Anales Científicos 80 (1): 53-59 (2019)

Análes Científicos

ISSN 2519-7398 (Versión electrónica)
Website: http://revistas.lamolina.edu.pe/index.php/acu/index

Análisis de residuales en un modelo lineal mixto para estimar heredabilidad

Analysis of residuals in a linear mixed model to estimate heritability

Ana Vargas Paredes [1*]; Víctor Maehara Oyata[2]

[1] Universidad Nacional Agraria La Molina, Lima, Perú. Email: anavargas@lamolina.edu.pe; vmaehara@lamolina.edu.pe

Recepción: 06/03/2019 ; Aceptación: 05/06/2019

Resumen

Los modelos lineales mixtos son utilizados como métodos para la estimación de parámetros genéticos como heredabilidad, en los que se incorpora información genealógica de los individuos observados. En este trabajo, se analizan los resultados de un modelo lineal mixto ajustado para estimar heredabilidad vía el enfoque de máxima verosimilitud restringida (REML), enfocado en el análisis de residuales mediante herramientas exploratorias en función a tres tipos de residuales (marginal, condicional y efectos aleatorios), siguiendo la propuesta dada por Singer adaptando algunas funciones hechas en "R" para incorporar la información genealógica en el modelo.

Palabras clave: análisis de residuales; modelo lineal mixto; heredabilidad.

Abstract

Linear mixed models are used as methods to estimate genetic parameters such as heritability in which genealogical information from observed individuals is incorporated into the model. In this work, we analyze the results of linear mixed model adjusted to estimate heritability via the restricted maximum likelihood (REML) approach, focused on the analysis of residuals through exploratory tools according to three types of residuals (marginal, conditional and random effects). Following the proposal given by Singer we adapted some functions made in "R" to incorporate the genealogical information into the model.

Keywords: Residual analysis; linear mixed models; heritability.

Forma de citar el artículo: Vargas, A.; Maehara, V. 2019. Análisis de residuales en un modelo lineal mixto para estimar heredabilidad. Anales Científicos 80 (1): 53-59 (2019).

DOI:http://dx.doi.org/10.21704/ac.v80i1.1375
Autor de correspondencia: Ana Vargas Paredes. Email: anavargas@lamolina.edu.pe

1. Introducción

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

$$y = X\beta + Zu + e \quad (1)$$

donde, X y Z son matrices de incidencias conocidas, u y e vectores de efectos aleatorios tal que $\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}\right)$, y, G y R son matrices de varianzas y covarianzas que están en función de los parámetros de dispersión. El vector u incluye el efecto genético aditivo entre otros efectos y la matriz G incluye la matriz A de relaciones genéticas aditivas entre los individuos, la cual se construye a partir de la información genealógica o pedigrí.

El método de estimación más popular para estimar componentes de varianza es el de máxima verosimilitud restringida (REML). Para verificar la validez de los supuestos y evaluar la confiabilidad de la inferencia estadística se realiza el análisis de residuales. En un modelo lineal gaussiano los residuales son usados para verificar linealidad de efectos, normalidad, independencia, homocedasticidad de errores y presencia de observaciones atípicas.

2. Materiales y métodos

Diagnóstico del modelo: análisis de residuales

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

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

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

Los diferentes usos para los tres tipos de residuales son resumidos por Singer *et al.* (2013), quienes lo adaptaron de Nobre y Singer (2007) y son presentados en la Tabla 1.

Lesaffre y Verbeke, citado por Singer *et al.* (2013), comentaron que cuando la estructura dentro de las unidades es adecuada, $\mathbf{V}_i = \left\| I_{m_i} - \mathbf{R}_i \mathbf{R}_i^T \right\|^2$, donde $\hat{\mathbf{R}}_i = \hat{\phi}^{1/2}_i \hat{\Lambda}_i$ con

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}(\theta) = \mathbf{Z}\boldsymbol{\Sigma}_\theta \mathbf{Z}^T \sigma_e^2 \mathbf{I} = \mathbf{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^T \mathbf{Z}^T \sigma^2 + \sigma_e^2 \mathbf{I}$$

debe ser cercana a cero. Unidades con valores grandes de $\mathbf{V}_i$ indicaría que la estructura de covarianza puede no ser adecuada para dichas observaciones. Singer *et al*. (2013) recomienda reemplazar $\mathbf{R}_i$ en $\mathbf{V}_i$ con el residual marginal estandarizado

$\hat{\xi}_i^* = \left[V(\hat{\xi}_i)\right]^{-1/2} \hat{\xi}_i$, donde $\hat{\xi}_i$ corresponden al elemento de la diagonal $\boldsymbol{\Omega} - \mathbf{X}(\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T$ asociado con la i-ésima unidad. Además, recomendaron utilizar $\mathbf{v}_i^* = \sqrt{\mathbf{v}_i / m_i}$ como una medida estandarizada de adecuación de la estructura de covarianza dentro de las unidades.

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

Nobre y Singer (2007) observaron que los residuales condicionales pueden tener varianzas diferentes, por lo que sugirieron graficar los residuales estandarizados condicionales $\hat{\mathbf{e}}_{ij}^* = \hat{\mathbf{e}}_{ij} / diag(\mathbf{Q})^{1/2}$, donde $\mathbf{Q} = \mathbf{U} - \mathbf{X}(\mathbf{X}^T \mathbf{U}^{-1} \mathbf{X})^{-1} \mathbf{X}^T$ versus los valores ajustados para chequear homocedasticidad de los errores condicionales o versus índice de observaciones para chequear observaciones atípicas.

Tabla 1. Usos de residuales para propósitos de diagnóstico

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

Fuente: Singer *et al.* (2013).

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

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

donde $1 \geq \lambda_1 \geq ... \geq \lambda_N$ son valores ordenandos de $\mathbf{\Lambda}_\theta$ obtenidos de la descomposición del valor singular $\mathbf{Q} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^T$, $\mathbf{L}^T\mathbf{L} = \mathbf{I}$, l_k, y representa la k-ésima columna de L. Los residuales mínimos confundidos estandarizados $\mathbf{c}_k^T\hat{\mathbf{e}}_{ij}^{*}$ pueden ser obtenidos dividiendo $\mathbf{c}_k^T\hat{\mathbf{e}}_{ij}^{*}$ por la raíz cuadrada de l o s elementos correspondiente en CQC^T donde $C = (C_1, ..., C_{N-P})^T$. El gráfico QQ-plot de los residuales mínimos confundidos estandarizados, $\mathbf{c}_k^T\hat{\mathbf{e}}_{ij}^{*}$, se emplea para chequear normalidad.

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

Descripción de los datos

Los datos son registros de 3397 lactaciones del primer al quinto parto de 1359 vacas Holstein, hijas de 38 toros en 57 rebaños. Todos los registros corresponden a vacas con al menos 100 días de leche. La información genealógica, pedigrí, de estas vacas comprende 5 generaciones con un total de 6547 animales. Toda esta información ha sido descargada desde United State Department of Agriculture USDA, http://www.aipl.arsusda.gov/, 2010 y están disponibles en el conjunto de datos *milk* y *pedCows* de la librería *pedigreemm* en *R*. (Vázquez *et al.*, 2010).

A partir del modelo animal formulado por Vázquez *et al.* (2010) se estimaron los tres tipos de residuales estandarizados y otros, para luego obtener los gráficos con fines de diagnóstico como se especificó en la Tabla 1, adaptando el código y las funciones en R proporcionadas por Singer *et al.* (2013) para este modelo en particular.

2. Resultados y discusión

La Figura 1 permite revisar la linealidad del efecto fijo de la covariable días en leche en el modelo, no se muestra un patrón por lo que indicaría que el logaritmo de días en leche tiene un efecto lineal en la producción de leche. Otro gráfico que evalúa la linealidad de los efectos fijos se muestra en la Figura 2, donde no se observa algún patrón definido por lo que no se descartaría la relación lineal con

los efectos fijos; asimismo, el histograma de la distribución de los residuales marginales estandarizados muestra un comportamiento simétrico.

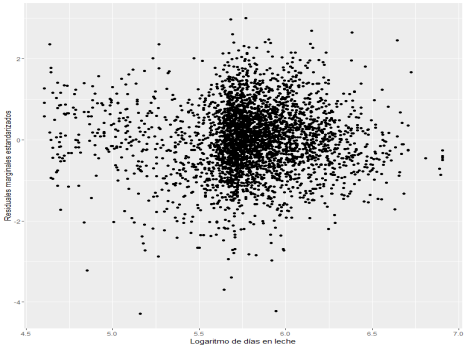


Figura 1. Residuales marginales estandarizados vs log (días en leche)

La Figura 3 presenta los residuales marginales ajustados versus los índices de observación, a partir de este gráfico se encuentra que hay 95 observaciones atípicas de 45 animales. Se consideraron atípicas aquellas observaciones con residuales, en términos absolutos, mayores que 2.

El gráfico de los residuales condicionales estandarizados versus las observaciones estimadas, Figura 4, sugiere que no habría homocedasticidad de los errores condicicionales, puesto que no se observa un patrón aleatorio, así también el histograma nos sugiere un comportamiento simétrico de estos residuales.

En la Figura 5, se muestra a los residuales condicionales ajustados versus los índices de observación, desde el cual se encuentra que hay 234 observaciones atípicas de 109 animales, las observaciones consideradas como atípicas son aquellas con residuales, en términos aboslutos, mayores que 2.

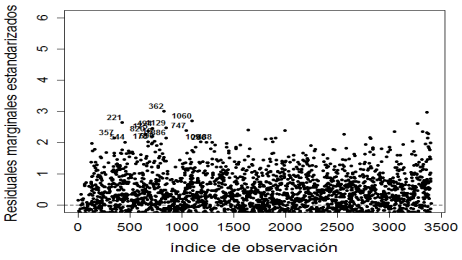


Figura 3. Residuales marginales estandarizados vs índices de observación

En la Figura 6 se observa que 215 animales pueden considerarse atípicos. Se consideraron individuos atípicos aquellos cuya distancia de Mahalanobis fue mayor a dos veces la media de las distancias estimadas para cada animal.

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

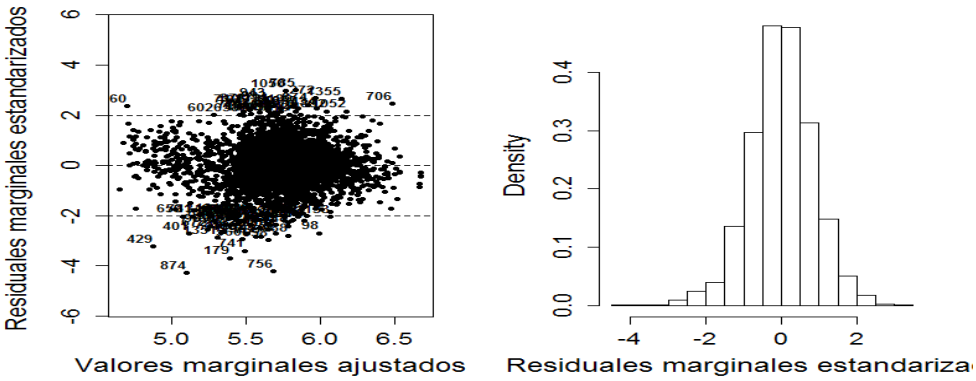


Figura 2. Residuales marginales estandarizados vs ajustados e histograma de los residuales marginales

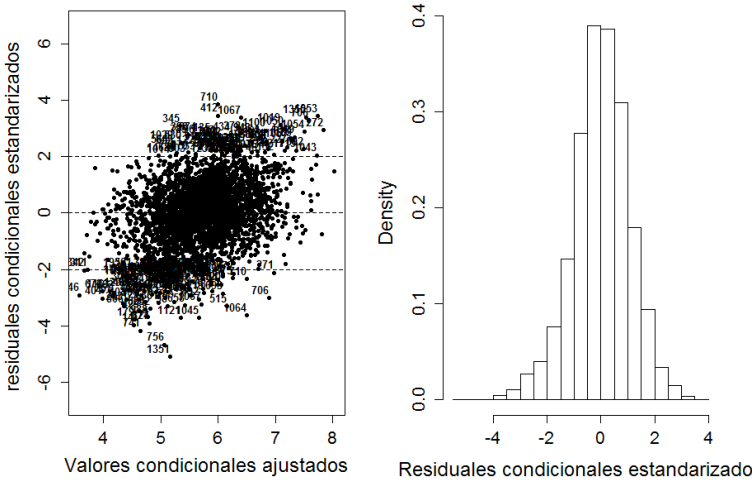


Figura 4. Residuales condicionales estandarizados vs ajustados e histograma de los residuales condicionales

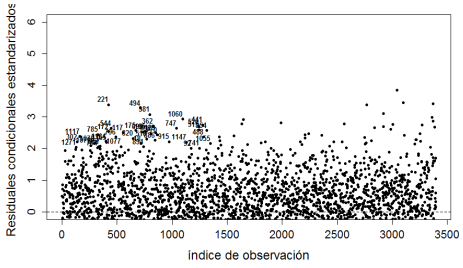


Figura 5. Residuales condicionales estandarizados vs índices de observación

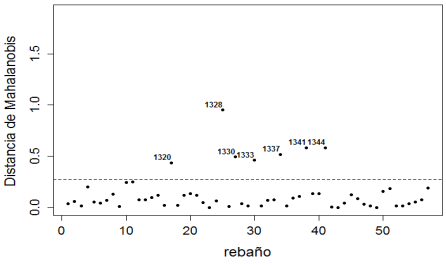


Figura 7. Distancia estandarizada de Mahalanobis vs índices de rebaño

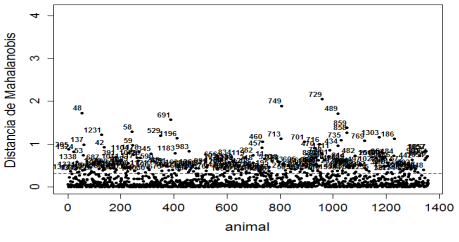


Figura 6. Distancia estandarizada de Mahalanobis vs índices de animal

El gráfico QQ-plot chi-cuadrado para la distancia de Mahalanobis para el efecto aleatorio de animal mostrado en la Figura 8, muestra un comportamiento que ajusta a una distribución normal. Sin embargo, el gráfico QQ-plot chi-cuadrado para la distancia de Mahalanobis para el efecto aleatorio rebaño mostrado en la Figura 9, muestra un comportamiento que podría no ajustar a una distribución normal.

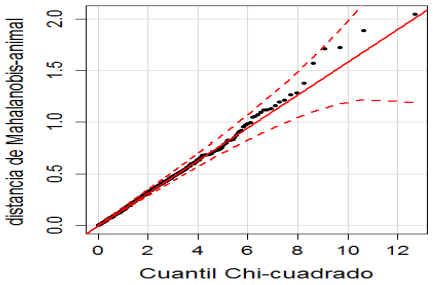


Figura 8. QQ plot chi-cuadrado para distancia estandarizada de Mahalanobis – animal

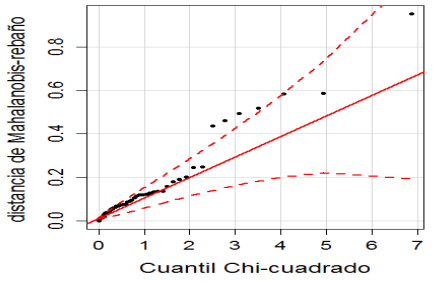


Figura 9. QQ plot chi-cuadrado para distancia estandarizada de Mahalanobis – rebaño

La Figura 10 muestra las distancias estandarizadas de Lesaffre y Verbeke versus los animales, de este gráfico se tiene que 132 animales tienen distancias mayores a dos veces el valor de la distancia promedio estimada, por lo que para estos animales la estructura de varianzas y covarianza no sería muy adecuada.

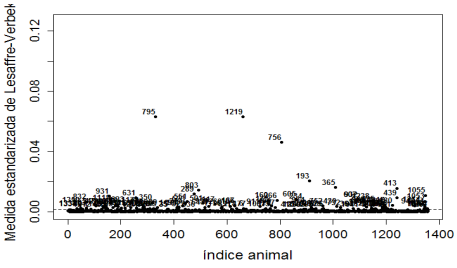
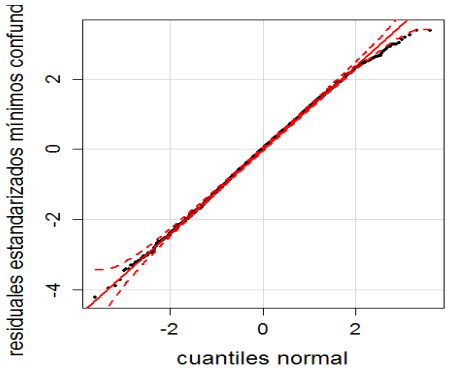


Figura 10. Medida estandarizada de Lesaffre-Verbeke vs animal

La Figura 11 muestra, a través del gráfico QQ plot normal para los residuales estandarizados condicionales mínimos confundidos, que estos no tendrían un comportamiento normal, pese a que tienen una forma simétrica como muestra su respectivo histograma.



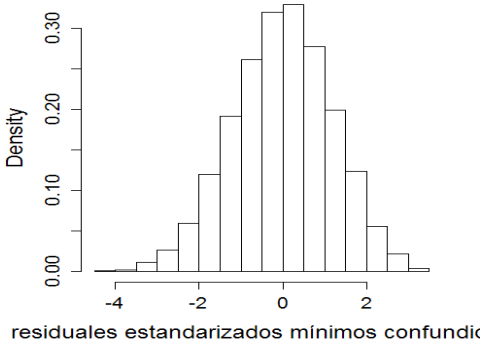


Figura 11. QQplot normal para los residuales estandarizados mínimos confundidos e histograma

3. Conclusiones

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

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

4. Literatura citada

Henderson, C.R.; Searle, S.R.; Kempthorne, O.; vonKrosigk, C.M. 1959. Estimation of Enviromental and Genetic Trends from Records Subject to Culling. Biometrics, 15: 192-218.

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

Singer, J.M.; Nobre, J.S.; Rocha, F. 2013. Diagnostic and treatment for linear mixed models. Session CPS203 Proceedings of the ISI World Statistics Congress (59), Hong Kong). Hong Kong, República Popular China.

Vázquez, A.I.; Bates, D.; Rosa, G.J.; Gianola, D.; Weigel, K.A. 2010. Technical note: An R package for fitting generalized linear mixed models in animal breeding. Journal of Animal Science 88 (2): 497-504. Disponible en https://www.alsciencepublications.org/publications/jas/abstracts/88/2/497

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