, que conlleva la ocurrencia de datos extremos (Strandén y Gianola 1999). Eliminar este tipo de registros parecería ser una práctica inapropiada, porque implica una pérdida de información importante para el análisis. En la medida que la asimetría pueda atribuirse a efectos no genéticos, la metodología aquí descripta se posiciona como una alternativa adecuada para estimar parámetros genéticos en caracteres lecheros.

Por otra parte, y dado que la asimetría de los efectos aumenta la dispersión de los estimadores (Banks et al. 1985), el modelo normal asimétrico permitiría reducir la varianza residual y, consecuentemente, incrementar la heredabilidad del carácter. La heredabilidad aquí estimada para IEP1 fue 0.13 con el modelo Normal asimétrico, valor que se encuentra dentro del rango de 0.04 a 0.15 informado en otros estudios con la raza Holstein (Dong y van Vleck 1989; Veerkamp et al. 2001). Por su parte, la heredabilidad para IEPR fue 0.04. Similar al IEP1, el valor se halla dentro del rango reportado para el carácter (Eghbalsaied 2011; Ghiasi et al. 2011; Haile-Mariam et al. 2013). Si bien los parámetros genéticos para el IEP1 y el IEPR pertenecen a la misma población, se espera que la heredabilidad para IEP1 sea mayor debido principalmente a que el IEPR depende de un mayor efecto ambiental por ser una medida repetida en el tiempo y considerar que una vaca presente un primer intervalo entre partos menor (días que transcurren entre el parto 1 y el parto 2) no garantiza que está presente menores intervalos entre partos posteriormente.

Para concluir, en el presente capítulo se implementó un modelo animal mixto con errores asimétricos, para caracteres asociados con el intervalo entre partos en bovinos lecheros. La metodología empleada permitió obtener estimaciones más precisas de los parámetros genéticos incluyendo solamente un parámetro adicional dentro del modelo.

Capítulo 4

Respuesta a la selección genómica en la población Holstein de Colombia

4.1. Introducción

El mejoramiento genético animal se realiza sobre la base de seleccionar los mejores individuos para uno o varios caracteres de importancia económica, para luego utilizarlos como padres del siguiente ciclo de producción. El progreso genético depende principalmente de la variabilidad genética, la intensidad de selección, la confiabilidad de los valores de cría y el intervalo generacional (Rendel y Robertson 1950; Burnside et al. 1992). El diseño de un esquema de mejoramiento eficiente en bovinos lecheros consiste en comparar el progreso genético de diferentes estrategias de selección y determinar cuál de ellas se traduce, efectivamente, en un mayor avance a nivel genético y económico de la población (Pryce y Daetwyler 2012; Thomasen et al. 2014).

Los programas de mejoramiento genético de bovinos lecheros en Colombia se basan en la importación de pajuelas para inseminación artificial (**IA**) de toros extranjeros, fundamentalmente norteamericanos. En los Estados Unidos dichos toros han sido tradicionalmente selectos a partir de una prueba de progenie (**PP**) con altas confiabilidades pero con un intervalo generacional considerablemente largo (Van Tassell y Van Vleck 1991; Pryce et al. 2010). En los últimos años, sin embargo, se han incorporado modificaciones sustanciales en el sistema de evaluación genética y en los programas de mejoramiento genético, a partir de la incorporación de los valores de cría genómicos (**VCG**) como criterio de selección. Los VCG se estiman combinando información fenotípica y genealógica con información genómica proveniente de un gran número de marcadores moleculares del tipo SNP (polimorfismos de nucleótido simple) distribuidos a lo largo del genoma. Esta técnica se conoce como 'Selección genómica' (**SG**, (Meuwissen et al. 2001; Wiggans et al. 2011)). Los valores de cría genómicos permiten una selección temprana y confiable de toros, hecho que se traduce en una reducción considerable del intervalo generacional: ya no es necesario esperar los primeros resultados de sus hijas en producción (Schaeffer 2006; Buch et al. 2012). Así, la SG ofrece a los criadores la oportunidad de reducir los costos de evaluación al obtener, a una edad temprana, confiabilidades que pueden variar entre 0,6 a 0,8 para caracteres con una heredabilidad de 0,2, cuando se cuenta con una cantidad superior a 4.000 registros fenotípicos para el carácter (Hayes et al. 2009). Por otro lado, la SG permite reducir los niveles de endogamia, debido a la posibilidad de evaluar un mayor número de individuos y, además, abrir la posibilidad de evaluar caracteres de difícil medición (Van Raden et al. 2009; Pryce y Daetwyler 2012).

Se ha evaluado el progreso genético potencialmente producido mediante distintos programas de mejoramiento genético para toros lecheros. Algunos programas se basan exclusivamente en selección genómica, otros en pruebas de progenie y algunos en la combinación de ambas estrategias (Pryce et al. 2010; Thomasen et al. 2014; Yamazaki et al. 2014). Los resultados de estos estudios indican una superioridad en el progreso genético basado en la selección genómica de los individuos, principalmente atribuida a la disminución del intervalo generacional. El objetivo de este capítulo fue calcular el progreso genético proyectado en la población Holstein de Colombia a partir de la utilización de pajuelas de toros genómicos estadounidenses. Para ello, se llevó a cabo una simulación determinística utilizando el enfoque presentado por Hill (1974).

4.2. Materiales y métodos

En los Estados Unidos el esquema de mejoramiento involucra cuatro vías de selección: padres de toros, madres de toros, padres de hembras y madres de hembras. En Colombia, en cambio, la mayoría de toros (aproximadamente el 90%) provienen de la importación de semen de toros, en su mayoría criados en Estados Unidos y Canadá; y las hembras son el resultado de la reposición interna de las madres. El modelo determinístico de la dinámica del progreso genético de un programa de selección genómica utilizando las 4 vías de selección en el núcleo (Figura 4.1) fue desarrollado a través de funciones programadas con el software *R* (*R* Core Team, 2015). La simulación consideró un único carácter que puede interpretarse, sin perder generalidad, como un genotipo agregado derivado de un índice de selección para varios caracteres de importancia económica.

$$\begin{bmatrix} \text{N♂ a N♂} & \text{N♀ a N♂} \\ \text{N♂ a N♀} & \text{N♀ a N♀} \end{bmatrix} = \begin{bmatrix} mm & mf \\ fm & ff \end{bmatrix}$$

Figura 4.1. Matriz de flujo de genes de mm: padres de machos, mf: padres de hembras, fm: madres de machos y ff: madres de las hembras.

Teniendo en cuenta la figura anterior cada bloque corresponde a una vía de selección representando la proporción de genes que provienen de las diferentes clases de edad. De esta manera, podemos construir la matriz de flujo de genes **P** donde sus filas y columnas representan una categoría de sexo y edad en las que son clasificados los individuos de la población. Esta matriz tiene dimensión $(h+k)\times(h+k)$, con h = 5 representando las clases de edad de los machos y k = 5 a las hembras (cuadro 4.1). De esta manera, $\mathbf{P}_{i,j}$ representa la proporción de genes de los individuos en la *i*-ésima categoría (sexo-edad) al momento *t* provenientes de individuos de la *j*-ésima categoría (sexo-edad) en el momento $t-1$. En el cuadro 4.1 se presenta la matriz de flujo de genes del núcleo bovino estudiado, obtenida a partir de las bases de datos de toros genómicos jóvenes (High Ranking Genomic Young) de la Asociación Holstein de Estados Unidos (www.holsteinusa.com). Por simplicidad, se asumirán ciclos reproductivos de un año y se utilizarán números correlativos para las diferentes clases de edad. Así, en la clase de edad 1 (o "año 1") se encuentran los individuos que apenas comienzan su vida reproductiva y que tendrán sus primeros hijos al año 2.

Cuadro 4.1. Matriz P de transmisión o flujo de genes dentro de la población del núcleo estudiada.

		Padres					Madres				
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Machos											
	1	0	0,32	0,14	0,02	0,02	0	0,35	0,12	0,02	0,01
	2	1	0	0	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0	0	0
	4	0	0	1	0	0	0	0	0	0	0
	5	0	0	0	1	0	0	0	0	0	0
Hembras											
	1	0	0,4	0,05	0,05	0	0	0,37	0,07	0,06	0
	2	0	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	0	0	1	0	0	0
	4	0	0	0	0	0	0	0	1	0	0
	5	0	0	0	0	0	0	0	0	1	0

De acuerdo a la metodología descrita por Hill (1974) luego de un único ciclo de selección (i.e., cuando $t = 1$) la respuesta a la selección será igual a

$$\mathbf{r}_1 = \mathbf{E_m}\,\mathbf{s}_m + \mathbf{E_f}\,\mathbf{s}_f$$

Donde las matrices $\mathbf{E_m}$ y $\mathbf{E_f}$, corresponden al paso de los genes por reproducción e indican la participación de las clases de sexo-edad en la producción de machos y hembras, respectivamente, permitiendo que estas participaciones sean diferentes (primera fila de $\mathbf{E_m}$ diferente a la fila *h+1* de $\mathbf{E_f}$). las matrices $\mathbf{E_m}$ y $\mathrm{E_f}$ son iguales a **P** pero con todos sus elementos iguales a 0, excepto para la filas 1 y *h+1*, respectivamente.

$\mathbf{S_m}$ y $\mathbf{S_f}$ corresponden a los vectores de diferenciales de selección genética de los machos y hembras de reemplazo, respectivamente, calculados de la siguiente manera.

$$\mathbf{s}_m = \boldsymbol{m}_{(0)} DS_{mm} + \boldsymbol{f}_{(0)} DS_{fm} \quad \text{y} \quad \mathbf{s}_f = \boldsymbol{m}_{(0)} DS_{mf} + \boldsymbol{f}_{(0)} DS_{ff}$$

Los vectores $m_{(0)}$ y $f_{(0)}$ corresponden a la proporción de genes derivados de machos y hembras, respectivamente, de clase de edad 1 en el tiempo *t* = 0. Para la matriz **P** que se presenta en el Cuadro 4.1, $m'_{(0)}$ y $f'_{(0)}$ corresponden a

$$m'_{(0)} = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$f'_{(0)} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]$$

En términos generales, el diferencial de selección (DS) para la *j*-ésima vía está dado por

$$DS_j = i_j \times r_j \times \sigma_g .$$

Donde j=1 corresponde a la vía (*mm*), j=2, 3, 4 corresponden a las vías *fm*, *mf* y *ff*, respectivamente donde el valor i_j corresponde a la intensidad de selección para cada una de las vias, r_j a las confiabilidades de la evaluación y σ_g al desvío estándar genético para el carácter evaluado. En este estudio se asumió *r* = 0,74 teniendo en cuenta las confiabilidades de los individuos de la base de datos de toros y hembras jóvenes con información genómica analizadas en EEUU. Adicionalmente, se evaluó la respuesta a la selección por generación teniendo en cuenta diferentes confiabilidades variando entre 0,1 a 0,99. Sobre la base de los DS, el progreso genético asintótico anual (ΔG_a) para las cuatro vías de selección fue calculado a través de la fórmula de Rendel y Robertson (1950):

$$\Delta G_a = \frac{\sum_j DS_j}{\sum_j L_j}$$

El valor L_j representa al intervalo generacional de la *j*-ésima vía de selección.

El intervalo generacional de cada una de las vías en el núcleo (Cuadro 4.2) fue calculado como el promedio de edad en años de los padres al nacimiento de su progenie seleccionada (Burnside et al. 1992). Las fechas de nacimiento de los padres y de la progenie selecta fueron obtenidas a partir de la base de datos genealógica de la asociación Holstein de Estados Unidos (www.holsteinusa.com). Las intensidades fueron calculadas de acuerdo al porcentaje de individuos seleccionados.

Cuadro 4.2. Intervalos generacionales (L) de cada una de las 4 vías de selección en el núcleo.

Vía de selección	L	i (% seleccionado)
Padre del Macho del Núcleo (*mm*)	2,48	2,06 (5%)
Madre del Macho del Núcleo (*fm*)	2,30	2,42 (2%)
Padre de la Hembra del Núcleo (*mf*)	2,38	1,40 (20%)
Madre de la Hembra del Núcleo (*ff*)	2,38	1,40 (20%)

La respuesta a la selección repetida puede obtenerse mediante el cálculo de las siguientes formulas, donde de manera general, para $t > 1$ la respuesta a la selección ($\mathbf{r}_t$) es igual a

$$\mathbf{r}_t = \mathbf{P}\mathbf{r}_{t-1} + \mathbf{E_m}\mathbf{Q}^{t-1}\mathbf{s}_m + \mathbf{E_f}\mathbf{Q}^{t-1}\mathbf{s}_f \quad (1)$$

Con una respuesta a la selección continúa igual a

$$\mathbf{R}_n = \sum_{t=1}^{n} \mathbf{r}_t \quad (2)$$

Q es una matriz igual a **P,** pero con las filas 1 y h + 1 con todos sus elementos iguales a cero. Hill (1974) la denominó la matriz de "envejecimiento", porque describe el flujo de genes a través del cambio de los individuos de una clase de edad a la siguiente. El vector **s** de dimensión $(h+k)$ contiene los diferenciales de selección de cada categoría de edad para machos y hembras.

4.2.1. Respuesta genética en Colombia utilizando toros jóvenes importados con evaluación genómica.

Tradicionalmente en Colombia, el progreso genético de la población Holstein se debe principalmente a la utilización de pajuelas de toros extranjeros - principalmente de EEUU – los que son machos del núcleo. Sin embargo, el intervalo generacional es mayor debido a que los toros llegan al país luego de ser empleados en EEUU y/o Canadá, donde son evaluados. En consecuencia, se calculó diferencialmente la respuesta genética anual extendiendo el esquema de cuatro vías de Hill (1974) a uno de seis vías de selección, cuya matriz **P** de flujo génico se muestra en la Figura 4.2. Las dos vías adicionales corresponden a los padres (mfc) y a las madres ($fcfc$) de las hembras comerciales. Los primeros representan a toros genómicos selectos anualmente del núcleo estadounidense, de los que se

importa material genético (principalmente semen en pajuelas). Estos toros llegan al país cuando tienen aproximadamente entre 2 y 4 años de edad y sus hijos nacen cuando tienen entre 3 a 5 años de edad. Por su parte, las madres de las vacas comerciales son hembras provenientes de los mismos hatos lecheros. Los intervalos generacionales de las vías mfc y $fcfc$ fueron obtenidos utilizando los catálogos de toros genómicos y una base de datos de Colombia, respectivamente. Por su parte, las confiabilidades empleadas fueron 0,74 para mfc y 0,50 para $fcfc$.

$$\begin{bmatrix} \text{N}\male \text{ a N}\male & \text{N}\female \text{ a N}\male & \varnothing \\ \text{N}\male \text{ a N}\female & \text{N}\female \text{ a N}\female & \varnothing \\ \text{N}\male \text{ a C}\female & \varnothing & \text{C}\female \text{ a C}\female \end{bmatrix} = \begin{bmatrix} mm & fm & \varnothing \\ mf & ff & \varnothing \\ mfc & \varnothing & fcfc \end{bmatrix}$$

Figura 4.2. Vías de selección consideradas para calcular la diseminación del progreso genético a la población comercial colombiana. El símbolo $\varnothing$ indica una submatriz de ceros.

Para el cálculo de la respuesta del progreso genético, la matriz **P** fue expandida en *q* filas y columnas, siendo *q* el número de clases de edad de las madres de vacas comerciales. Para este estudio, se trabajó con $q = 5$ y todos los elementos de los nuevos bloques de **P** corresponden a cero con excepción de:

$$\mathbf{P}[11,1{:}5] = [0,000 \quad 0,060 \quad 0,147 \quad 0,176 \quad 0,117]$$

$$\mathbf{P}[11,11{:}15] = [0,000 \quad 0,000 \quad 0,250 \quad 0,150 \quad 0,100]$$

$$\mathbf{P}[12,11] = \mathbf{P}[13,12] = \mathbf{P}[14,13] = \mathbf{P}[15,14] = 1$$

Para obtener la respuesta genética en la población comercial colombiana se calculó la diferencia esperada en performance entre los machos del núcleo y sus correspondientes contemporáneos en el hato comercial (Bichard et al. 1973; Hill 1974). Siguiendo a estos autores, la respuesta a la selección asintótica en la población comercial es igual a **Ds**, con **s** igual al vector de diferenciales de selección de cada categoría de edad y sexo en el núcleo y

$$\mathbf{D} = (\mathbf{I} - \mathbf{B})^{-1} - (\mathbf{I} - \mathbf{Q})^{-1}$$

donde $\mathbf{B} = \mathbf{P} - \mathbf{A}$ y $\mathbf{A} = \lim_{t \to \infty} \mathbf{P}^t$ representando la respuesta asintótica anual en el núcleo. Para expresar estas diferencias en unidades de tiempo (el denominado "retraso" o *lag*) es necesario multiplicar la matriz **D** por los intervalos generacionales promediados con respecto a todas las vías de selección; i.e.,

$$\mathbf{C} = 2\ L \times \mathbf{D}$$

con $L = \frac{L_{mm} + L_{mf} + L_{mfc} + L_{fm} + L_{ff} + L_{fcfc}}{6}$. En particular, el retraso de los machos de la clase de edad 1 del núcleo respecto al rodeo comercial corresponde a $lag_{mnc} = c_{1,1} + c_{1,\mathrm{h}+\mathrm{k}+1}$.

4.3. Resultados

De acuerdo a los cálculos obtenidos en este trabajo, en el núcleo Holstein americano la proporción de genes derivados de cualquier vía de selección (macho o hembra) presentes en los individuos nacidos en los años siguientes a la selección fluctúa inicialmente, pero se estabiliza en 0,21 aproximadamente a los 20 años. Sobre la base de dichos cálculos, el progreso genético anual en el núcleo luego de un solo ciclo de selección y un fracción de padres de machos selectos de 5%, 20% para la fracción de padres de las hembras, 20% para las madres de las hembras y 2% en la fracción selecta de madres de machos, es igual a $\Delta G_a = 0,56\sigma_g / año$. Considerando que el intervalo generacional promedio en el núcleo es de 2,39 años, el progreso genético acumulado por generación es de $\Delta G_{generación} = 1,34\sigma_g$. En este trabajo se asumió una confiabilidad promedio entre las vías de selección de 0,74. El progreso genético anual en desviaciones genéticas por año teniendo en cuenta diferentes confiabilidades se presenta en la Figura 4.3.

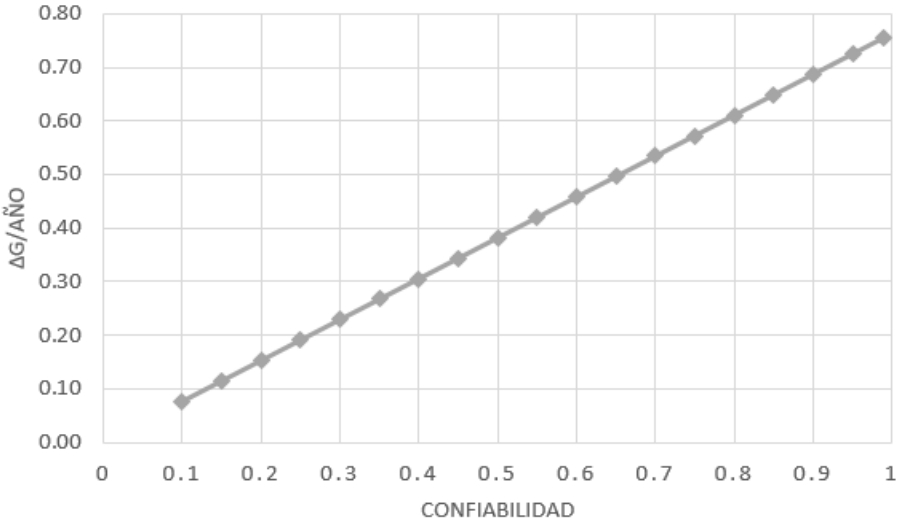

Figura 4.3. Respuesta a la selección por año (Δg/anual en σ_g /año) en el núcleo variando la confiabilidad de los valores de cría genómicos.

La respuesta genética en 5 años luego de un único ciclo de selección se presenta en el Cuadro 4.3. Al primer año, la respuesta a la selección en el núcleo fue de $0,956\sigma_g$, y luego muestra el patrón errático característico que recién se estabilizará a los 20 años con una respuesta de $0,560\sigma_g$. La respuesta acumulada a los 5 años teniendo un proceso de selección continua fue $2,413\sigma_g$. En términos generales se observa una mayor contribución al progreso genético de parte de los machos con una diferencia de $0,158\sigma_g$ con respecto a las hembras cuando ambos tienen 2 años de edad.

Cuadro 4.3. Respuesta en unidades de desviación típica genética en el núcleo luego de un ciclo de selección (machos, hembras y ambos) y la respuesta acumulada teniendo en cuenta la selección continúa.

Tiempo	**Machos**	**Hembras**	**Ambos**	**Acumulado**
1	0,000	0,000	0,000	0,000
2	0,557	0,399	0,956	0,956
3	0,214	0,062	0,276	1,232
4	0,434	0,343	0,777	2,009
5	0,282	0,122	0,404	2,413

4.3.1. Respuesta en la población colombiana

En el caso de la población comercial colombiana, los intervalos generacionales calculados para las vías de padres de hembras comerciales (*mfc*) y madres de hembras comerciales (*fcfc*) fueron 3,74 y 3,70 años, respectivamente. Teniendo en cuenta las 6 vías de selección, el intervalo generacional promedio es de L = 2,83 años. Por su parte, las intensidades de selección empleadas fueron $i_{mfc} = 1,40$ para la vía *mfc* e $i_{fcfc} = 0.4237$ para la vía *fcfc*. En este último caso se asumió una retención del 75% de las hembras nacidas cada año para reposición. El retraso o *lag* calculado entre los machos jóvenes del núcleo y sus contemporáneos en la población comercial fue 2,83 años lo cual corresponde a 1,18 generaciones con respecto al núcleo.

El progreso genético asintótico anual en la población comercial utilizando toros genómicos del núcleo fue $\Delta G_{ac} = 0.391\sigma_g / año$. Utilizando el enfoque de Hill (1974), la utilización de toros "genómicos" del núcleo en Colombia se traduciría en un progreso genético acumulado de $1,442\sigma_g$ en 5 años. Este valor difiere en $0,968\sigma_g$ con respecto al calculado en el núcleo. En la Figura 4.4 se presentan las diferencias entre el progreso genético acumulado en el núcleo y el rebaño comercial. Los cálculos fueron realizados asumiendo un proceso de selección continua de toros genómicos luego de 5 años.

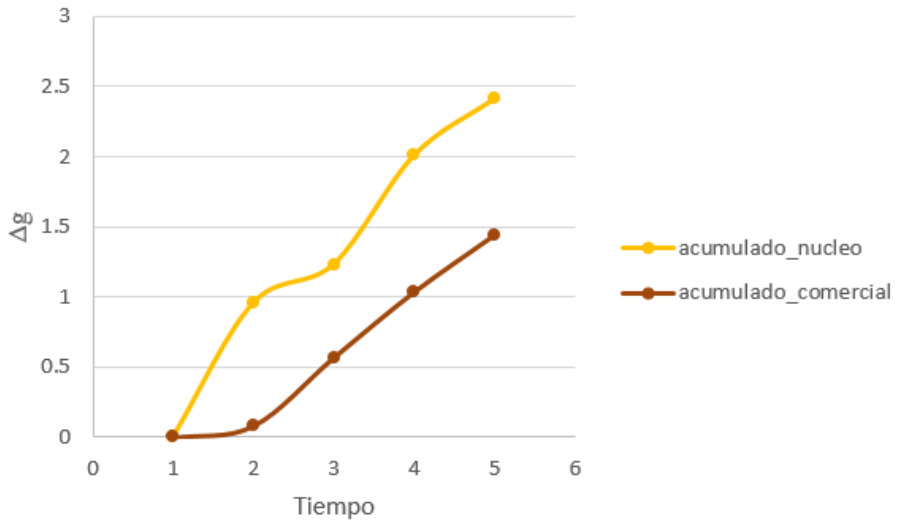

Figura 4.4. Progreso genético acumulado en desviaciones típicas genéticas para el núcleo y el rodeo comercial.

4.5. Discusión

En el presente estudio se calculó el progreso genético esperado al utilizar toros jóvenes, seleccionados en el núcleo por selección genómica, dentro de la población Holstein de Colombia, mediante el enfoque determinístico de Hill (1974). Utilizando datos reales de las bases de la población Holstein americanas se estimó un intervalo generacional en el núcleo de 2,39 años, valor similar al reportado por Schaeffer (2006) y Thomasen et al. (2014). Este intervalo reduce en aproximadamente 1,2 generaciones (3 años) al de la selección por prueba de progenie en el núcleo (Schaeffer 2006; Pryce et al. 2010). La reducción en el intervalo generacional beneficia el progreso genético del núcleo y, por diseminación, el del rodeo comercial. En este último caso, al disminuir los años que transcurren para que el material genético de toros jóvenes (con buenos valores de cría y una confiabilidad aceptable) llegue a ser utilizado en los hatos comerciales. Sin embargo, es necesario tener en cuenta que esta disminución podría resultar en un aumento en la consanguinidad cuando existe un número pequeño de candidatos a la selección (Hayes et al. 2009). Con respecto a la población comercial, una reducción en el intervalo generacional conlleva una disminución en el retraso ("lag", Hill 1974) de la repuesta a la selección en la población comercial contemporánea con los toros jóvenes selectos en el núcleo. Este retraso es inferior al esperado por la utilización de toros seleccionados mediante pruebas de progenie, las que - aunque más confiables - presentan intervalos generacionales mayores a 5 años para todas las vías de selección.

El progreso genético asintótico anual en el núcleo, calculado a través de la fórmula de Rendel and Robertson (1950) teniendo en cuenta selección continua fue superior en 0,02 σ_g respecto al reportado por Pryce et al. (2010), de 0,1 σ_g respecto al estimado por Schaeffer (2006) y de 0,17 σ_g con respecto al progreso genético esperado en el rodeo comercial. La fórmula de Rendel y Robertson (1950) asume que los padres de diferentes edades tienen igual oportunidad de producir progenie (Bichard et al. 1973), convirtiéndola en una excelente herramienta para evaluar estrategias de selección a largo plazo, pero no

refleja con precisión el progreso genético en los primeros años luego de aplicar un esquema de selección. En este estudio se calculó una diferencia en la respuesta genética en el primer año de aplicarse la selección en el núcleo igual a $0{,}39\,\sigma_g$ con respecto al progreso asintótico, indicando que si se realizara una selección anual de los individuos, el progreso genético sería efectivamente superior al que se reporta en la literatura (Pryce et al. 2010).

Los cálculos de diseminación del progreso genético por selección genómica del núcleo al rodeo comercial Holstein de Colombia indican altas tasas de respuesta y, en consecuencia, alientan a los productores colombianos a adaptarse al cambio de las estrategias actuales a través de la introducción de material genético importado de toros jóvenes con evaluación genómica para ser utilizados por medio de inseminación artificial. Por otra parte, el retraso generacional de la población comercial respecto al núcleo sugiere que podría obtenerse un mayor progreso genético si se realizaran evaluaciones genómicas de toros y vacas jóvenes en Colombia. Sin embargo, aunque puede esperarse un progreso genético superior en pequeñas poblaciones realizando evaluación genómica en comparación con esquemas de selección basados en pruebas de progenie (Thomasen et al. 2014), en el caso de Colombia sería conveniente plantear programas de selección genómica en colaboración con otros países, fundamentalmente por la necesidad de contar con una buena población de referencia que permita, a su vez, obtener una alta confiabilidad en la prueba. Por otra parte, la discusión sobre la conveniencia de implementar un programa de selección genómica para la población Holstein colombiana debería complementar el presente análisis con una detallada evaluación económica.

En conclusión, el progreso genético esperado al utilizar toros jóvenes seleccionados en el núcleo por selección genómica en la población Holstein de Colombia tendría potencial para producir una respuesta genética acelerada y una disminución considerable de la diferencia genética entre el hato comercial colombiano y el núcleo.

Capítulo 5

Discusión general

La presente tesis se ha centrado en el desarrollo e implementación de modelos estadísticos para la estimación de parámetros genéticos en bovinos lecheros, por un lado, y en la estimación del progreso genético a partir de la diseminación de toros evaluados mediante selección genómica, por otro. En primer lugar, se presentaron contribuciones teóricas y aplicadas relacionadas con la selección del orden del polinomio para los efectos genéticos aditivos y ambientales permanentes en un modelo de regresión aleatoria. En particular, se planteó la utilización de un método de selección de modelos novedoso y se ejemplificó su utilidad con una base de datos de producción de leche proveniente de vacas Holstein colombianas. Adicionalmente, se presentó la aplicación de un modelo Normal asimétrico a datos de intervalo entre partos en bovinos lecheros con el fin de evitar problemas de estimación de parámetros genéticos, cuando la distribución de los valores fenotípicos muestra asimetría. Por último, y a luz de la incorporación de la información genómica en los programas de evaluación de toros lecheros, se estimó el progreso genético esperado por diseminación utilizando toros genómicos en la población Holstein de Colombia.

5.1. Selección del orden del polinomio de Legendre

En el primer capítulo se abordó el problema de la selección de un orden del polinomio apropiado para los efectos genéticos aditivos y ambientales permanentes en un modelo de regresión aleatoria. El desarrollo fue enmarcado en un problema de selección de modelos. Tradicionalmente, para la estimación del orden del polinomio se han utilizado los criterios de información Akaike (AIC; Akaike 1974) y el criterio de información bayesiano (BIC, Schwaarz 1978). Sin embargo, en el contexto de la selección del orden del polinomio, los criterios AIC y BIC tienden a seleccionar o bien el modelo con el mayor orden del polinomio evaluado, o bien un modelo diferente cada uno de ellos (Bignardi et al. 2009; Herrera et al. 2013). Como alternativa, en este capítulo se introdujo el criterio PAL y se comparó su desempeño contra los criterios tradicionales (AIC y BIC). Los resultados indicaron un excelente desempeño del criterio PAL para seleccionar el modelo más apropiado bajo diferentes escenarios simulados, particularmente en términos de habilidad predictiva. Para la base de datos de producción de leche en bovinos Holstein de Colombia, los órdenes del polinomio para los efectos ambientales permanentes y genéticos aditivos fueron 6 y 3, respectivamente. La utilización del criterio PAL requiere una estructura anidada de los modelos a comparar. Esto implica que, para su implementación, se requiere la evaluación de todos los modelos posibles, desde el más básico hasta aquel que tiene un mayor orden del polinomio.

5.2. Asimetría de los residuales en un modelo animal

En el segundo capítulo se abordó el problema de estimación de parámetros genéticos para caracteres que muestran una distribución asimétrica de valores fenotípicos, un problema común en caracteres de importancia económica en lechería. En estos casos, los modelos lineales mixtos comúnmente utilizados asumen erróneamente una distribución normal de los errores. Sin embargo, lo apropiado es utilizar distribuciones alternativas (Stranden y Gianola 1999; Varona et al. 2008). En este capítulo se introdujo el modelo Normal asimétrico como alternativa y se describió la implementación de un análisis bayesiano jerárquico para obtener estimaciones de los parámetros genéticos. A modo de

ilustración, se ajustó el modelo a datos de intervalo entre partos en bovinos lecheros y se obtuvieron estimaciones de los parámetros genéticos. En comparación con el modelo que asume errores normales, el modelo Normal asimétrico permitió obtener estimaciones más precisas de los parámetros genéticos, reducir la varianza residual y, como consecuencia, aumentar la heredabilidad del carácter analizado, a costa de incorporar un único parámetro adicional. En el contexto de la producción animal, otros estudios utilizaron este enfoque para estimar el efecto de la depresión endogámica sobre la longevidad en cerdas (Casellas et al. 2008) y para la estimación de parámetros genéticos para tamaño de la camada también en cerdos (Varona et al. 2008).

5.3. Selección genómica en la población colombiana

Finalmente, en el capítulo 3 se estimó el progreso genético esperado en la población Holstein de Colombia por diseminación de toros genómicos provenientes del núcleo de la raza. El desarrollo siguió las tendencias actuales en el sentido de evaluar toros lecheros por selección genómica y su posterior utilización, para países en los cuales se importa material genético (i.e., pajuelas de semen). Los cálculos permiten predecir un aumento acelerado en la respuesta genética en la población comercial y una reducción considerable de la diferencia genética con el núcleo. Si bien es dable esperar un mayor progreso si el país implementase un programa nacional de evaluación genómica de toros y vacas, el éxito de tal estrategia requeriría no sólo un buen diseño del programa para asegurar una buena población de referencia, sino también un pormenorizado estudio de la relación costo-beneficio.

5.4. Conclusión general

Las metodologías empleadas en los capítulos de esta tesis presentan soluciones a los problemas de selección y estimación de los parámetros genéticos en bovinos lecheros. La introducción del PAL como criterio de selección del orden del polinomio en un modelo de regresión aleatoria mostró un excelente desempeño para seleccionar un modelo apropiado dentro de los distintos escenarios simulados. Al mismo tiempo, la introducción del modelo normal asimétrico utilizando una metodología bayesiana permitió obtener estimaciones más precisas de los parámetros genéticos, reducir la varianza residual y, como consecuencia, incrementar la heredabilidad del carácter analizado. Finalmente, el análisis de progreso genético utilizando toros genómicos en la población Holstein de Colombia permitió dilucidar un aumento acelerado en la respuesta por selección y una reducción considerable de la diferencia genética entre el hato comercial y el núcleo.

BIBLIOGRAFIA

Akaike, H. 1974. A New Look at the Statistical Model Identification. IEEE Trans. Autom. Control 19:716–723.

Banks, B. D., I. L. Mao, and J. P. Walter. 1985. Robustness of the Restricted Maximum Likelihood Estimator Derived Under Normality as Applied to Data with Skewed Distributions. J. Dairy Sci. 68:1785–1792.

Bichard, M., A. H. R. Pease, P. H. Swales, and K. Özkütük. 1973. Production : Selection in a population with overlapping. Anim. Prod. 17:215–227.

Bignardi, A. B., L. El Faro, V. L. Cardoso, P. F. Machado, and L. G. de Albuquerque. 2009. Random regression models to estimate test-day milk yield genetic parameters Holstein cows in Southeastern Brazil. Livest. Sci. 123:1–7.

Bohmanova, J., F. Miglior, J. Jamrozik, I. Misztal, and P. G. Sullivan. 2008. Comparison of random regression models with Legendre polynomials and linear splines for production traits and somatic cell score of Canadian Holstein cows. J. Dairy Sci. 91:3627–3638.

Buch, L. H., M. K. Sørensen, P. Berg, L. D. Pedersen, and A. C. Sørensen. 2012. Genomic selection strategies in dairy cattle : Strong positive interaction between use of genotypic information and intensive use of young bulls on genetic gain. J. Anim. Breed. Genet. 129:138–151.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65:23–35.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A practical Information-Theoretic Approach. 2nd ed. Springer Verlag, New York, USA.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol. Methods Res. 33:261–304.

Burnside, E., G. Jansen, G. Civati, and D. E. 1992. Observed and theoretical genetic trends in a large dairy population under intensive selection. J. Dairy Sci. 75:2242–2253.

Cantet, R. J. C., A. N. Birchmeier, A. W. C. Cayo, and C. Fioretti. 2005. Semiparametric animal models via penalized splines as alternatives to models with contemporary groups. J. Anim. Sci. 83:2482–2494.

Carabaño, M. J., C. Díaz, C. Ugarte, and M. Serrano. 2007. Exploring the use of random regression models with legendre polynomials to analyze measures of volume of ejaculate in Holstein bulls. J. Dairy Sci. 90:1044–57.

Casella, G., and R. Berger. 2002. Statistical Inference. 2nd ed. Duxbury, Thomson Learning, CA,USA.

Casellas, J., L. Varona, N. Ibáñez-Escriche, R. Quintanilla, and J. L. Noguera. 2008. Skew distribution of founder-specific inbreeding depression effects on the longevity of Landrace sows. Genet. Res. cambridge 90:499–508.

Darwash, a O., G. E. Lamming, and J. a Woolliams. 1997. Estimation of genetic variation in the interval from calving to postpartum ovulation of dairy cows. J. Dairy Sci. 80:1227–34.

Davidson, R., and J. Mackinnon. 2004. Econometric Theory and Methods. Oxford University Press, New York, USA.

Dong, M. C., and L. D. van Vleck. 1989. Correlations Among First and Second Lactation Milk Yield and Calving Interval. J. Dairy Sci. 72:1933–1936.

Eghbalsaied, S. 2011. Estimation of genetic parameters for 13 female fertility indices in Holstein dairy cows. Trop. Anim. Health Prod. 43:811–6.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2009. Bayesian Data Analysis. Vol 2. Chapman & Hall/CRC.

Geweke, J. 1992. Bayesian Statistics: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: J. Bernardo, J. Berger, P. Dawid, and A. Smith, editors. 4th ed. Oxford University Press, New York. p. 169–193.

Ghiasi, H., A. Pakdel, A. Nejati-Javaremi, H. Mehrabani-Yeganeh, M. Honarvar, O. González-Recio, M. J. Carabaño, and R. Alenda. 2011. Genetic variance components for female fertility in Iranian Holstein cows. Livest. Sci. 139:277–280.

Gianola, D., and R. L. Fernando. 1986. Bayesian Methods in Animal Breeding Theory. J. Anim. Sci. 63:217–244.

Gilmour, A. R., R. Thompson, and B. R. Cullis. 1995. Linear Mixed Models Algorithm for Average Information REML : An Efficient in Linear Mixed Models Variance Parameter Estimation. Biometrics 51:1440–1450.

Haile-Mariam, M., P. J. Bowman, and J. E. Pryce. 2013. Genetic analyses of fertility and predictor traits in Holstein herds with low and high mean calving intervals and in Jersey herds. J. Dairy Sci. 96:655–67.

Harville, D. 1997. Matrix algebra from a statistician's perspective. Springer, New York, USA.

Hayes, B. J., P. J. Bowman, a J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–443.

Heinrichs, a. J., and M. Vazquez-Anon. 1993. Changes in First Lactation Dairy Herd Improvement Records. J. Dairy Sci. 76:671–675.

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada.

Henderson, Crj. 1982. Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. Biometrics 38:623–640.

Herrera, A. C., O. D. Munera, and M. F. Cerón-Muñoz. 2013. Variance components and genetic parameters for milk production of Holstein cattle in Antioquia (Colombia) using random regression models. Rev. Colomb. Ciencias Pecu. 26:90–97.

Hill, W. G. 1974. Prediction and evaluation of response to selection with overlapping

generations. Anim. Prod. 18:117–139.

Hober, J., and G. Casella. 1996. The effect of improper prior on gibbs sampling in hierarchical linear mixed models. J. Am. Stat. Assoc. 91:1461–1473.

Hurvich, C. M., and C. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Jakobsen, J. H., P. Madsen, J. Jensen, J. Pedersen, L. G. Christensen, and D. a Sorensen. 2002. Genetic parameters for milk production and persistency for Danish Holsteins estimated in random regression models using REML. J. Dairy Sci. 85:1607–1616.

Jamrozik, J., D. Gianola, and L. R. Schaeffer. 2001. Bayesian estimation of genetic parameters for test day records in dairy cattle using linear hierarchical models. Livest. Prod. Sci. 71:223–240.

Jamrozik, J., and L. R. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. J. Dairy Sci. 80:762–770.

Jamrozik, J., I. Strandén, and L. R. Schaeffer. 2004. Random regression test-day models with residuals following a Student's-t distribution. J. Dairy Sci. 87:699–705.

Kass, R., B. Carlin, A. Gelman, and R. Neal. 1998. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. Am. Stat.:93–100.

Kirkpatrick, M., D. Lofsvold, and M. Bulmer. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. Genetics 124:979–93.

Kuhn, M. T., P. J. Boettcher, and a. E. Freeman. 1994. Potential Biases in Predicted Transmitting Abilities of Females from Preferential Treatment. J. Dairy Sci. 77:2428–2437.

Lachos, V. H., D. K. Dey, and V. G. Cancho. 2009. Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. J. Stat. Plan. Inference 139:4098–4110.

Liu, Y. X., J. Zhang, L. R. Schaeffer, R. Q. Yang, and W. L. Zhang. 2006. Short communication: Optimal random regression models for milk production in dairy cattle. J. Dairy Sci. 89:2233–2235.

López-Romero, P., and M. Carabano. 2003. Comparing alternative random regression models to analyse first lactation daily milk yield data in Holstein – Friesian cattle. Livest. Prod. Sci. 82:81–96.

López-Romero, P., R. Rekaya, and M. Carabaño. 2003. Assessment of homogeneity vs. heterogeneity of residual variance in random regression test-day models in a Bayesian analysis. J. Dairy Sci. 86:3374–3385.

McQuarrie, A., R. Shumway, and C.-L. Tsai. 1997. The model selection criterion AICu. Stat. Probab. Lett. 34:285–292.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157:1819–1829.

Meyer, K., and W. G. Hill. 1997. Estimation of genetic and phenotypic covariance

functions for longitudinal or “repeated” records by restricted maximum likelihood. Livest. Prod. Sci. 47:185–200.

Meyer, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. Genet. Sel. Evol. 30:221–240.

Meyer, K. 2007. WOMBAT – A tool for mixed model analyses in quantitative genetics by REML. J Zhejiang Univ Sci B 8:815–821.

Müller, S., J. L. Scealy, and a. H. Welsh. 2013. Model Selection in Linear Mixed Models. Stat. Sci. 28:135–167.

Odegård, J., J. Jensen, G. Klemetsdal, P. Madsen, and B. Heringstad. 2003. Genetic analysis of somatic cell score in Norwegian cattle using random regression test-day models. J. Dairy Sci. 86:4103–14.

Oliveira, D., and J. Bueno Filo. 2010. Análise bayesiana de modelos mistos normais assimétricos em dados de expressao génetica originados de pedigrees complexos. Rev. Bras. Biometria 28:137–160.

Pereira, R. J., A. B. Bignardi, L. El Faro, R. S. Verneque, A. E. Vercesi Filho, and L. G. Albuquerque. 2013. Random regression models using Legendre polynomials or linear splines for test-day milk yield of dairy Gyr (Bos indicus) cattle. J. Dairy Sci. 96:565–574.

Plaizier, J. C., K. D. Lissemore, D. Kelton, and G. J. King. 1998. Evaluation of overall reproductive performance of dairy herds. J. Dairy Sci. 81:1848–54.

Pool, M. H., and T. H. E. Meuwissen. 2000. Reduction of the number of parameters needed for a polynomial random regression test day model. Livest. Prod. Sci. 64:133–145.

Pryce, J. E., and H. D. Daetwyler. 2012. Designing dairy cattle breeding schemes under genomic selection : a review of international research. :107–114.

Pryce, J. E., M. E. Goddard, H. W. Raadsma, and B. J. Hayes. 2010. Deterministic models of breeding scheme designs that incorporate genomic selection. J. Dairy Sci. 93:5455–5466.

R, C. T. 2015. R: A language and environment for statistical computing. Available from: http://www.r-project.org/

Rekaya, R., M. Carabano, and M. Toro. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. Livest. Prod. Sci. 57:203–217.

Rendel, J., and A. Robertson. 1950. The use of progeny testing with artificial insemination in dairy cattle. J. Genet. 50:21–31.

Rodriguez-Zas, S. L., D. Gianola, and G. E. Shook. 1998. Bayesian analysis via Gibbs sampling of susceptibility to intramammary infection in Holstein cattle. J. Dairy Sci. 81:2710–22.

Rönnegård, L., M. Felleki, W. F. Fikse, H. a Mulder, and E. Strandberg. 2013. Variance component and breeding value estimation for genetic heterogeneity of residual variance in Swedish Holstein dairy cattle. J. Dairy Sci. 96:2627–36.

Sahu, K., D. K. Dey, and M. D. Branco. 2003. A new class of multivariate skew

distributions with applications to Bayesian regression models. Can. J. Stat. 31:129–150.

Schaeffer, L. R. 2004. Application of random regression models in animal breeding. Livest. Prod. Sci. 86:35–45.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. 123:218–223.

Schwarz, G. 1978. Estimating the Dimension of a Model. Ann. Stat. 6:461–464.

Searle, S. 1982. Matrix Algebra Useful for Statistics. John Wiley & Sons, New York, USA.

Shibata, R. 1981. An optimal selection of regression variables. Biometrika 68:45–54.

Smith, B. J. 2007. boa : An R Package for MCMC Output Convergence. J. Stat. Softw. 21:1–37.

Sorensen, D., and D. Gianola. 2002. Likelihood, bayesian and MCMC methods in quantitative genetics. Springer Verlag, New York, USA.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Statistical Methodol. 64:583–639.

Stoica, P., and P. Babu. 2013. Model order estimation via penalizing adaptively the likelihood (PAL). Signal Processing 93:2865–2871.

Strabel, T., J. Szyda, E. Ptak, and J. Jamrozik. 2005. Comparison of random regression test-day models for Polish Black and White cattle. J. Dairy Sci. 88:3688–3699.

Strandén, I., and D. Gianola. 1999. Mixed effects linear models with t-distributions for quantitative genetic analysis : Bayesian approach. Genet. Sel. Evol. 31:25–42.

Van Tassell, C. P., and L. Van Vleck. 1991. tes of genetic selection differentials and generation intervals for four paths of selection. J. Dairy Sci. 74:1078–1086.

Thomasen, J. R., A. Willam, B. Guldbrandtsen, M. S. Lund, and A. C. Sørensen. 2014. Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. J. Dairy Sci. 97:458–470.

Tiezzi, F., C. Maltecca, M. Penasa, a Cecchinato, Y. M. Chang, and G. Bittante. 2011. Genetic analysis of fertility in the Italian Brown Swiss population using different models and trait definitions. J. Dairy Sci. 94:6162–72.

Vanraden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, and J. F. Taylor. 2009. Invited review : Reliability of genomic predictions for North American Holstein bulls. :16–24.

Varona, L., N. Ibañez-Escriche, R. Quintanilla, J. L. Noguera, and J. Casellas. 2008. Bayesian analysis of quantitative traits using skewed distributions. Genet. Res. (Camb). 90:179–90.

Veerkamp, R. F., E. P. C. Koenen, and G. De Jong. 2001. Genetic Correlations Among Body Condition Score, Yield, and Fertility in First-Parity Cows Estimated by Random Regression Models. J. Dairy Sci. 84:2327–2335.

Weigel, K. a, and R. Rekaya. 2000. Genetic parameters for reproductive traits of Holstein cattle in California and Minnesota. J. Dairy Sci. 83:1072–80.

Van Der Werf, J. H. J., M. E. Goddard, and K. Meyer. 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. J. Dairy Sci. 81:3300–3008.

Wiggans, G. R., P. M. Vanraden, and T. A. Cooper. 2011. The genomic evaluation system in the United States : Past , present , future. J. Dairy Sci. 94:3202–3211.

Winkelman, a M., D. L. Johnson, and a K. MacGibbon. 1999. Estimation of heritabilities and correlations associated with milk color traits. J. Dairy Sci. 82:215–24.

Yamazaki, T., K. Togashi, S. Iwama, S. Matsumoto, K. Moribe, T. Nakanishi, K. Hagiya, and K. Hayasaka. 2014. effects of a breeding scheme combined by genomic pre-selection and progeny testing on annual genetic gain in a dairy cattle population. Anim. Sci. J. 85:639-649.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. Biometrika 92:937–950.