UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2017, Vol. 10, No. 2, pp. 247–280

El impacto de especificar incorrectamente la distribución de los efectos aleatorios en las estimaciones de modelos lineales generalizados mixtos[1]

The impact of misspecifying random efects distribution on the estimation of generalized linear mixed models

Diana María Arango Botero[a]
dmarangob@unal.edu.co

Freddy Hernádez Barajas[b]
fhernanb@unal.edu.co

Resumen

La inferencia en modelos lineales generalizados mixtos está basada principalmente en la teoría de máxima verosimilitud, la cual asume que las estructuras tanto para la parte de los efectos fijos como de los efectos aleatorios están correctamente especificadas. Algunos autores han mostrado la sensibilidad de las estimaciones de los efectos fijos a especificaciones incorrectas de los efectos aleatorios. El objetivo de esta investigación es identificar, vía simulación, el impacto de la especificación incorrecta de la distribución de los efectos aleatorios en los modelos lineales generalizados mixtos con variable respuesta distribuida Poisson y binomial negativa.

***Palabras clave*:** Distribución binomial negativa, distribución Poisson, efectos aleatorios, especificación incorrecta, modelos lineales generalizados mixtos.

Abstract

Inference in generalized linear mixed models is often based on maximum likelihood theory, which assumes that structures of both fixed effects and random effects is correctly specified. Some authors have shown sensitivity of estimates of fixed effects to random-effects misspecifications. This research aims to identify, using simulation, the impact of misspecifying random-effects distribution in generalized linear mixed models, specifically for the cases of Poisson and Negative Binomial

[1]DOI: http://dx.doi.org/10.15332/s2027-3355.2017.0002.04
Arango, D.,Hernández, F. (2017) El impacto de especificar incorrectamente la distribución de los efectos aleatorios en las estimaciones de modelos lineales generalizados mixtos. *Comunicaciones en Estadística*, **10**(2), 247-280.

[a]Magister, Escuela de estadística, Universidad Nacional de Colombia, sede Medellín
[b]Profesor asociado, Escuela de estadística, Universidad Nacional de Colombia, sede Medellín

distributions.

***Keywords*:** Generalized linear mixed models, misspecification, negative binomial distribution, Poisson distribution random effects.

1. Introducción

La estimación en los modelos lineales generalizados mixtos (GLMM, por sus siglas en inglés) se basa principalmente en la teoría de máxima verosimilitud. Neuhaus & McCulloch (2011a) consideran dos enfoques populares para la estimación de los parámetros fijos vía máxima verosimilitud (condicional y marginal). La estimación y las inferencias basadas en estos enfoques dependen de la suposición de que la estructura de los efectos aleatorios está correctamente especificada (Alonso et al. 2008, Molenberghs & Verbeke 2005).

En la clase general de modelos de efectos mixtos (incluyendo modelos de efectos mixtos generalizados), se hace un supuesto específico sobre la distribución paramétrica para los efectos aleatorios (por ejemplo, gamma, normal), ya sea por razones convencionales (Tsonaka et al. 2010) o por consideraciones computacionales (Litière et al. 2007). Sin embargo, a menudo hay poca información acerca de la forma de la distribución conjunta de los efectos aleatorios, de modo que el supuesto de la distribución para estos efectos aleatorios no observados no se puede evaluar directamente (Xiang et al. 2012); por esta razón, una preocupación natural en el uso de GLMM es la especificación incorrecta del modelo para los efectos aleatorios (Huang 2009).

En un modelo de regresión pueden aparecer diferentes tipos de especificaciones incorrectas, algunos de los cuales son más difíciles de detectar que otros (Spiessens et al. 2002). Aunque la mala especificación de la distribución de los efectos aleatorios puede afectar gravemente la estimación y los procedimientos inferenciales en GLMM, otros tipos de especificación incorrecta de la estructura de los efectos aleatorios también son importantes (Alonso et al. 2008). Por ejemplo, McCulloch & Neuhaus (2011) identifican otros dos aspectos de la especificación incorrecta: la distribución de los efectos aleatorios puede depender de una covariable y la dependencia de la distribución de los efectos aleatorios sobre el tamaño de muestra del conglomerado.

En los modelos lineales generalizados mixtos, la distribución de los efectos aleatorios usualmente se asume normal (Alonso et al. 2010) y esta distribución es frecuentemente usada en los paquetes de *software* populares como `SAS`, `Stata` y `R` (McCulloch & Neuhaus 2011). Aunque lahipótesis de normalidad para los efectos aleatorios a menudo se da por sentada, es importante darse cuenta que, dado que los efectos aleatorios son cantidades hipotéticas latentes no observables, este supuesto no se puede evaluar directamente, y no parece haber un consenso general sobre el impacto de especificar incorrectamente la distribución de los efectos aleatorios (Verbeke & Molenberghs 2013).

Para los modelos lineales mixtos, Verbeke & Lesaffre (1997) mostraron que los estimadores de máxima verosimilitud (ML, por sus siglas en inglés) de los efectos fijos y los componentes de varianza, obtenidos bajo el supuesto de efectos aleatorios normales, son consistentes y asintóticamente normales, incluso cuando está mal especificada la distribución de efectos aleatorios. Sin embargo, la investigación llevada a cabo en los últimos años ilustra que resultados similares no son válidos para GLMM (Litière et al. 2008). Por ejemplo, Komàrek & Lesaffre (2008) indicaron que, en contraste con el modelo lineal mixto, la mala especificación de la distribución de los efectos aleatorios en GLMM podría influir en la inferencia de los efectos fijos, que son generalmente de interés primario, pero la situación no es clara. Litière et al. (2008) trataron de ilustrar que la especificación incorrecta de la distribución de los efectos aleatorios en GLMM puede tener un efecto sobre los estimadores ML y los procedimientos de inferencia. Sus simulaciones indican que diferentes aspectos del modelo se ven afectados de diferentes maneras y en diferentes grados; es importante destacar que esta conclusión parece ser independiente de la estrategia adoptada para estudiar la especificación errónea. El impacto parece depender de la complejidad de la estructura de los efectos aleatorios, la varianza de la distribución de los efectos aleatorios subyacente y los parámetros de interés.

Litière et al. (2007) exponen que hay una amplia variedad de opiniones sobre el impacto de la especificación incorrecta de los efectos aleatorios en GLMM. Según Huang (2009), investigaciones anteriores para abordar esta preocupación han sugerido que la especificación incorrecta de los modelos para los efectos aleatorios, por lo general, solo da lugar a una pequeña cantidad de sesgo en los estimadores de máxima verosimilitud (MLE, por sus siglas en inglés) para los efectos fijos. Sin embargo, varios autores han afirmado sensibilidad a la especificación paramétrica de una distribución de efectos aleatorios (McCulloch & Neuhaus 2011). Por ejemplo, Agresti et al. (2004) llevaron a cabo estudios empíricos sobre el impacto de la especificación incorrecta del modelo para los efectos aleatorios en GLMM, mostrando que los MLE por los efectos fijos pueden ser muy sensibles al modelo asumido para los efectos aleatorios.

Alonso et al. (2015) exponen que, en general, si la selección del modelo está mal especificada entonces las estimaciones de los parámetros en el modelo pueden estar sesgadas y los procedimientos de inferencia, al igual que los intervalos de confianza, se pueden afectar también. Por lo tanto, un análisis de sensibilidad para evaluar la estabilidad de los resultados es siempre altamente recomendada (Geneletti et al., 2011; citado por Alonso et al. (2015)). Un artículo muy citado es el de Heckman y Singer (1984; citado por McCulloch & Neuhaus (2011)), el cual hace referencia a que las estimaciones de los parámetros estructurales obtenidos de los procedimientos convencionales son muy sensibles a la elección de la mezcla de distribuciones.

Según Litière et al. (2007) para estudiar el impacto de la especificación incorrecta de la distribución de los efectos aleatorios en inferencias, los investigadores suelen utilizar diseños de simulación en los que se consideran varias opciones para la verdadera distribución subyacente de los efectos aleatorios, mientras que la distribución asumida se mantiene fija. Ellos ilustran que la potencia puede ser

seriamente alterada, dependiendo de la forma y la varianza de la distribución subyacente de los efectos aleatorios. Aunque, Neuhaus et al. (2011) expusieron que el trabajo de Litière et al. (2007) contiene una falacia lógica que invalida esta afirmación, pues para demostrar los efectos de la especificación incorrecta, se necesita variar la distribución ajustada asumida, mientras se mantiene constante la verdadera distribución. Ellos presentan estudios de simulación lógicamente correctos que demuestran poco aumento en el error de tipo II, en consonancia con el trabajo anterior que muestra poco sesgo en las estimaciones de los efectos de covarianza debido a la especificación incorrecta. Además, la evidencia más fuerte para apoyar las conclusiones de (Litière et al. 2007) proviene de simulaciones que fueron incapaces de replicar, a pesar de programación muy cuidadosa.

Se sabe que los estimadores de máxima verosimilitud y los procedimientos inferenciales asociados pueden ser afectados por especificaciones incorrectas de la estructura de efectos aleatorios en GLMM (Alonso et al. 2008); por esa razón, muchos autores se han preocupado por pruebas para detectar la especificación incorrecta. Huang (2009) propuso un método de diagnóstico de dos etapas para detectar la especificación incorrecta del modelo de los efectos aleatorios en GLMM, este método utiliza los datos observados y unos reconstruidos creados a partir de los datos observados. Alonso et al. (2010) propusieron dos pruebas de diagnóstico que se basan en dos representaciones equivalentes de la matriz de información del modelo. Ellos evaluaron el poder de ambas pruebas usando consideraciones diagnóstico, así como vía simulación. Waagepetersen (2006; citado por Alonso et al. (2010)) propuso una prueba basada en la simulación para evaluar la idoneidad de la elección de la distribución de los efectos aleatorios, mediante la generación de efectos aleatorios mientras condiciona sobre las observaciones. Tchetgen y Coull (2006; citado por Alonso et al. (2010)) introdujeron una prueba de diagnóstico para evaluar la distribución asumida de los efectos aleatorios, mediante la comparación de estimadores ML marginales y condicionales de un subconjunto de efectos fijos en el modelo. Muchos autores han considerado probar la especificación incorrecta en los modelos mixtos, por ejemplo, mediante la comparación de inferencias robustas y basadas en el modelo (Alonso et al. 2008), mediante la comparación de las estimaciones de máxima verosimilitud marginales y condicionales (Tchetgen y Coull , 2006; citado por Verbeke & Molenberghs (2013)), mediante la comparación de inferencias basadas en el original y en los resultados obtenidos (Huang 2009), o mediante la comparación de las distribuciones de los residuales y/o efectos aleatorios predichos con sus distribuciones esperadas bajo el modelo asumido (Ritz, 2004; Pan & Lin, 2005; citados por Verbeke & Molenberghs (2013)).

Komàrek & Lesaffre (2008) trataron de mostrar cómo la "mezcla gaussiana penalizada"(PGM, por sus siglas en inglés) se puede utilizar como una herramienta de diagnóstico para comprobar supuestos paramétricos sobre la distribución de los efectos aleatorios. El enfoque se basa en la idea de suavizamiento penalizado, promovido por Eilers y Marx (1996; citado por Komàrek & Lesaffre (2008)). Verbeke & Molenberghs (2013) desarrollaron una herramienta de diagnóstico exploratoria sencilla para comprobar gráficamente la idoneidad de un supuesto paramétrico específico (a menudo la normalidad) acerca de la distribución de los efectos alea-

torios en diversos tipos de modelos mixtos. Su técnica no requiere ningún cálculo, además de los cálculos necesarios para ajustar el modelo, y en caso de cualquier evidencia de especificación errónea, su método indica cómo el modelo paramétrico puede ser mejorado para describir mejor los datos observados.

Es importante señalar que se han sugerido algunos enfoques diferentes para tratar con la especificación incorrecta de la distribución de los efectos aleatorios. Un área de trabajo con un enfoque ligeramente diferente ha sido el de la estimación de la forma de la distribución de los efectos aleatorios, además de establecerhipótesis de ajustes más flexibles de la distribución para los efectos aleatorios (McCulloch & Neuhaus 2011). Chen et al. (2002; citado por Litière et al. (2008)) sugirieron una distribución de los efectos aleatorios semi-paramétrica, permitiendo que la densidad de los efectos aleatorios sea sesgada, multimodal, de cola delgada o pesada, e incluyendo la normal como un caso especial. Lee y Thompson (2007; citado por Litière et al. (2008)) utilizaron métodos MCMC (Monte Carlo Markov Chain) para ajustar modelos con efectos aleatorios siguiendo una distribución t-student, y extensiones de la normal y la distribución t. Otro enfoque consiste en la sustitución de la distribución normal de los efectos aleatorios mediante mezclas de distribuciones normales (Magder & Zeger, 1996; Caffo, An & Rohde, 2007; citados por McCulloch & Neuhaus (2011)) y ajustes suaves no paramétricos (Laird, 1978; Davidian & Galán, 1993; Zhang & Davidian, 2001; Ghidey, Lesaffre & Filers, 2004; citados por McCulloch & Neuhaus (2011)). Litière et al. (2008) y Verbeke & Molenberghs (2013) también utilizaron un enfoque con mezclas de distribuciones normales para ajustar las distribuciones de los efectos aleatorios.

En la literatura estadística, algunos autores han abordado los efectos de la especificación incorrecta de la distribución de los efectos aleatorios en los modelos lineales generalizados mixtos con respuesta normal y binaria (Neuhaus et al. 1992, Heagerty & Kurland 2001, Neuhaus & McCulloch 2006, Litière et al. 2007, Komàrek & Lesaffre 2008, Huang 2009, Neuhaus & McCulloch 2011b), pero han sido pocos los trabajos en los que se han analizado modelos lineales generalizados mixtos con respuesta Poisson (Fabio et al. 2012, Milanzi et al. 2012, Cook et al. 2007) y con respuesta binomial negativa (Kondo et al. 2015, Zhao et al. 2014). Por lo anterior, es que el objetivo de este artículo identificar el impacto de la especificación incorrecta de la distribución de los efectos aleatorios en las inferencias sobre los parámetros fijos y los componentes de varianza en los modelos lineales generalizados mixtos. Para llevar a cabo dicho objetivo, se procedió con un estudio de simulación que incluyen dos tipos de variable respuesta: Poisson y binomial negativa, donde se variá la verdadera distribución de los efectos aleatorios, mientras se mantuvo constante la distribución asumida. En la primera parte se presenta una breve descripción de los modelos lineales generalizados mixtos. Luego, se describe cómo es el proceso de inferencia en dichos modelos y los paquetes computacionales que son utilizados para su ajuste. Después se presenta el estudio de simulación y finalmente los resultados y conclusiones.

2. Modelos lineales generalizados mixtos

Los modelos lineales generalizados mixtos (GLMM, por sus siglas en inglés) amplían la clase de modelos lineales generalizados por la adición de efectos aleatorios al predictor lineal (Neuhaus & McCulloch, 2011c), los cuales permiten modelar las observaciones correlacionadas. De igual forma que los modelos lineales generalizados, los GLMM pueden ser formulados usando una especificación de tres partes (Fitzmaurice et al. 2011), donde Y_{ij} representa la observación del j-ésimo individuo $(j = 1, 2, ..., n_i)$ dentro del i-ésimo conglomerado $(i = 1, 2, ..., m)$:

1. La distribución condicional de cada Y_{ij}, dado un vector $q \times 1$ de efectos aleatorios $\mathbf{b}_i$, pertenece a la familia exponencial de distribuciones (binomial, binomial negativa, Poisson, normal, gamma, entre otras). La $Var(Y_{ij}|\mathbf{b}_i) = \phi\upsilon(E(Y_{ij}|\mathbf{b}_i))$, donde $\upsilon(\cdot)$ es una función conocida para la varianza, una función de la media condicional, $E(Y_{ij}|\mathbf{b}_i)$ y ϕ es un parámetro escalar que puede ser conocido o ser necesario estimarlo. En adición, dado los efectos aleatorios $\mathbf{b}_i$, se asume que los Y_{ij} son independientes entre sí, lo cual es la asunción de independencia condicional.

2. La media condicional de Y_{ij}, que depende de los efectos fijos $\boldsymbol{\beta}$ y los efectos aleatorios $\mathbf{b}_i$, se relaciona con el predictor lineal η_{ij}, vía la aplicación de una función de enlace conocida, $g(\cdot)$, la cual es monótona y diferenciable (Gad & El Kholy 2012), de la siguiente manera:

$$g\{E(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij})\} = \eta_{ij} = \mathbf{x}_{ij}^\top\boldsymbol{\beta} + \mathbf{z}_{ij}^\top\mathbf{b}_i \tag{1}$$

   donde $\mathbf{x}_{ij}$ y $\mathbf{z}_{ij}$ son dos vectores de covariables $p + 1$ dimensionales y q dimensionales, respectivamente.

   Aunque cualquier función puede ser escogida para $g(\cdot)$, cada distribución que pertenece a la familia exponencial tiene una función de enlace especial llamada función de enlace canónica. La función de enlace canónica es definida como esa función $g(\cdot)$ tal que

$$g\{E(y_{ij})\} = \theta_i, \tag{2}$$

   donde θ_i es el parámetro de localización canónico.

3. En principio, cualquier distribución multivariada puede ser asumida para los $\mathbf{b}_i$. En la práctica, es común asumir que los $\mathbf{b}_i$ tienen una distribución normal multivariada, con media cero y matriz de covarianza $\mathbf{D}$ de dimensiones $q \times q$. Adicionalmente los efectos aleatorios $\mathbf{b}_i$ son asumidos para ser independientes de las covariables $\mathbf{X}_i$.

2.1. Distribución Poisson y binomial negativa

Dentro de la familia exponencial, como ya se mencionó anteriormente, se encuentran las distribuciones Poisson y binomial negativa (BN). En este apartado se profundizará sobre cada una de ellas, debido a que son las distribuciones consideradas en el estudio de simulación.

La distribución Poisson sobre la que la regresión Poisson está basada, se origina desde el trabajo de Simeon Poisson (1781-1840; citado por Hilbe (2011)); él introdujo la distribución como un caso límite de la binomial en su "Research on the Probability of Judgments in Criminal and Civil Matters" (1838).

Una variable aleatoria Y tiene distribución Poisson con media $\lambda > 0$ si la distribución de probabilidad de masa es como sigue (DeGroot & Schervish 1988):

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \text{ para } y = 0, 1, 2, ... \tag{3}$$

donde $E(Y) = \lambda$ y $Var(Y) = \lambda$.

La binomial negativa tradicional se deriva de una mezcla de distribución Poisson-gamma, pero tal mezcla de distribuciones es solo una de las maneras en la que la función de probabilidad de masa de la BN puede ser definida. La característica interesante de esta parametrización es que permite modelar la heterogeneidad de la Poisson (Hilbe 2011). Como se detalló anteriormente, la media y la varianza de la Poisson son iguales, cuanto mayor es el valor de la media, mayor es la variabilidad de los datos, medido por la varianza estadística. Esta característica de los datos se denomina equidispersión y es un supuesto de la distribución de los datos de Poisson. Inherente a esta suposición está el requisito de que los conteos sean independientes unos de otro. Cuando no es así, las propiedades de la distribución Poisson son violados, lo que resulta en extra-dispersión. La media y la varianza ya no pueden ser idénticas. La forma de extra-dispersión es casi siempre una de sobredispersión, es decir, la varianza es mayor, en valor, que la de la media. El modelo BN, como un modelo de mezcla Poisson-gamma, es apropiado de utilizar cuando la sobredispersión en un modelo de Poisson está presente (Hilbe 2011). Es así, como la distribución BN depende de un parámetro extra comparado con la distribución Poisson, el cual permite que la sobredispersión sea tenida en cuenta. Este parámetro es denotado con la letra $\alpha(\alpha > 0)$ y entre más grande sea su valor, mayor será la sobredispersión.

La función de probabilidad de masa de la BN está dada por:

$$f(y|\mu, \alpha) = \binom{y + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \text{ para } x = 0, 1, 2, ... \tag{4}$$

donde $E(Y) = \mu$ y $Var(Y) = \mu + \alpha\mu^2$, con $\mu > 0$ y $\alpha > 0$.

3. Inferencias en los GLMM

Los efectos sujeto-específicos $\mathbf{b}_i$ son asumidos con frecuencia que distribuyen normal con media cero y matriz de varianza-covarianza $\mathbf{D}$. El ajuste del modelo requiere maximización de la verosimilitud marginal, la cual es obtenida integrando sobre los efectos aleatorios.

Sea $f(\mathbf{Y_i}|\mathbf{b_i}, \boldsymbol{\eta_i}, \phi) = \prod_{j=1}^{\eta_i} f(y_{ij}|\mathbf{b_i}, \boldsymbol{\eta_i}, \phi)$ la distribución condicional de un vector $\mathbf{Y_i}$ de observaciones del conglomerado i dado $\mathbf{b_i}$, donde y_{ij} corresponde a la función de densidad dada en la expresión (3) si y_{ij} es Poisson o la expresión (4) si se está en el caso de binomial negativa. Además, $g(\mathbf{b_i}|\mathbf{D})$ es la distribución marginal de los efectos aleatorios $\mathbf{b_i}$. Entonces la distribución marginal de los $\mathbf{Y_i}$'s es obtenida integrando la densidad conjunta de $\mathbf{Y_i}$ y $\mathbf{b_i}$ con respecto a $\mathbf{b_i}$,

$$f(\mathbf{Y_i}|\boldsymbol{\theta}) = \int \int \cdots \int f(\mathbf{Y_i}|\mathbf{b_i}, \boldsymbol{\eta_i}, \phi) g(\mathbf{b_i}|\gamma) d\mathbf{b_i} \tag{5}$$

El logaritmo de esta distribución marginal es la contribución de log-verosimilitud, $l_i(\boldsymbol{\theta})$, para el individuo i. Las estimaciones de máxima verosimilitud se calculan para maximizar la log-verosimilitud, $l = \sum_{i=1}^{n} l_i(\boldsymbol{\theta})$. La dificultad para calcular las estimaciones de máxima verosimilitud radica en la integración de la distribución conjunta para obtener la contribución de log-verosimilitud. Las soluciones analíticas de la integral multidimensional no son posibles en general, y los métodos de cuadratura del producto se vuelven inviables desde el punto de vista computacional a medida que aumenta la dimensión de la integral.

La elección de la distribución normal para estos efectos aleatorios generalmente conduce a funciones de verosimilitud intratables, con la excepción del modelo lineal mixto (LMM, por sus siglas en inglés), donde la variable de respuesta tiene una distribución normal (Alonso et al. 2008). En respuesta, varias aproximaciones numéricas a la verosimilitud se han implementado en los paquetes de *software* disponibles (tabla 1)

En el caso más simple de efectos aleatorios anidados (GLMM), hay varios métodos disponibles para obtener estimaciones de máxima verosimilitud (ML), incluida la linealización (MQL, PQL) e integración numérica, como la aproximación de Laplace (MLLA) y la cuadratura gaussiana adaptativa (AGQ) (Grilli & Innocenti 2016).

Los métodos bayesianos generalmente tienen un mejor rendimiento en modelos de efectos aleatorios complejos (Karim & Zeger 1992). Sin embargo, el método bayesiano estándar, llamado MCMC, tiene algunas limitaciones prácticas debido a la carga computacional y las dificultades para evaluar la convergencia. Una posible solución está representada por INLA, es decir, aproximaciones de Laplace anidadas integradas. De hecho, INLA directamente aproxima la distribución posterior, evitando así los métodos complejos basados en la simulación (Grilli & Innocenti 2016).

Tabla 1: *Métodos de estimación en los GLMM. Fuente: Adaptado de Bolker, Brooks, Clark, Geange, Poulsen, Stevens and White (2009).*

Métodos para la estimación de los parámetros en GLMM	Ventajas	Desventajas	Paquetes computacionales
Cuasi-verosimilitud penalizada (PQL)	Flexible, implementada ampliamente	inferencia de la verosimilitud puede ser inapropiada; sesgo para varianzas grandes o medias pequeñas	`PROC GLIMMIX` (`SAS`), `GLMM` (`GenStat`), `glmmPQL` (`R:MASS`), `ASREML-R`
Aproximación de Laplace	Mejor aproximación que PQL	más lenta y menos flexible que PQL	`glmer` (`R:lme4,lme4a`), `glmm.admb` (`R:glmmADMB`), `AD Model Builder`, `HLM`
Cuadratura Gauss-Hermite	Mejor aproximación que Laplace	más lenta que Laplace; limitada a 2-3 efectos aleatorios	`PROC NLMIXED` (`SAS`), `glmer` (`R:lme4, lme4a`), `glmmML` (`R:glmmML`), `xtlogit` (`Stata`)
Cadenas de Markov de Monte Carlo	Altamente flexibles, número arbitrario de efectos aleatorios	Muy lento, técnicamente desafiante, marco de referencia bayesiano	`MCMCglmm` (`R:MCMCglmm`), `MCMCpack` (`R`), `WinBUGS/OpenBUGS` (`R:BRugs/R2WinBUGS`), `JAGS` (`R:rjags/R2jags`), `AD Model Builder` (`R:R2admb`), `glmm.admb1` (`R:glmmADMB`)

De acuerdo a Raudenbush et al. (2000), la aproximación multivariada de Laplace es una buena estrategia para evaluar las verosimilitudes en modelos lineales generalizados con efectos aleatorios anidados.

Los ajustes de los modelos utilizados en este estudio se hicieron con el paquete `glmmADMB` del *software R*, el cual es un paquete construido sobre código abierto AD Model Builder (ADMB), que es el *software* más rápido y más poderoso para el desarrollo y ajuste de modelos generales estadísticos no lineales (Fournier 2011). El ADMB utiliza la aproximación de Laplace (Skaug & Fournier 2006) para el cálculo de las verosimilitudes marginales por iteración entre la maximización de la verosimilitud con respecto a los efectos aleatorios, y la actualización del valor de los hiperparámetros usando las estimaciones de los efectos aleatorios obtenidos en la maximización de la verosimilitud (Fournier et al. 2012).

4. Estudio de simulación

Para estudiar el impacto sobre las estimaciones de los parámetros cuando se especifica incorrectamente la distribución de los efectos aleatorios, se realizó un estudio de simulación, considerando en la primera parte modelos mixtos Poisson y binomial negativa con intercepto aleatorio; en la segunda, se incluyó tanto un intercepto como una pendiente aleatoria en las simulaciones para dichos modelos.

4.1. Modelos Poisson y binomial negativa con intercepto aleatorio

Para el estudio de simulación se generaron respuestas Poisson y binomial negativa a partir de los GLMM con intercepto aleatorio, donde se consideraron $m = 100$ conglomerados de cinco tamaños diferentes ($n_i = 3, 6, 9, 12, 15$).

Para el caso Poisson, se consideró el siguiente modelo con intercepto aleatorio:

$$\begin{aligned} y_{ij}|b_i &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij}), \\ log(\mu_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + b_i, \end{aligned} \tag{6}$$

con $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $i = 1, 2, ..., m$ y $j = 1, 2, ..., n_i$.

En la Tabla 2 se presentan algunos datos simulados de un modelo mixto Poisson con $m = 100$, $n_i = 6$, $\sigma^2 = 4$ e intercepto aleatorio normal, correspondiente a la ecuación 5.

Para el caso binomial negativa, el siguiente modelo mixto con intercepto aleatorio fue considerado:

Tabla 2: *Datos simulados del modelo mixto Poisson con* $m = 100$, $n_i = 6$, $\sigma^2 = 4$ *e intercepto aleatorio normal. Fuente: elaboración propia.*

X_1	X_2	Y	Conglomerado
−0.44	0.26	3	7
−0.17	0.65	3	7
−0.07	0.81	0	7
0.37	0.40	2	7
2.95	0.20	1	7
−0.73	0.26	1	7
−1.05	0.69	2	26
−1.39	0.87	3	26
−0.48	0.28	1	26
−0.65	0.75	6	26
−0.22	0.50	4	26
−0.87	0.00	2	26

$$\begin{aligned} y_{ij}|b_i &\overset{ind.}{\sim} \text{BN}(\mu_{ij}, \alpha), \\ log(\mu_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + b_{0i}, \\ \alpha &= 0.5, \end{aligned} \tag{7}$$

donde $\beta_0 = 1, \beta_1 = 2$ y $\beta_2 = 3$. Las covariables x_{1ij} y x_{2ij} representan covariables dentro de los conglomerados (covariables cuyos valores cambian para cada conglomerado i y cada observación j), con $x_1 \sim N(0, 1)$ y $x_2 \sim U(0, 1)$.

Para los modelos de las expresiones 5 y 6 los b_{0i} fueron generados a partir de cuatro distribuciones diferentes con media cero y cuatro valores de varianza σ^2 (1, 2, 4 y 16): normal, mezcla de dos normales, uniforme y lognormal (figura 1) (Alonso et al. 2008, Spiessens et al. 2002, Verbeke & Lesaffre 1997). El vector de parámetros de interés para los dos modelos de las ecuaciones 5 y 6 es $\boldsymbol{\omega} = (\beta_0, \beta_1, \beta_2, \sigma^2)^\top$.

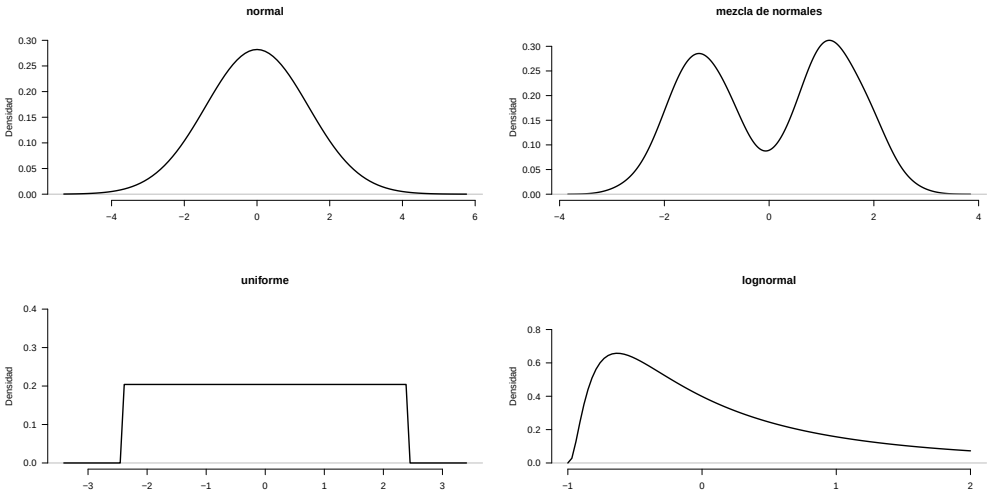

Figura 1: *Distribuciones consideradas para el intercepto aleatorio: normal, mezcla de normales, uniforme y lognormal, con media igual a 0 y varianza igual a 2. Fuente: elaboración propia.*

Para cada conjunto de datos simulados se ajustó un modelo Poisson o BN de efectos mixtos, donde se asumió un intercepto aleatorio que sigue una distribución normal. Para el ajuste del modelo y obtención del vector de parámetros $\hat{\omega}$ se utilizó la función `glmmadmb` del paquete `glmmADMB` de `R`. El impacto de la especificación incorrecta se evaluó con la distancia relativa DR entre el verdadero valor del parámetro simbolizado por ω y su valor estimado $\hat{\omega}$ (Verbeke & Lesaffre 1997). A mayor valor del indicador, mayor es el impacto de la especificación incorrecta.

$$DR = \frac{|\hat{\omega} - \omega|}{|\omega|} \tag{8}$$

4.2. Modelos Poisson y binomial negativa con intercepto y pendiente aleatoria

El impacto de la especificación incorrecta de la distribución de los efectos aleatorios también se evaluó en datos provenientes de poblaciones con respuesta Poisson y BN, con intercepto y pendiente aleatoria; para lo cual se consideraron dos covariables dentro de los conglomerado: $x_1 \sim N(0,1)$ y $x_2 \sim U(0,1)$. Se generaron $m = 100$ conglomerados con cinco tamaños diferentes ($n_i = 3, 6, 9, 12, 15$).

El modelo considerado en este caso es:

$$y_{ij}|\mathbf{b}_i \overset{ind.}{\sim} \text{Poisson}(\mu_{ij}) \quad \text{ó} \quad y_{ij}|\mathbf{b}_i \overset{ind.}{\sim} \text{BN}(\mu_{ij}, \alpha = 0.5),$$
$$log(\mu_{ij}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{1ij} + \beta_2 x_{2ij}, i = 1, 2, ..., m = 100, \quad j = 1, 2, ..., n_i \tag{9}$$

Los $\mathbf{b}_i = (b_{0i}, b_{1i})^\top$ fueron generados a partir de cuatro distribuciones diferentes con $\mu_{b_0} = \mu_{b_1} = 0$ y 4 valores de varianza $\sigma^2_{b_0} = var(b_0) = \sigma^2_{b_1} = var(b_1) = 0.5, 1, 2, 4$. Las distribuciones consideradas para $\mathbf{b_i}$ se muestran a continuación, las cuales fueron tomadas del trabajo de Neuhaus et al. (2012).

1. $\mathbf{b}_i \sim$ normal bivariada.

2. $\mathbf{b}_i \sim$ t-student bivariada con 3 grados de libertad.

3. $\mathbf{b}_i \sim$ exponencial bivariada (1).

4. $\mathbf{b}_i \sim$ Tukey bivariada ($g = 0.446$, $h = 0.05$), donde el parámetro g controla la cantidad y dirección de asimetría, mientras que el parámetro h controla la cantidad de elongación (curtosis) de la distribución Tukey bivariada (Valencia 2014).

Para medir el impacto de la especificación incorrecta de la distribución de los efectos aleatorios se fijaron los valores de $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$ y se asumió una correlación entre b_0 y b_1 de 0.5. Al igual que para el intercepto aleatorio, se utilizó la distancia relativa para evaluar el impacto de la especificación incorrecta. Los valores estimados fueron obtenidos a través de los ajustes de un modelo Poisson o BN de intercepto y pendiente aleatoria asumiendo para este caso una distribución normal bivariada. Para el ajuste se utilizó la función `glmmadmb`, del paquete `glmmADMB` de `R`.

Para la generación de las distribuciones exponencial y Tukey bivariada se partió de lo propuesto por Neuhaus et al. (2012), en donde $(b_0, b_1)^\top$ son generadas mediante la transformación de las marginales de una normal bivariada, $(Z_0, Z1)^\top$, con $E(Z_0) = E(Z_1) = 0$, $\sigma^2_{b_0} = var(b_0) = \sigma^2_{b_1} = var(b_1) = 0.5, 1, 2, 4$, $cov(Z_0, Z_1) = 0.5$.

5. Resultados

En esta sección se presentan los resultados para los modelos descritos en la sección anterior.

5.1. Resultados para el caso de modelos con intercepto aleatorio

Para el caso de GLMM con intercepto aleatorio y variable de respuesta Poisson, los resultados se presentan en las figuras 2, 3 y 4.

La figura 2 muestra las medianas de DR para las estimaciones del parámetro β_0. En el caso de la varianza de 16 se presenta un menor DR cuando la verdadera distribución es la lognormal, seguido de las distribuciones normal y mezcla de normales y, por último, se encuentra uniforme. Para las varianzas de 1, 2 y 4 no se observan diferencias en la DR.

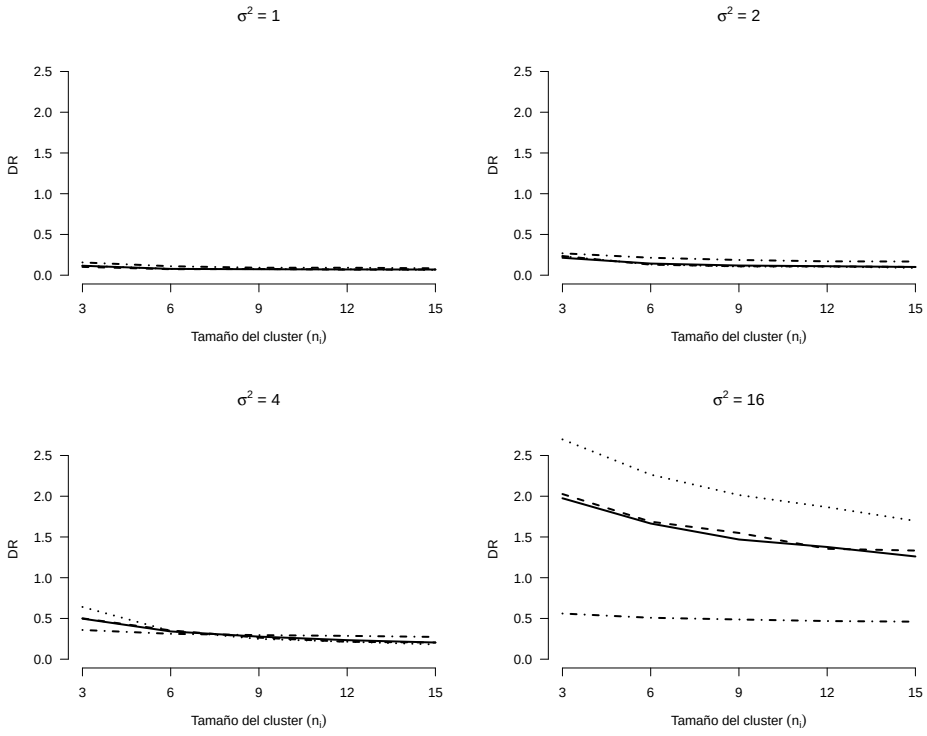

Figura 2: *Mediana de las distancias relativas para las estimaciones de β_0 en un GLMM Poisson, con $\sigma^2 = 1, 2, 4, 16$ y cuatro distribuciones para el intercepto aleatorio: —normal, · · · uniforme, - - - mezcla de normales, — · — lognormal. Fuente: elaboración propia.*

De acuerdo a la figura 3, el impacto de la especificación incorrecta de las distribuciones de los efectos aleatorios para la estimación del parámetro β_1 es indiferente para las cuatro distribuciones consideradas, puesto que la DR presenta el mismo

comportamiento con tendencia decreciente a medida que aumenta el tamaño del conglomerado n_i.

Al analizar el comportamiento de DR para las estimaciones de β_2 en una figura (no mostrada aquí), se encontró un patrón similar al observado en la figura 3, lo cual indica que no se encontró un impacto de la especificación incorrecta de la distribución del efecto aleatorio.

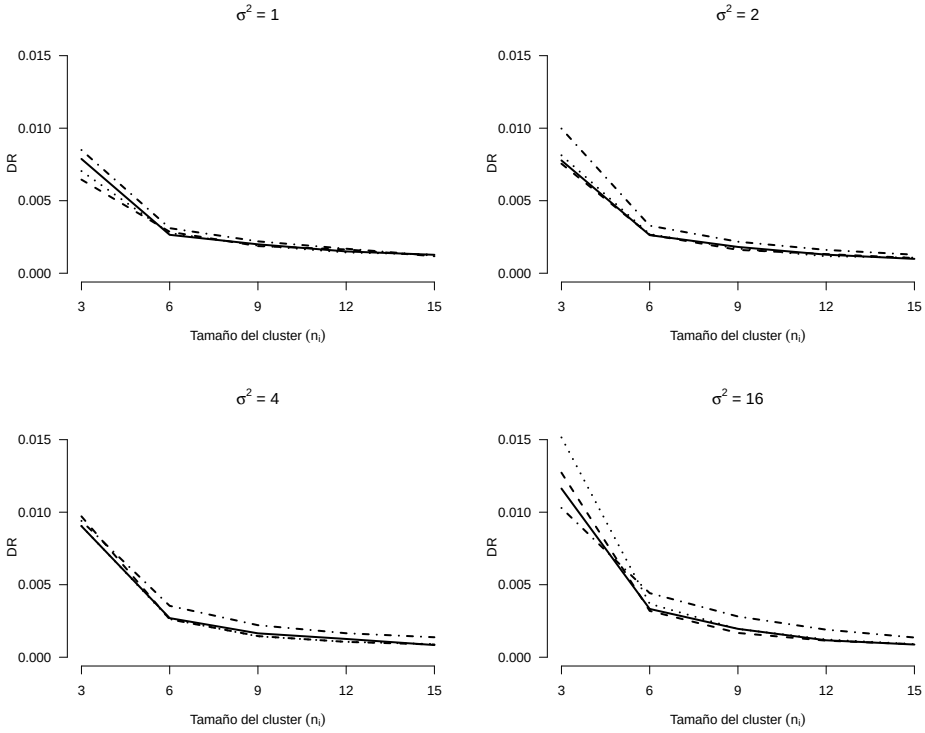

Figura 3: *Mediana de las distancias relativas para las estimaciones de β_1 en un GLMM Poisson, con $\sigma^2 = 1, 2, 4, 16$ y cuatro distribuciones para el intercepto aleatorio: —normal, · · · uniforme, - - - mezcla de normales, — · — lognormal. Fuente: elaboración propia.*

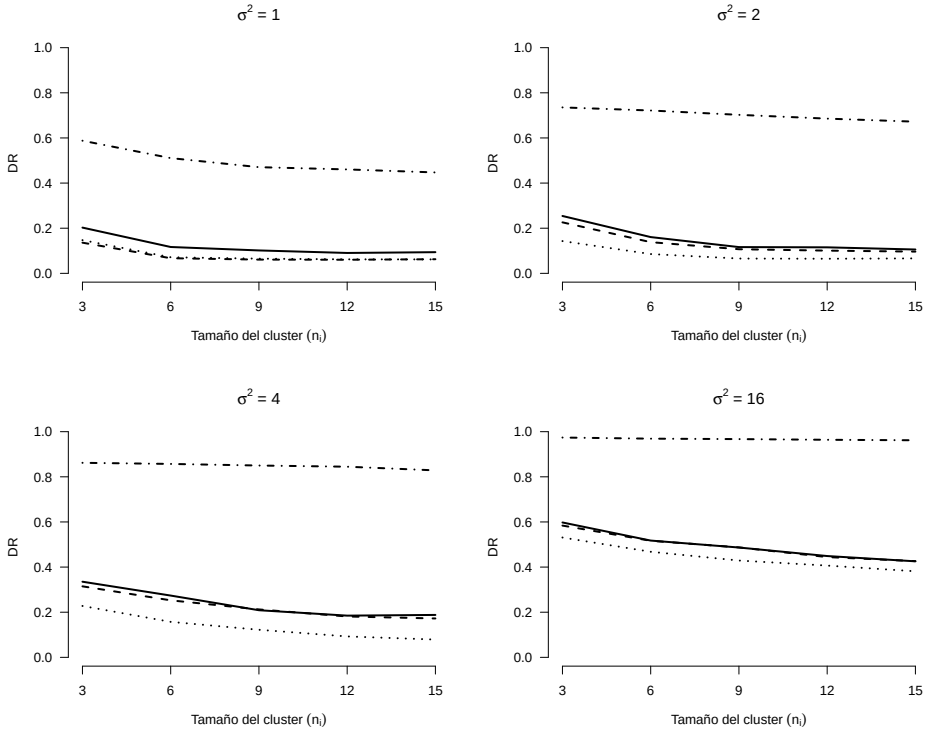

Figura 4: *Mediana de las distancias relativas para las estimaciones de* σ^2 *en un GLMM Poisson, con* $\sigma^2 = 1, 2, 4, 16$ *y cuatro distribuciones para el intercepto aleatorio: —normal,* $\cdots$ *uniforme, - - - mezcla de normales,* $-\cdot-$ *lognormal. Fuente: elaboración propia.*

El impacto de la especificación incorrecta de la distribución de los efectos aleatorios es mayor para la lognormal que para las otras distribuciones si se comparan las estimaciones para σ^2 de dicho intercepto, lo cual es mucho más evidente a medida que se aumenta el verdadero valor de σ^2, tal y como se muestra en la figura 4. Además, para cada una de las varianzas el impacto de la especificación incorrecta decrece a medida que se aumenta el tamaño de los conglomerados n_i.

En las figuras 5, 6 y 7 se muestran los resultados del estudio de simulación correspondientes al GLMM con intercepto aleatorio y respuesta BN.

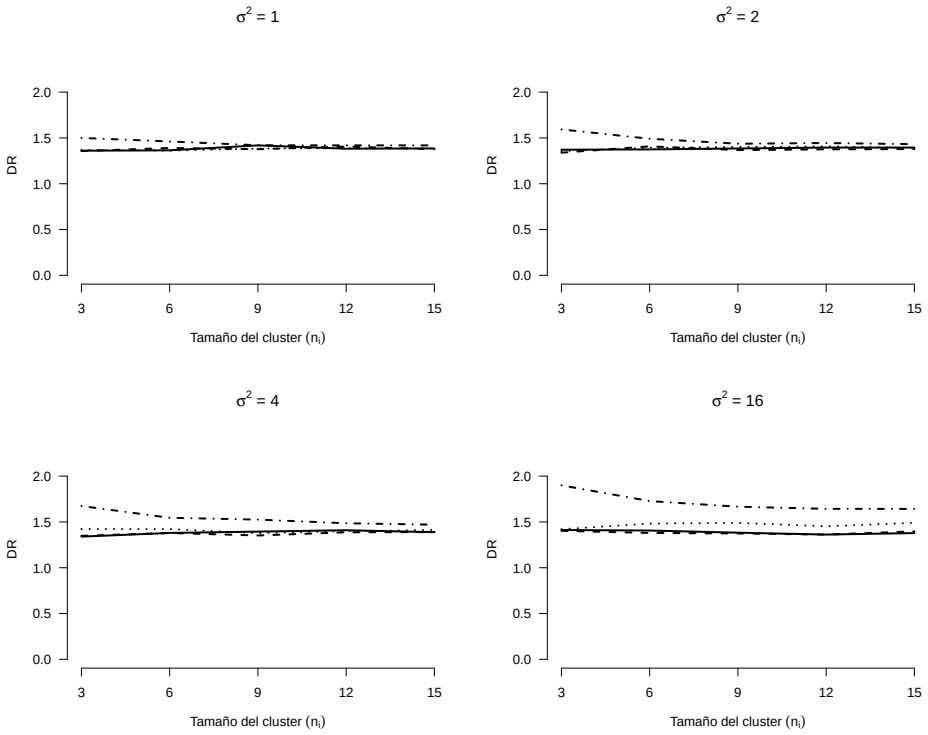

Figura 5: *Mediana de las distancias relativas para las estimaciones de* β_0 *en un GLMM BN, con* $\sigma^2 = 1, 2, 4, 16$ *y cuatro distribuciones para el intercepto aleatorio: —normal,* $\cdots$ *uniforme, - - - mezcla de normales,* $-\cdot-$ *lognormal. Fuente: elaboración propia.*

En el caso de respuesta BN se encontró que las estimaciones para el parámetro β_0 presentan una DR ligeramente mayor cuando la verdadera distribución del efecto aleatorio es la lognormal. La DR tiende a ser mayor cuando aumenta la varianza y tiende a disminuir con el aumento del tamaño del conglomerado n_i como se muestra en la Figura 5.

De la figura 6, se observa que para las estimaciones del parámetro β_1, no hay grandes diferencias entre las DR para las cuatro distribuciones de los efectos aleatorios. Se evidencia también una tendencia decreciente de las medianas de las distancias relativas a medida que aumenta el tamaño del conglomerado n_i. Las estimaciones correspondientes al parámetro β_2 presentaron resultados muy similares a las de β_1 y por esa razón no se presenta una figura en el artículo.

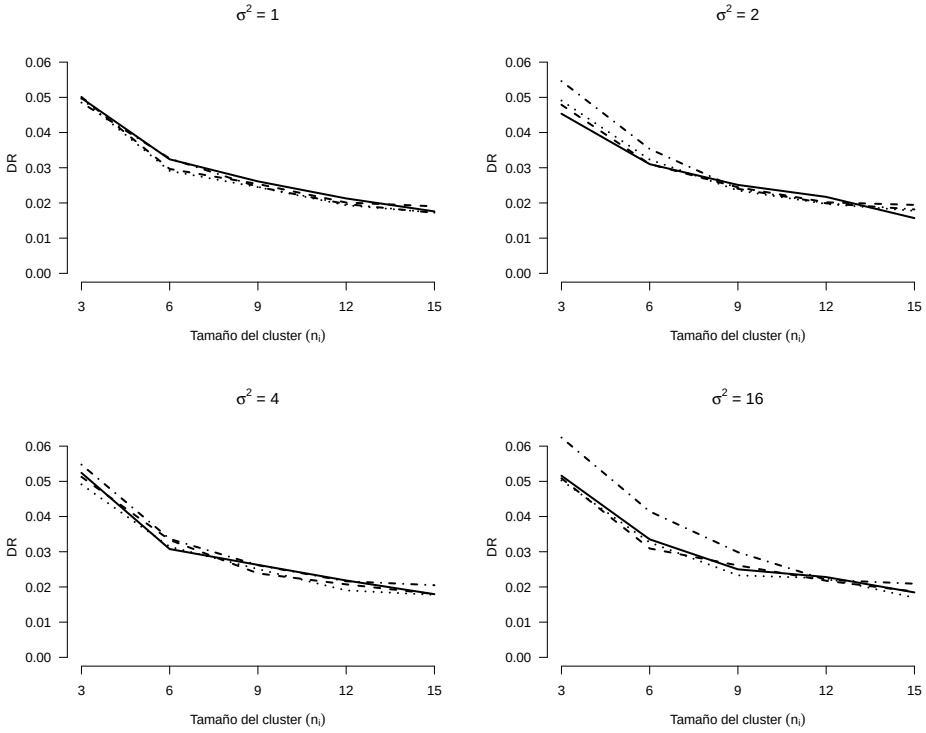

Figura 6: *Mediana de las distancias relativas para las estimaciones de* β_1 *en un GLMM BN, con* $\sigma^2 = 1, 2, 4, 16$ *y cuatro distribuciones para el intercepto aleatorio: —normal,* $\cdots$ *uniforme, - - - mezcla de normales,* $-\cdot-$ *lognormal. Fuente: elaboración propia.*

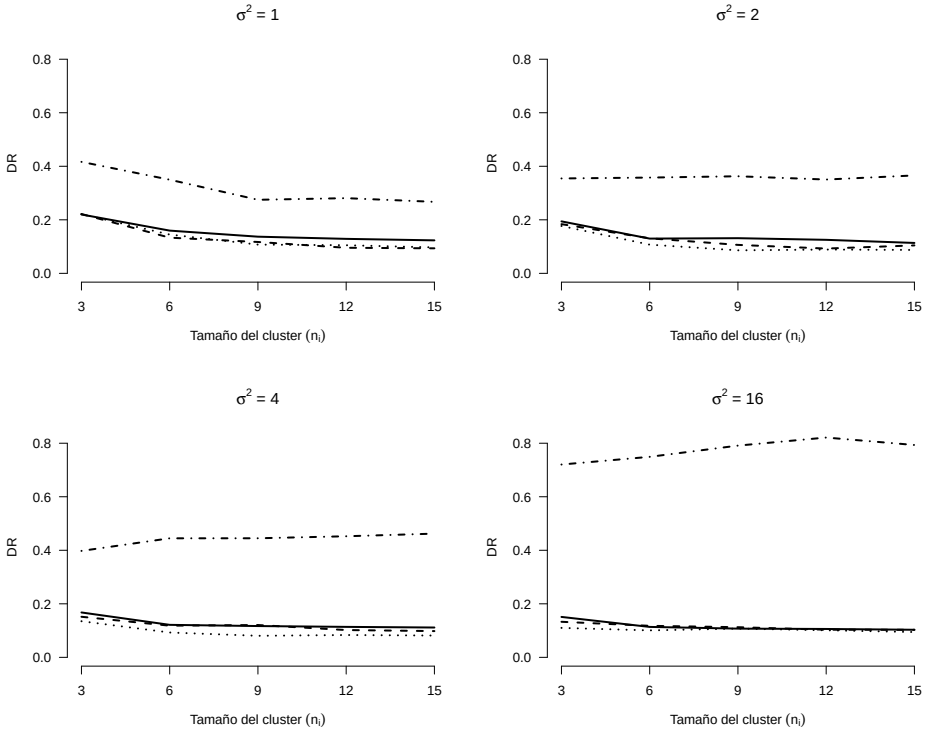

Figura 7: *Mediana de las distancias relativas para las estimaciones de* σ^2 *en un GLMM BN, con* $\sigma^2 = 1, 2, 4, 16$ *y cuatro distribuciones para el intercepto aleatorio: —normal · · · uniforme - - - mezcla de normales — · — lognormal. Fuente: elaboración propia.*

Lo observado en cuanto a las estimaciones de σ^2 del intercepto aleatorio (figura 7) es similar que para el caso de respuesta Poisson, ya que de acuerdo a los valores de las medianas de las distancias relativas, no solo se observa un impacto mayor cuando la verdadera distribución del efecto aleatorio es lognormal, sino también un aumento de dicho impacto a medida que aumenta el valor varianza.

5.2. Resultados para el caso de modelos con intercepto y pendiente aleatoria

En la figura 8 se presentan los resultados de las medianas de las distancias relativas de la estimación del parámetro β_0 para los diferentes tamaños de conglomerado ($n_i = 3, 6, 9, 12$) y las cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria considerados (normal, t-student, exponencial y Tukey). Se observa

que para las distribuciones Tukey y t-student bivariadas, se presentan los mayores valores de las medianas de las distancias relativas. Por tanto, se evidencia un impacto de la especificación incorrecta de la distribución de los efectos aleatorios para dicho parámetro poblacional.

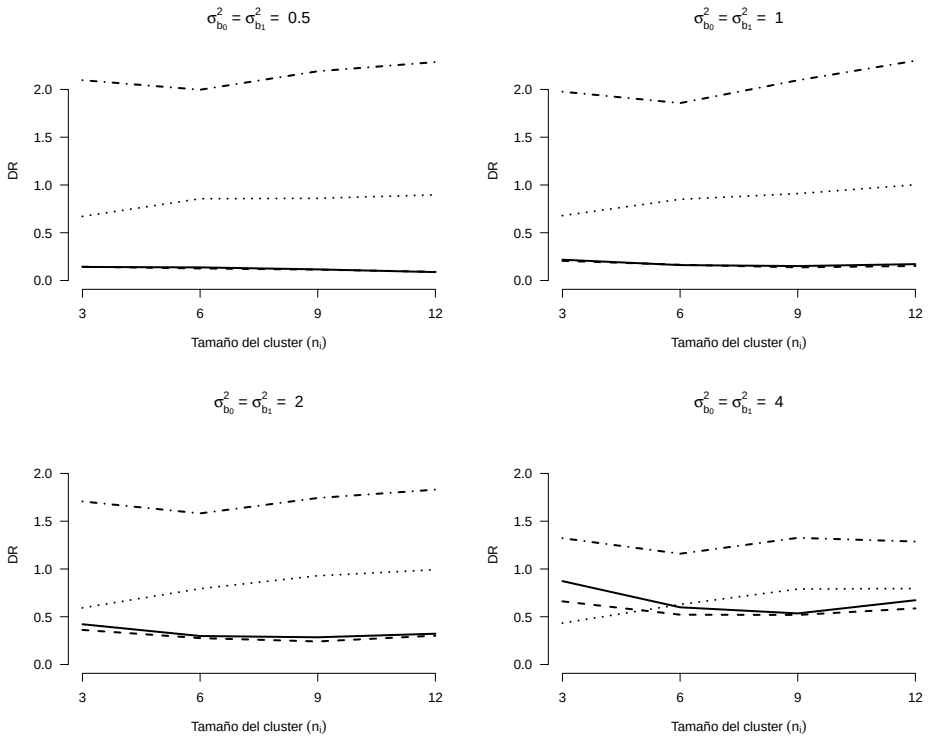

Figura 8: *Mediana de las distancias relativas para las estimaciones de β_0 en un GLMM Poisson, con $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal, · · · t-student, - - - exponencial, —·— Tukey. Fuente: elaboración propia.*

El mismo comportamiento de la figura 8 se presenta para las estimaciones del parámetro β_1 (figura 9), en donde nuevamente las distribuciones verdaderas de los efectos aleatorios Tukey y t-student bivariadas son las que presentan mayores valores de las medianas de las distancias relativas, y con ello, un mayor impacto de la especificación incorrecta en dichas distribuciones.

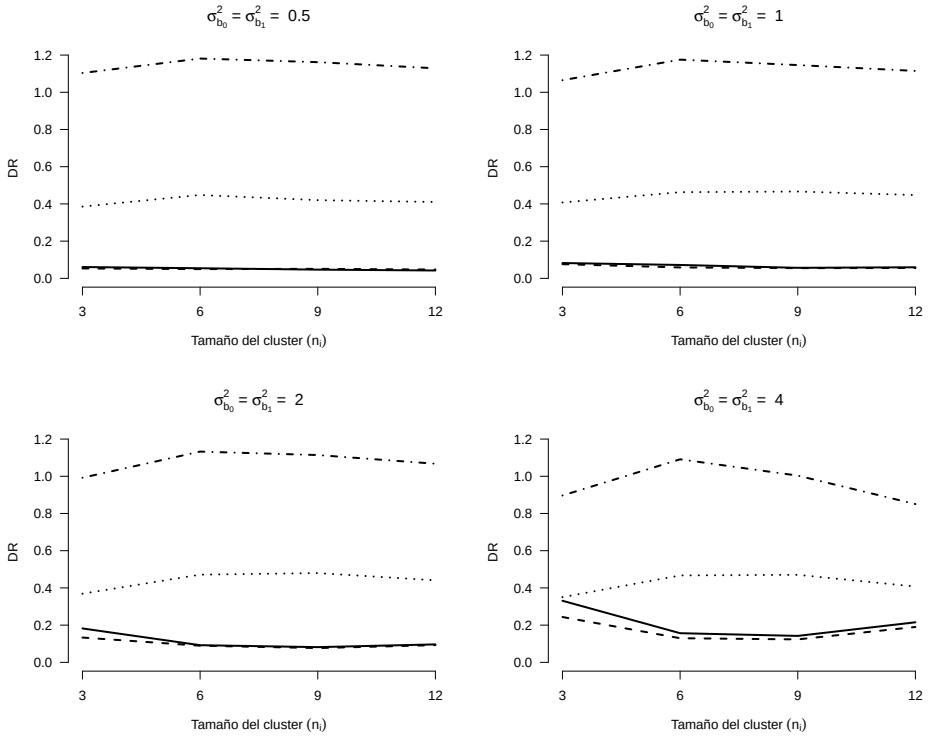

Figura 9: *Mediana de las distancias relativas para las estimaciones de* β_1 *en un GLMM Poisson, con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $-\cdot-$ *Tukey. Fuente: elaboración propia.*

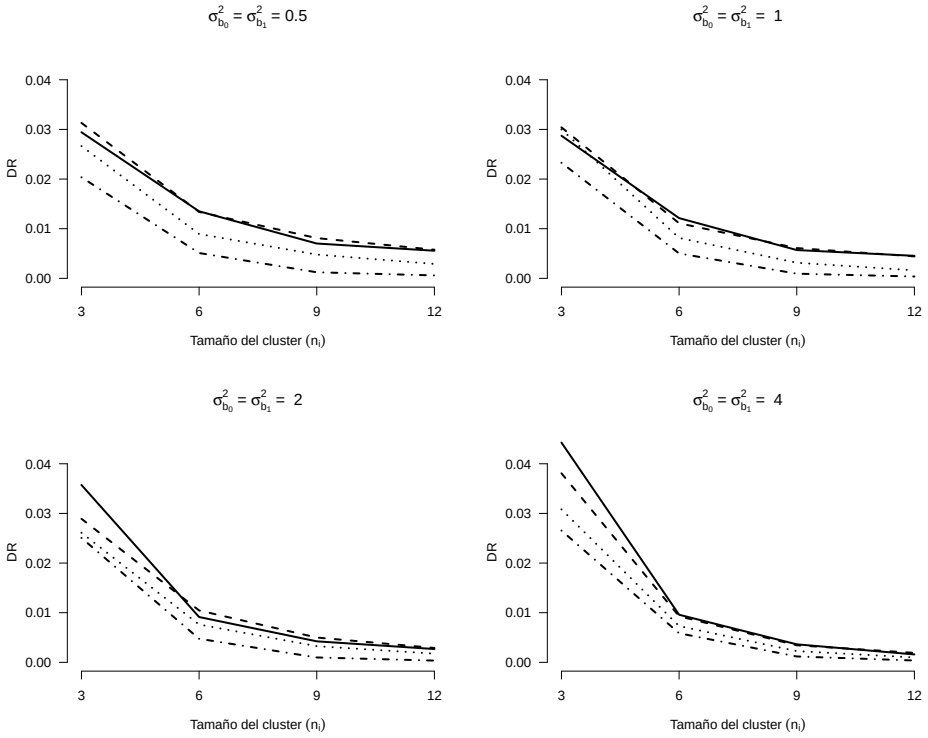

Figura 10: *Mediana de las distancias relativas para las estimaciones de β_2 en un GLMM Poisson, con $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal, · · · t-student, - - - exponencial, —·— Tukey. Fuente: elaboración propia.*

La estimación del parámetro β_2 (figura 10) resultó ser el menos afectado por la especificación incorrecta de la distribución de los efectos aleatorios con valores que oscilan entre 0 % y el 4 % para todas las configuraciones consideradas, rescatándose que los valores de las medianas de las distancias relativas son muy similares y que decrecen a medida que aumenta el tamaño del conglomerado n_i. Contrario a los resultados encontrados para el modelo mixto con intercepto aleatorio, se observan diferencias en las DR para las estimaciones de los parámetros β_1 y β_2, puesto que aquí hay una pendiente aleatoria b_1 asociada con el primero.

Las estimaciones de los componentes de varianza ($\sigma^2_{b_0}$ y $\sigma^2_{b_1}$) de ambos efectos aleatorios se vieron ampliamente afectados por la especificación incorrecta de las distribuciones de dichos efectos. En la figura 11 se observan los resultados de las estimaciones para $\sigma^2_{b_0}$, donde claramente se evidencia que la distribución Tukey bivariada es la que presenta los mayores valores de DR para todos los casos. Un resultado interesante es que cuando $\sigma^2_{b_0}$ es pequeña y la distribución verdadera

de los efectos aleatorios es Tukey, las estimaciones para $\sigma^2_{b_0}$ tienden a tener altos valores de DR; además, a medida que $\sigma^2_{b_0}$ aumenta la estimación de este parámetro mejora considerablemente.

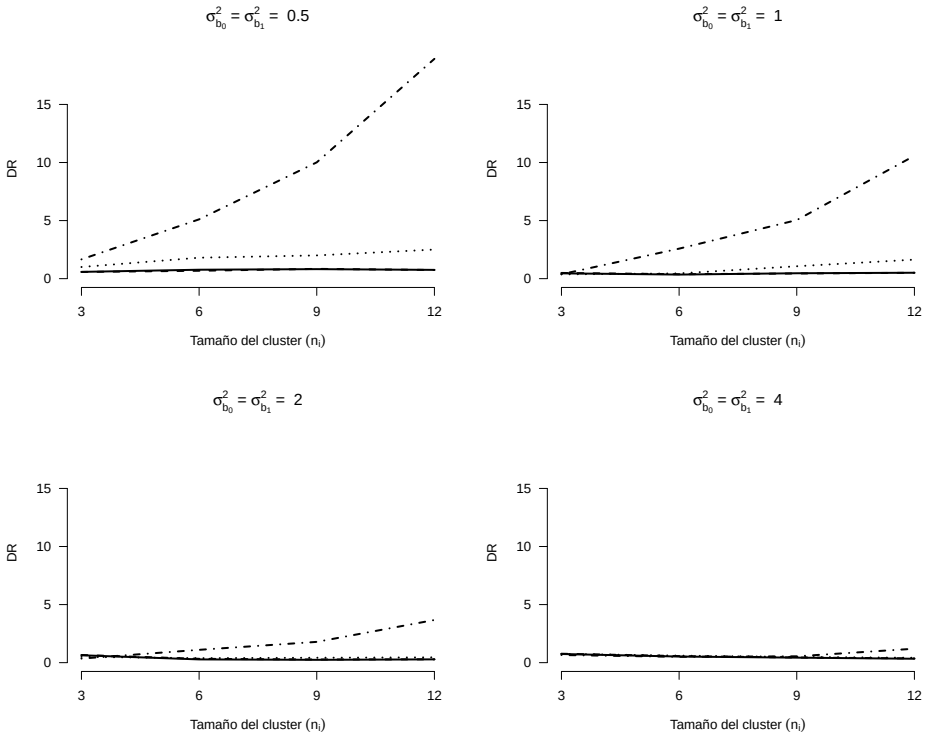

Figura 11: *Mediana de las distancias relativas para las estimaciones de* $\sigma^2_{b_0}$ *en un GLMM Poisson, con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $- \cdot -$ *Tukey. Fuente: elaboración propia.*

Para la estimación del componente de varianza $\sigma^2_{b_1}$ (figura 12) también se observa un impacto de la especificación incorrecta, pero en menor proporción que la de la estimación de $\sigma^2_{b_0}$, en donde nuevamente la distribución Tukey bivariada es la que presenta los mayores valores de las medianas de las distancias relativas cuando $\sigma^2_{b_1} = 0.5, 1$. Adicionalmente, se observa que las segundas mayores DR se dan cuando la distribución es la t-student.

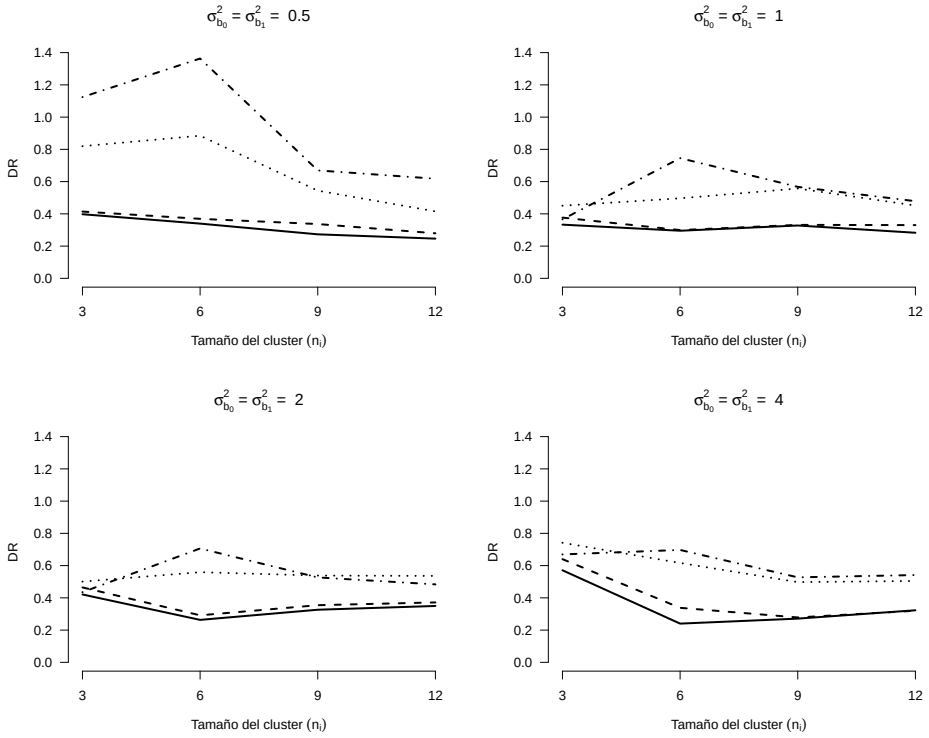

Figura 12: *Mediana de las distancias relativas para las estimaciones de* $\sigma^2_{b_1}$ *en un GLMM Poisson, con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $- \cdot -$ *Tukey. Fuente: elaboración propia.*

Las figuras 13, 14, 15, 16 y 17 presentan los resultados de las medianas de las distancias relativas para las simulaciones del modelo mixto BN con intercepto y pendiente aleatoria. La figura 13 corresponde al parámetro de β_0, allí se observa que, contrario a todos los resultados encontrados hasta aquí, la distribución normal bivariada es la que presenta los mayores valores de las medianas de las distancias relativas, sabiendo que para esta distribución, que es la distribución asumida para el ajuste del modelo mixto con intercepto y pendiente aleatoria no hay especificación incorrecta. Un comportamiento similar al de la normal bivariada lo presenta la distribución exponencial bivariada. El menor impacto de la especificación incorrecta de la distribución de los efectos aleatorios se presentó cuando la distribución verdadera fue la t-student bivariada.

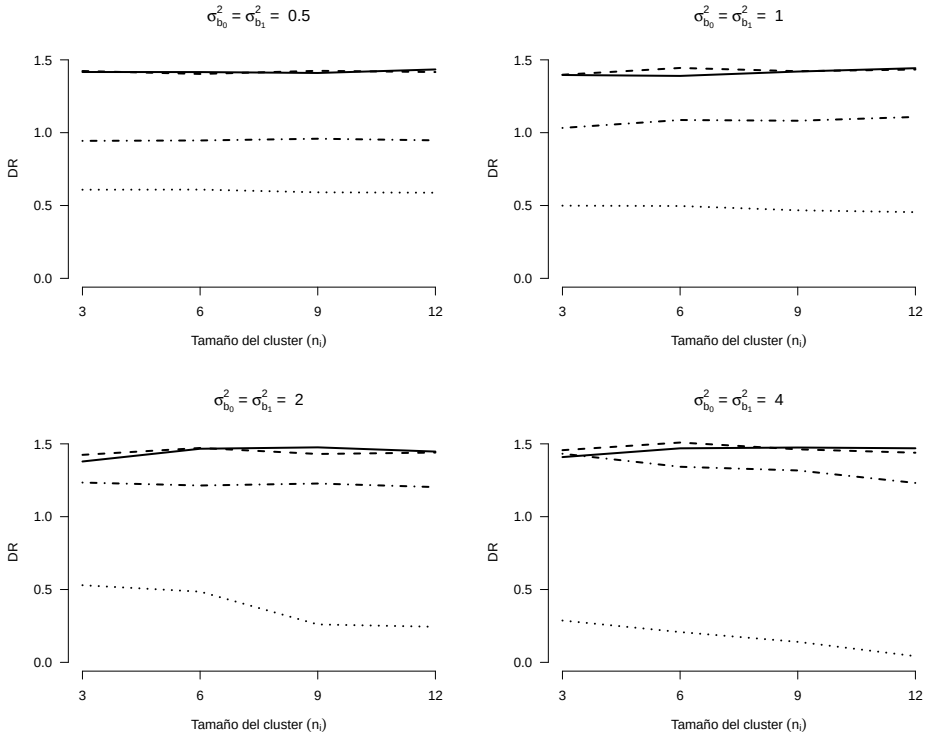

Figura 13: *Mediana de las distancias relativas para las estimaciones de* β_0 *en un GLMM BN con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $-\cdot-$ *Tukey. Fuente: elaboración propia.*

En cuanto al impacto de la especificación incorrecta para el parámetro β_1 (figura 14), se observa que la distribución Tukey bivariada es la que presenta los mayores valores de las distancias relativas, y así, los mayores impactos de la especificación incorrecta. Le sigue la distribución t-student bivariada con valores que oscilan entre el 44 % y el 90 %.

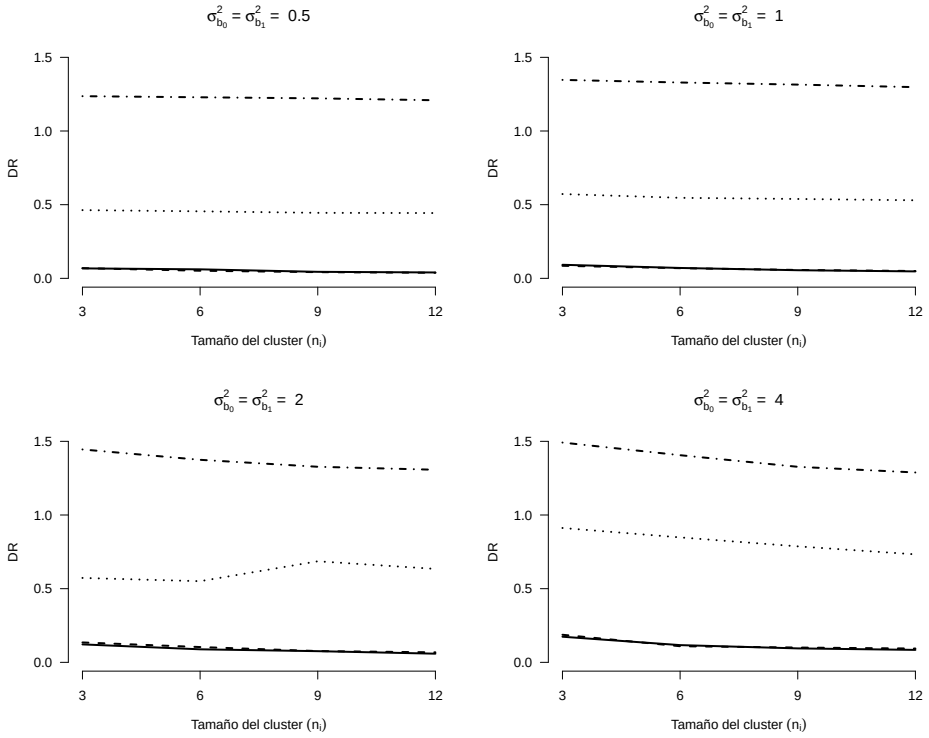

Figura 14: *Mediana de las distancias relativas para las estimaciones de* β_1 *en un GLMM BN con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $-\cdot-$ *Tukey. Fuente: elaboración propia.*

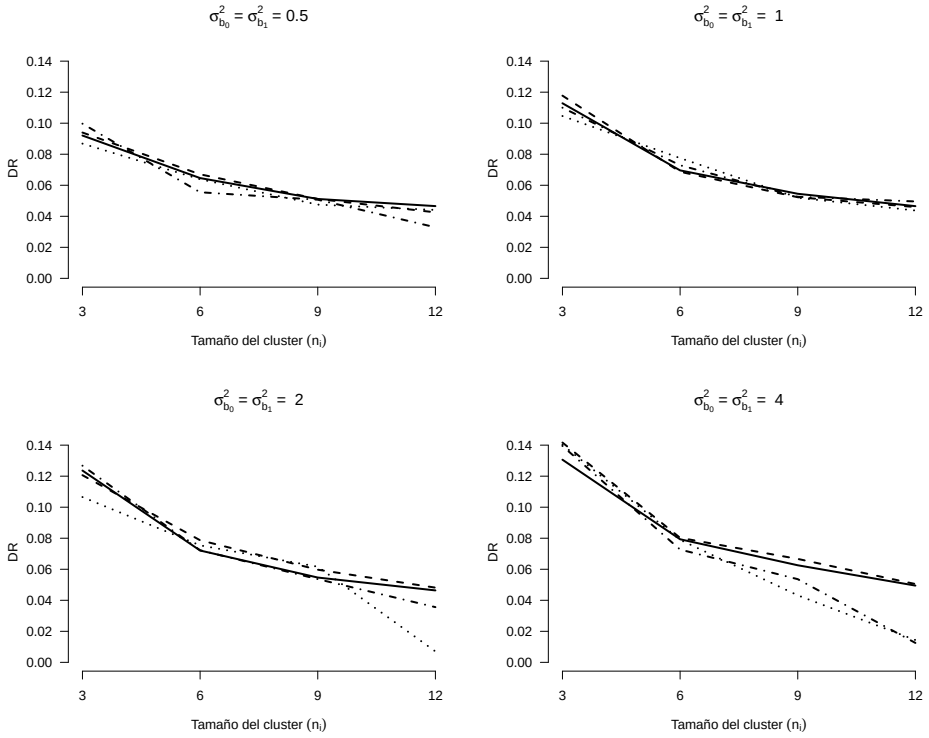

Figura 15: *Mediana de las distancias relativas para las estimaciones de β_2 en un GLMM BN con $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal, · · · t-student, - - - exponencial, — · — Tukey. Fuente: elaboración propia.*

La figura 15 contiene la representación de las medianas de las distancias relativas de la estimación del parámetro β_2, cuyos valores no sobrepasen el 15 % y que no muestra diferencias entre las cuatro distribuciones bivariadas consideradas tanto para el intercepto como para la pendiente aleatoria.

La estimación del componente de varianza correspondiente al intercepto aleatorio $\sigma^2_{b_0}$ y las medidas de cuánto se aleja del verdadero valor de este se puede visualizar en la figura 16, donde se observa un comportamiento muy parecido al que se presenta para las estimaciones de β_0, con los mayores valores de las distancias relativas cuando las verdaderas distribuciones de los efectos aleatorios son la normal y la exponencial bivariadas, sabiendo que para la primera, no hay especificación incorrecta. Vale la pena resaltar que los valores no superan el 60 %, contrario a lo obtenido en las estimaciones del componente de varianza $\sigma^2_{b_0}$ del modelo mixto Poisson, cuyos valores estuvieron por el orden de 1800 %.

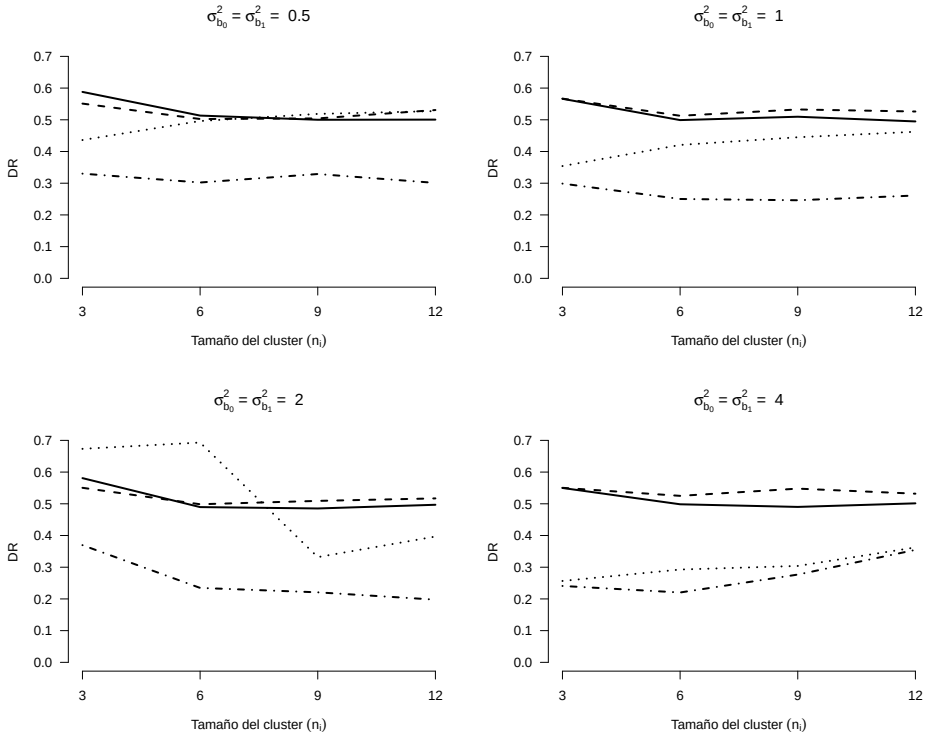

Figura 16: *Mediana de las distancias relativas para las estimaciones de* $\sigma^2_{b_0}$ *en un GLMM BN con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $-\cdot-$ *Tukey. Fuente: elaboración propia.*

Para la estimación del componente de varianza $\sigma^2_{b_1}$ (figura 17) es posible observar que efectivamente hay un impacto de la especificación incorrecta si la distribución verdadera es Tukey bivariada, puesto que es la que presenta los mayores valores de las distancias relativas, excepto cuando $\sigma^2_{b_1} = 2$.

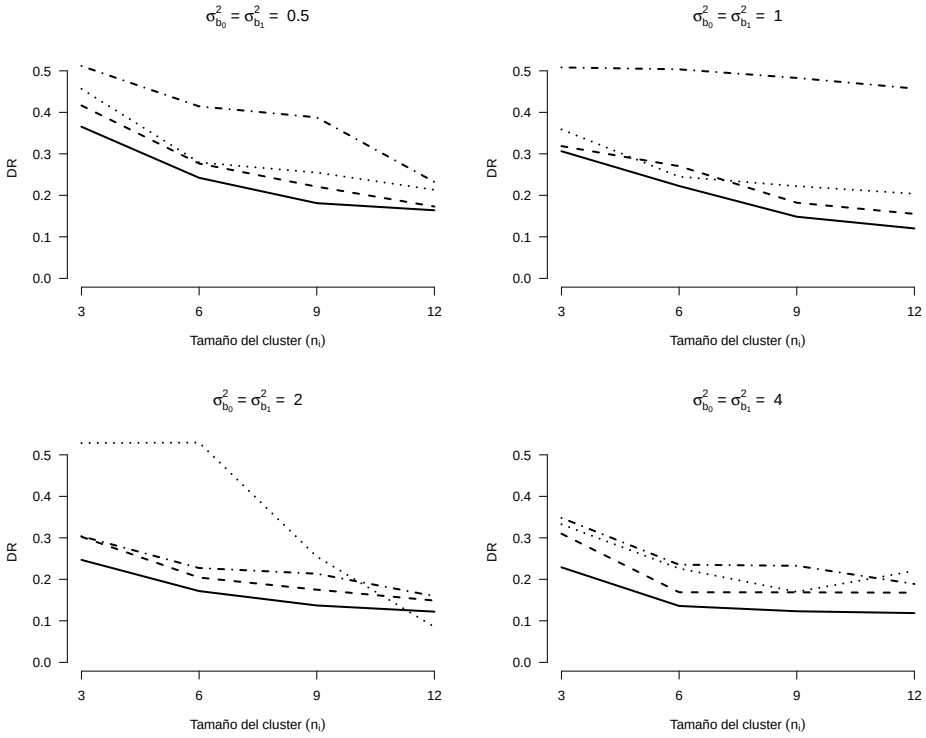

Figura 17: *Mediana de las distancias relativas para las estimaciones de* $\sigma^2_{b_1}$ *en un GLMM BN con* $\sigma^2_{b_0} = \sigma^2_{b_1} = 0.5, 1, 2, 4$ *y cuatro distribuciones bivariadas para el intercepto y pendiente aleatoria: —normal,* $\cdots$ *t-student, - - - exponencial,* $-\cdot-$ *Tukey. Fuente: elaboración propia.*

Finalmente las tablas 3 y 4 presentan la información resumida de todos los casos analizados en el presente artículo. La notación (k/4) indica la cantidad de veces k que la distribución correspondiente tuvo mayores valores de DR dentro de cada uno de los páneles de las figuras.

Tabla 3: *Resultados de los modelos mixtos con intercepto aleatorio. Fuente: elaboración propia.*

parámetros	Caso 1: Poisson con b_0	Caso 2: BN con b_0
β_0	uniforme (1/4)	lognormal (3/4)
β_1	No diferencias	No diferencias
β_2	No diferencias	No diferencias
σ^2	lognormal (4/4)	lognormal (4/4)

Tabla 4: *Resultados de los modelos mixtos con intercepto y pendiente aleatoria. Fuente: elaboración propia.*

parámetros	Caso 3: Poisson con b_0 y b_1	Caso 4: BN con b_0 y b_1
β_0	Tukey (3/4) y t-student (3/4)	normal (4/4) y exponencial (4/4)
β_1	Tukey (4/4) y t-student (4/4)	Tukey (4/4) y t-student (4/4)
β_2	No diferencias	No diferencias
$\sigma^2_{b_0}$	Tukey (3/4)	normal (2/4) y exponencial (2/4)
$\sigma^2_{b_1}$	Tukey (4/4) y t-student (4/4)	Tukey (2/4)

6. Conclusiones

En el caso del modelo mixto Poisson con intercepto aleatorio (b_0), no se observó efecto de especificar incorrectamente la distribución del b_0 en las estimaciones de β_1 ni de β_2, mientras que para las estimaciones de σ^2 siempre se observó efecto cuando la verdadera distribución del intercepto aleatorio fue la lognormal.

Para el caso del modelo mixto BN con intercepto aleatorio (b_0), y al igual que en el caso anterior, no se observó efecto de especificar incorrectamente la distribución del b_0 en las estimaciones de β_1 ni de β_2. Para las estimaciones de β_0 y σ^2 sí se observó un efecto de especificación incorrecta y se dio cuando la verdadera distribución de b_0 fue la lognormal, esto posiblemente se debe a que la lognormal es una distribución asimétrica considerada.

En términos generales, para los dos primeros casos se encontró que las estimaciones del componente de varianza σ^2 fueron las más afectadas por la especificación incorrecta de la distribución de b_0 al ajustar tanto un modelo mixto Poisson o BN y eso se dio cuando la distribución de b_0 fue la lognormal. también hay que resaltar de los resultados del estudio de simulación que el impacto de la especificación incorrecta disminuye a medida que se aumenta el tamaño del conglomerado n_i, al igual que lo observado con el modelo mixto Poisson de intercepto aleatorio.

Para los casos de modelos mixtos Poisson y BN con intercepto y pendiente aleatoria se observó que efectivamente hay un impacto de la especificación incorrecta de las distribuciones de dichos efectos. Para el modelo Poisson los mayores impactos se presentaron en la estimación de los parámetros β_0, β_1, $\sigma^2_{b_0}$ y $\sigma^2_{b_1}$. también se observó que en casi todas las configuraciones, la distribución Tukey bivariada y t-student fueron las que presentaron los mayores valores de las medianas de las distancias relativas.

En los ajustes de un modelo BN con intercepto y pendiente aleatoria se encontró que para las estimaciones de β_1 y de $\sigma^2_{b_1}$sí hubo un impacto de la especificación incorrecta y dicho impacto fue mayor cuando la verdadera distribución de los efectos aleatorios fue la Tukey bivariada. Para las estimaciones del parámetro β_0 y de $\sigma^2_{b_0}$ los mayores valores de las distancias relativas resultaron ser para las distribuciones normal y t-student bivariadas. Finalmente, las estimaciones del parámetro β_2 resultaron ser muy similares al verdadero valor, marcando con esto valores de las distancias relativas no superiores al 15 %.

Con este estudio de simulación que incluía modelos mixtos con intercepto aleatorio únicamente o intercepto y pendiente aleatoria, se logró identificar, en términos generales, que sí hay un impacto de la especificación incorrecta de la distribución de los efectos aleatorios y que dicho impacto se presentó en mayor medida para β_0 y las componentes de varianza en todos los casos considerados. Adicionalmente, para el parámetro β_1 en el caso de intercepto y pendiente aleatoria se observó un impacto de especificación incorrecta. también se logró observar que los mayores valores de DR se obtuvieron principalmente al ajustar los modelos de dos efectos aleatorios (intercepto y pendiente aleatoria) en comparación con los de un solo efecto (intercepto aleatorio). Por último, las distribuciones lognormal, Tukey t-student fueron las que presentaron mayores valores de DR en los casos estudiados.

Recibido: 14 de Septiembre de 2016
Aceptado: 8 de Octubre de 2017

Referencias

Agresti, A., Caffo, B. & Ohman-Strickland, P. (2004), 'Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies', *Computational Statistics and Data Analysis* **47**(3), 639–653.

Alonso, A., Litière, S. & Molenberghs, G. (2008), 'A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models', *Computational statistics and data analysis* **52**(9), 4474–4486.

Alonso, A., Litière, S. & Molenberghs, G. (2010), 'Testing for misspecification in generalized linear mixed models', *Biostatistics* **11**(4), 771–786.

Alonso, A., Milanzi, E., Molenberghs, G., Buyck, C. & Bijnens, L. (2015), 'A new modeling approach for quantifying expert opinion in the drug discovery process', *Statistics in medicine* **34**(9), 1590–1604.

Cook, R. J., Lee, K. A. & Li, H. (2007), 'Non-inferiority trial design for recurrent events', *Statistics in medicine* **26**(25), 4563–4577.

DeGroot, M. H. & Schervish, M. J. (1988), *Probabilidad y estadística*, Editorial Addison Wesley, Mexico.

Fabio, L. C., Paula, G. A. & De Castro, M. (2012), 'A Poisson mixed model with nonnormal random effect distribution', *Computational Statistics and Data Analysis* **56**(6), 1499–1510.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2011), *Applied longitudinal analysis*, segunda edn, John Wiley and Sons, Boston, Massachusetts.

Fournier, D. (2011), 'An introduction to AD Model Builder for use in nonlinear modeling and statistics. Version 10.0 (2011-01-18).'.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A. & Sibert, J. (2012), 'AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models', *Optimization Methods and Software* **27**(2), 233–249.

Gad, A. M. & El Kholy, R. B. (2012), 'Generalized Linear mixed models for Longitudinal Data', *International Journal of Probability and Statistics* **1**(3), 41–47.

Grilli, L. & Innocenti, F. (2016), 'Fitting logistic multilevel models with crossed random effects via Bayesian Integrated Nested Laplace Approximations: a simulation study', *arXiv preprint arXiv:1607.05981* .

Heagerty, P. J. & Kurland, B. F. (2001), 'Misspecified maximum likelihood estimates and generalised linear mixed models', *Biometrika* **88**(4), 973–985.

Hilbe, J. M. (2011), *Negative binomial regression*, Cambridge University Press.

Huang, X. (2009), 'Diagnosis of Random-Effect Model Misspecification in Generalized Linear Mixed Models for Binary Response', *Biometrics* **65**(2), 361–368.

Karim, M. R. & Zeger, S. L. (1992), 'Generalized linear models with random effects; salamander mating revisited', *Biometrics* pp. 631–644.

Komàrek, A. & Lesaffre, E. (2008), 'Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution', *Computational Statistics and Data Analysis* **52**(7), 3441–3458.

Kondo, Y., Zhao, Y. & Petkau, J. (2015), 'A flexible mixed-effect negative binomial regression model for detecting unusual increases in MRI lesion counts in individual multiple sclerosis patients', *Statistics in medicine* **34**(13), 2165–2180.

Litière, S., Alonso, A. & Molenberghs, G. (2007), 'Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models', *Biometrics* **63**(4), 1038–1044.

Litière, S., Alonso, A. & Molenberghs, G. (2008), 'The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models', *Statistics in medicine* **27**(16), 3125–3144.

McCulloch, C. E. & Neuhaus, J. M. (2011), 'Misspecifying the shape of a random effects distribution: why getting it wrong may not matter', *Statistical science* pp. 388–402.

Milanzi, E., Alonso, A. & Molenberghs, G. (2012), 'Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions', *Statistics in medicine* **31**(14), 1475–1482.

Molenberghs, G. & Verbeke, G. (2005), *Models for Discrete Longitudinal Data. Springer Series in Statistics*, Springer.

Neuhaus, J. M., Hauck, W. W. & Kalbfleisch, J. D. (1992), 'The effects of mixture distribution misspecification when fitting mixed-effects logistic models', *Biometrika* **79**(4), 755–762.

Neuhaus, J. M. & McCulloch, C. E. (2006), 'Separating between-and within-cluster covariate effects by using conditional and partitioning methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(5), 859–872.

Neuhaus, J. M. & McCulloch, C. E. (2011a), 'Estimation of covariate effects in generalized linear mixed models with informative cluster sizes', *Biometrika* **98**(1), 147–162.

Neuhaus, J. M. & McCulloch, C. E. (2011b), 'The effect of misspecification of random effects distributions in clustered data settings with outcome-dependent sampling', *Canadian Journal of Statistics* **39**(3), 488–497.

Neuhaus, J. M., McCulloch, C. E. & Boylan, R. (2011), 'A Note on Type II Error Under Random Effects Misspecification in Generalized Linear Mixed Models', *Biometrics* **67**(2), 654–656.

Neuhaus, J. M., McCulloch, C. E. & Boylan, R. (2012), 'Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes', *Statistics in medicine* **32**(14), 2419–2429.

Raudenbush, S. W., Yang, M. & Yosef, M. (2000), 'Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation', *Journal of computational and Graphical Statistics* **9**(1), 141–157.

Skaug, H. J. & Fournier, D. A. (2006), 'Automatic approximation of the marginal likelihood in non-gaussian hierarchical models', *Computational Statistics & Data Analysis* **51**(2), 699–709.

Spiessens, B., Lesaffre, E., Verbeke, G. & Kim, K. (2002), 'Group Sequential Methods for an Ordinal Logistic Random-Effects Model Under Misspecification', *Biometrics* **58**(3), 569–575.

Tsonaka, R., Rizopoulos, D., Verbeke, G. & Lesaffre, E. (2010), 'Nonignorable models for intermittently missing categorical longitudinal responses', *Biometrics* **66**(3), 834–844.

Valencia, A. (2014), 'El uso de la distribución gh en riesgo operativo', *Contaduría y administración* **59**(1), 123–148.

Verbeke, G. & Lesaffre, E. (1997), 'The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data', *Computational Statistics and Data Analysis* **23**(4), 541–556.

Verbeke, G. & Molenberghs, G. (2013), 'The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models', *Biostatistics* **14**(3), 477.

Xiang, L., Yau, K. K. & Lee, A. H. (2012), 'The robust estimation method for a finite mixture of Poisson mixed-effect models', *Computational Statistics and Data Analysis* **56**(6), 1994–2005.

Zhao, Y., Li, D. K., Petkau, A. J., Riddehough, A. & Traboulsee, A. (2014), 'Detection of unusual increases in MRI lesion counts in individual multiple sclerosis patients', *Journal of the American Statistical Association* **109**(505), 119–132.