Revista de la Facultad de Ciencias, *Rev. Fac. Cienc.* Universidad Nacional de Colombia, Sede Medellín
V 7 N°2 julio-diciembre de 2018 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 84 a 95
DOI: https://doi.org/10.15446/rev.fac.cienc.v7n2.66441

EFECTO DE LA ESPECIFICACIÓN INCORRECTA DE LA DISTRIBUCIÓN DE LOS EFECTOS ALEATORIOS EN EL MODELO DE REGRESIÓN BETA CON INTERCEPTO ALEATORIO[a]

EFFECT OF MISSPECIFYING THE RANDOM EFFECTS DISTRIBUTION IN RANDOM INTERCEPT BETA REGRESSION MODEL

OLGA CECILIA USUGA MANCO[b]

Recibido 13-02-2018, aceptado 14-06-2018, versión final 18-06-2018.
Artículo Investigación

RESUMEN: La estimación en el modelo de regresión beta con intercepto aleatorio esta usualmente basada en la teoría de máxima verosimilitud, asumiendo que el modelo esta correctamente especificado. Sin embargo, la validez de este supuesto algunas veces es difícil de verificar. El objetivo de este trabajo es estudiar el impacto de la especificación incorrecta de la distribución de los interceptos aleatorios de la media y la dispersión en la estimación de los parámetros del modelo a través de un estudio de simulación. Los resultados de las simulaciones mostraron la existencia de un efecto en las estimaciones de los parámetros cuando se usa la distribución mezcla de normales y cuando la cantidad de información por grupo es pequeña.

PALABRAS CLAVE: Efectos aleatorios; especificación incorrecta; distancia relativa; distribución de efectos aleatorios; regresión beta.

ABSTRACT: Estimation in random intercept beta regression model is often based on maximum likelihood theory, which assumes that the underlying probability model is correctly specified. However, the validity of this assumption is sometimes difficult to verify. The objective of this paper is to study the impact of random effects distribution misspecification on the parameter estimation. The simulation results showed the existence of an effect in the parameter estimates when mixed normal distribution are used and when the amount of information per group is small.

KEYWORDS: Random effects; misspecification; relative distance; random effects distribution; beta regression.

1. INTRODUCCIÓN

Los datos proporcionales provienen de estudios prácticos en medicina, ciencias sociales y educación, donde las respuestas se limitan a un intervalo (a, b) (Ferrari & Cribari-Neto, 2004). Por su parte, los datos longi-

[a]Usuga, O. C. (2018). Efecto de la especificación incorrecta de la distribución de los efectos aleatorios en el modelo de regresión beta con intercepto aleatorio. *Rev. Fac. Cienc.*, 7(**2**), 84–95. DOI: https://doi.org/10.15446/rev.fac.cienc.v7n2.66441

[b]INCAS, Departamento de Ingeniería Industrial, Facultad de Ingeniería, Universidad de Antioquia. olga.usuga@udea.edu.co

tudinales se obtienen a partir de observaciones repetidas de una variable respuesta a través del tiempo, lo cual permite analizar posibles alteraciones en las características de un grupo o un individuo (Verbeke & Molenberghs, 2000). Algunos estudios prácticos requieren el análisis de datos longitudinales donde la variable respuesta se limita a un intervalo (a,b) (Song *et al.*, 2004) y donde la distribución usual para modelar esta variable es la distribución beta.

Un método utilizado para el análisis de tales datos son los modelos de regresión beta mixtos. Un aspecto importante de estos modelos es la suposición de que la variabilidad observada en la variable respuesta se puede modelar a través de los efectos aleatorios, los cuales se asumen que tienen una distribución predeterminada. En estos modelos la estimación de los parámetros se obtiene usualmente maximizando la función de verosimilitud, en la cual intervienen los datos y las funciones de densidad de la variable respuesta y de los efectos aleatorios. Este tipo de modelos ha sido estudiado desde el punto de vista frecuentista por Usuga (2013) y desde el punto de vista bayesiano por Galvis *et al.* (2014).

En el análisis de modelos mixtos es común asumir que la distribución de los efectos aleatorios es normal debido a aspectos matemáticos y computacionales (Alonso *et al.*, 2010). Sin embargo, si se especifica de manera incorrecta esta distribución las estimaciones de máxima verosimilitud del modelo pueden sufrir alguna alteración en sus propiedades. En los modelos mixtos este problema ha sido estudiado ampliamente por Verbeke & Lessafre (1997), Heagerty & Kurland (2001), Agresti *et al.* (2004), Litiere *et al.* (2007), Alonso *et al.* (2008), Huang (2009), Alonso *et al.* (2010), McCulloch & Neuhaus (2011a), McCulloch & Neuhaus (2011b), Neuhaus *et al.* (2013), Verbeke & Molenberghs (2013), Efendi *et al.* (2014), Bartolucci *et al.* (2017) y Drikvandi *et al.* (2017). Sin embargo, para el caso del modelo de regresión beta mixto no se han reportado trabajos al respecto.

Los estudios de especificación incorrecta de los efectos aleatorios iniciaron a finales de la década de los 90 del siglo XX, cuando Verbeke & Lessafre (1997) encontraron resultados asintóticos bajo el supuesto de no normalidad de los efectos aleatorios en un modelo mixto para datos longitudinales. Estos autores mostraron que para los efectos fijos, sus errores estándar no se veían afectados por la especificación incorrecta, sin embargo, para los componentes de varianza se observaban diferencias. Seguido de este estudio, Heagerty & Kurland (2001) estudiaron el impacto de especificar incorrectamente la distribución de los efectos aleatorios en un modelo de regresión lineal generalizado sobre los coeficientes de regresión, específicamente a través del sesgo relativo asintótico, encontrando sesgo en el modelo estudiado cuando la distribución de los efectos aleatorios dependía de covariables medidas. En el caso de modelos mixtos para datos binarios y para datos de sobrevivencia, Agresti *et al.* (2004) estudiaron el efecto de asumir distribución normal para los efectos aleatorios cuando en realidad la verdadera distribución se alejaba de la normal, mostrando que existe pérdida de eficiencia en la predicción de la variable respuesta. Finalmente, en el análisis de un modelo de regresión logístico con intercepto aleatorio Litiere *et al.* (2007) analizaron el impacto de la especificación incorrecta de la distribución de los interceptos aleatorios en los errores tipo I y II, encontrando que la tasa de error tipo

I puede aumentar al cambiar la distribución del intercepto aleatorio.

McCulloch & Neuhaus (2011a), McCulloch & Neuhaus (2011b) y Neuhaus *et al.* (2013) evaluaron, en el contexto de los modelos lineales generalizados, el impacto de la especificación incorrecta de la distribución de los efectos aleatorios en la predicción de los efectos aleatorios y en las estimaciones de los parámetros.

Además de los estudios mencionados anteriormente, se han propuesto pruebas para identificar la especificación incorrecta en los modelos de regresión mixtos. Alonso *et al.* (2008) y Alonso *et al.* (2010) desarrollaron pruebas de diagnóstico para evaluar la especificación incorrecta basadas en los valores propios de las matrices de varianzas y covarianzas de las estimaciones de los efectos fijos y en representaciones de la matriz de información del modelo, Huang (2009) propuso un método de diagnóstico comparando las inferencias basadas en los datos originales y los reconstruidos. Verbeke & Molenberghs (2013) propusieron la función gradiente como una herramienta gráfica exploratoria para verificar la bondad de ajuste de la distribución de los efectos aleatorios en el modelo mixto y Efendi *et al.* (2014) desarrollaron una prueba de bondad de ajuste para la distribución de los efectos aleatorios en modelos mixtos basado en la función gradiente.

En este artículo se estudia el impacto que existe al especificar incorrectamente los efectos aleatorios en el modelo de regresión beta con intercepto aleatorio sobre las estimaciones de los parámetros. En la sección 2 se describe el modelo de regresión y cada uno de sus componentes. Luego, en la sección 3 se considera el método de estimación del modelo. Un estudio de simulación que considera diferentes escenarios se describe en la sección 3. Finalmente, en la sección 4 se presentan las conclusiones.

2. MODELO DE REGRESIÓN BETA CON INTERCEPTO ALEATORIO

Si y es una variable aleatoria con distribución beta, entonces una parametrización de su densidad en términos de la media μ y el parámetro de dispersión σ esta dada por

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{(1-\sigma^2)}{\sigma^2}\right)}{\Gamma\left(\mu\frac{(1-\sigma^2)}{\sigma^2}\right)\Gamma\left((1-\mu)\left(\frac{(1-\sigma^2)}{\sigma^2}\right)\right)} y^{\mu\left(\frac{(1-\sigma^2)}{\sigma^2}\right)-1}(1-y)^{(1-\mu)\left(\frac{(1-\sigma^2)}{\sigma^2}\right)-1}, \tag{1}$$

con $0 < y < 1$, $0 < \mu < 1$ y $0 < \sigma < 1$. En esta parametrización, $y \sim Be(\mu,\sigma)$ y la media y la varianza de y son $\mathrm{E}(y) = \mu$ y $\mathrm{Var}(y) = \sigma^2\mu(1-\mu)$.

La Figura 1 muestra algunas densidades de la distribución beta junto con los correspondientes valores de (μ,σ). Se destaca que las densidades exhiben formas muy diferentes dependiendo de los valores de los dos parámetros. En particular, la distribución puede ser simétrica, cuando μ =0.5, o asimétrica, cuando $\mu \neq 0.5$.

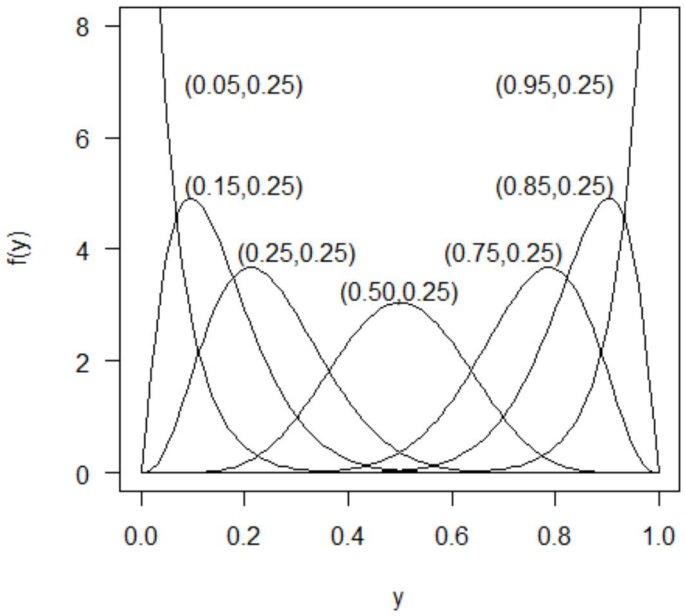

Figura 1: Densidades beta para diferentes combinaciones de $(\mu.\sigma)$. Fuente: Elaboración propia.

Sin pérdida de generalidad, se asumirá que $(0,1) = (a,b)$, donde a y b son escalares conocidos con $a < b$. Si la variable respuesta está limitada al intervalo (a,b), se podrá modelar $(y-a)/(b-a)$ en lugar de y.

2.1. Modelo

Sean y_{ij} las mediciones observadas en el i-ésimo grupo con $i = 1,2,\ldots,N$ y t_{ij} con $j = 1,2,\ldots,n_i$ los correspondientes tiempos en los cuales se toman las mediciones en cada grupo i. En el modelo de regresión beta con intercepto aleatorio se asume que la distribución condicional de y_{ij} dado $\mathbf{b}_i = (b_{i1}, b_{i2})^\top$ sigue una distribución beta con una densidad determinada por la expresión (1). Se asumirá el siguiente modelo:

$$\begin{aligned} y_{ij} \mid b_{i1}, b_{i2} &\sim \text{Be}(\mu_{ij}, \sigma_{ij}), \\ g_1(\mu_{ij}) &= \eta_{ij1} = \mathbf{x}_{ij1}^\top \beta_1 + b_{i1}, \\ g_2(\sigma_{ij}) &= \eta_{ij2} = \mathbf{x}_{ij2}^\top \beta_2 + b_{i2}, \end{aligned} \tag{2}$$

donde $\mathbf{x}_{ij1} = (x_{ij11}, x_{ij21}, \ldots, x_{ijp_11})^\top$ y $\mathbf{x}_{ij2} = (x_{ij12}, x_{ij22}, \ldots, x_{ijp_22})^\top$ son vectores de covariables, $\beta_1 = (\beta_{11}, \beta_{21}, \ldots, \beta_{p_11})^\top$ y $\beta_2 = (\beta_{12}, \beta_{22}, \ldots, \beta_{p_22})^\top$ son vectores de parámetros fijos no dependientes del tiempo, y b_{i1} y b_{i2} son los interceptos aleatorios. Las funciones de enlace conocidas $g_1 : (0,1) \to \Re$ y $g_2 : (0,1) \to \Re$ son estrictamente monótonas y doblemente diferenciables. Se puede usar la misma o diferentes funciones de enlace para la media y el parámetro de dispersión, por ejemplo, logit, probit, clog-log, log-log o cauchit. Para una discusión de estas funciones de enlace ver McCullagh & Nelder (1989).

Los interceptos aleatorios b_{i1} y b_{i2}, los cuales se comparten entre mediciones del mismo grupo, son variables aleatorias normales independientes e idénticamente distribuidas,

$$\begin{aligned} b_{i1} &\overset{\text{i.i.d}}{\sim} N(0,\tau_1^2), \\ b_{i2} &\overset{\text{i.i.d}}{\sim} N(0,\tau_2^2), \end{aligned} \tag{3}$$

donde τ_1^2 y τ_2^2 son las varianzas de los interceptos aleatorios. El caso particular en el cual $\tau_1^2 = 0$ muestra que la media de la variable respuesta se puede modelar sin intercepto aleatorio. El vector de parámetros para el modelo (2) está dado por $\theta = \left(\beta_1^\top, \beta_2^\top, \tau_1^2, \tau_2^2\right)^\top$.

2.2. Método de estimación

Sea θ el vector de parámetros, $f(y_{ij} \mid b_{i1}, b_{i2}; \beta_1, \beta_2)$ la función de densidad de probabilidad de las observaciones dados los efectos aleatorios, $f(b_{i1}; \tau_1^2)$ y $f(b_{i2}; \tau_2^2)$ las funciones de densidad de probabilidad de los inteceptos aleatorios. La distribución marginal de las observaciones $\mathbf{y}_i$ para el grupo i está dada por

$$f(\mathbf{y}_i;\theta) = \int\int_{\Re^2} \prod_{j=1}^{n_i} f(y_{ij} \mid b_{i1}, b_{i2}; \beta_1, \beta_2) \cdot f(b_{i1}; \tau_1^2) f(b_{i2}; \tau_1^2)\, \mathrm{d}b_{i1}\mathrm{d}b_{i2}, \tag{4}$$

y la función de verosimilitud de θ dados los datos observados $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^\top$ está dada por

$$L(\theta) = \prod_{i=1}^{N} \int\int_{\Re^2} \prod_{j=1}^{n_i} f(y_{ij} \mid b_{i1}, b_{i2}; \beta_1, \beta_2) \cdot f(b_{i1}; \tau_1^2) f(b_{i2}; \tau_2^2)\, \mathrm{d}b_{i1}\mathrm{d}b_{i2}. \tag{5}$$

A diferencia de un modelo lineal con distribución normal, la distribución marginal (4) y la función de verosimilitud (5) no tienen solución analítica. La principal dificultad de la inferencia basada en el método de máxima verosimilitud para un modelo de regresión beta con intercepto aleatorio es la evaluación de las integrales intratables en la función de verosimilitud (5). Los métodos más usados para su evaluación incluyen métodos de integración Monte Carlo, algoritmo EM, y métodos aproximados, ver Wu (2010). En este trabajo se usó la cuadratura de Gauss-Hermite multivariada para aproximar las integrales de la función de verosimilitud dada en (5). Así, la función de verosimilitud se puede escribir en forma aproximada como

$$L(\theta) \cong \prod_{i=1}^{N} \left(\sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f(y_{ij} \mid \sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2}; \beta_1, \beta_2) \frac{w_{k_1} w_{k_2}}{\pi} \right)$$

y la función de log-verosimilitud queda escrita como

$$\ell(\theta) \cong \sum_{i=1}^{N} \log \left(\sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid \sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2}; \beta_1, \beta_2) \frac{w_{k_1} w_{k_2}}{\pi} \right), \tag{6}$$

donde Q_1 y Q_2 son el número de puntos de cuadratura, z_{k_1} y z_{k_2} son los puntos de la cuadratura y w_{k_1} y w_{k_2} son los pesos de la cuadratura correspondientes. Para una discusión detallada de la cuadratura de Gauss-Hermite multivariada ver Fahrmeir & Tutz (2001).

3. ESTUDIO DE SIMULACIÓN

En esta sección se evalúa el impacto de la especificación incorrecta de la distribución de los efectos aleatorios en la consistencia de los estimadores de máxima verosimilitud a partir de un estudio de simulación. En este artículo se adoptó el enfoque utilizado por Verbeke & Lessafre (1997), Agresti *et al.* (2004), Litiere *et al.* (2008) y Alonso *et al.* (2008), en el cual se generan los valores de los interceptos aleatorios a partir de distribuciones normal, uniforme y mezcla de normales y se estiman los parámetros del modelo asumiendo distribución normal para los interceptos aleatorios.

3.1. Estructura del estudio de simulación

El modelo usado para simular los datos del estudio fue el siguiente:

$$\begin{aligned} y_{ij} \mid b_{i1}, b_{i2} &\overset{\text{ind}}{\sim} \text{Be}(\mu_{ij}, \sigma_{ij}), \\ g_1(\mu_{ij}) &= \eta_{ij1} = \beta_{11} + \beta_{21} x_{ij} + \beta_{31} t_i + b_{i1}, \\ g_2(\sigma_{ij}) &= \eta_{ij2} = \beta_{12} + \beta_{22} x_{ij} + \beta_{32} t_i + b_{i2}, \end{aligned} \tag{7}$$

donde $i = 1, \ldots, N$, $j = 1, \ldots, n$, $g_1(\cdot)$ y $g_2(\cdot)$ son las funciones de enlace logit y $\beta_1 = (\beta_{11}, \beta_{21}, \beta_{31})^\top$ y $\beta_2 = (\beta_{12}, \beta_{22}, \beta_{32})^\top$ son los vectores de parámetros fijos asociados a μ y σ, respectivamente.

Los interceptos aleatorios b_{i1} y b_{i2} son variables aleatorias independientes e idénticamente distribuidas,

$$\begin{aligned} b_{i1} &\overset{\text{i.i.d}}{\sim} G_T, \\ b_{i2} &\overset{\text{i.i.d}}{\sim} G_T, \end{aligned} \tag{8}$$

donde G_T corresponde a la verdadera distribución de los interceptos aleatorios b_{i1} y b_{i2} y en este estudio de simulación G_T fue normal, uniforme y mezcla de normales. En la Figura 2 se muestran las densidades para las tres distribuciones usadas para generar los interceptos aleatorios, cada distribución G_T se caracteriza por tener media 0 y varianza τ^2.

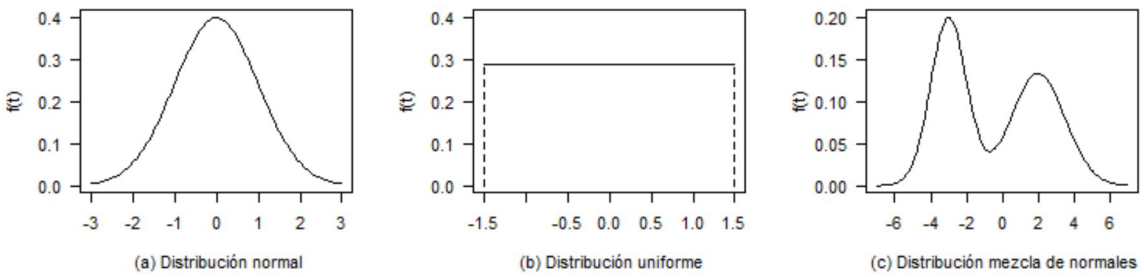

Figura 2: Distribuciones de probabilidad utilizadas para generar los interceptos aleatorios. Fuente: Elaboración propia.

Los valores de los parámetros fueron $\beta_1 = \beta_2 =$ (-0.15,0.15,-0.15)$^\top$. La variable x_{ij} se generó de acuerdo a una distribución uniforme $U(0,1)$, donde para cada par (i, j) se generó un valor diferente. La variable t_i,

con valores entre cero y uno, se generó como $t_i = (n-1)/n$ y tomó los mismos valores para cada grupo. Las varianzas de los interceptos aleatorios que se consideraron fueron τ^2 =1.0, 1.5, 2.0. Se analizaron todas las combinaciones de número de grupos, $N = 10.15, 20, 25, 35, 50, 65, 85, 100$ y número de observaciones por grupo, $n = 3, 5, 8, 12, 20, 40$.

Para el método de cuadratura de Gauss-Hermite, se usaron $Q_1 = Q_2 = 8$ puntos de cuadratura para estimar los interceptos aleatorios. Todos los análisis se llevaron a cabo en `R`, (R, 2017). Para calcular los puntos y pesos necesarios en la cuadratura de Gauss-Hermite se usó el paquete `glmmML` de `R` propuesto por Brostöm & Holmberg (2011) y para maximizar la función de verosimilitud (6) se usó la función `nlminb` del paquete `stats` de `R` desarrollada por Gay (1990). El número de simulaciones fue de 1000.

Para evaluar el comportamiento de las estimaciones de máxima verosimilitud, β_1, β_2, τ_1^2, y τ_2^2 se calculó la distancia relativa entre el vector de parámetros y el vector de estimaciones (Verbeke & Lessafre, 1997),

$$\text{DR} = \frac{\|\hat{\theta} - \theta\|}{\|\theta\|}.$$

En los resultados de simulaciones se presenta la distancia relativa promedio para evaluar el impacto de la especificación incorrecta de los interceptos aleatorios en la estimación de los parámetros. Valores altos de la distancia relativa promedio indican que existe diferencia entre las estimaciones de los parámetros y los valores reales de los parámetros debido a la especificación incorrecta de la distribución de los interceptos aleatorios.

Los interceptos aleatorios se ajustaron considerando como distribución la normal $b_{i1} \sim N(0, \tau^2)$ y $b_{i2} \sim N(0, \tau^2)$.

3.2. Resultados

La Figura 3 muestra el comportamiento de la distancia relativa para $N = 10$ grupos con diferente número de observaciones por grupo $n = 3, 5, 8, 12, 20$ y 40. Adicionalmente, en la figura se muestran tres paneles que corresponden a varianzas de τ^2=1.0, 1.5 y 2.0, y por medio de tres tipos de líneas se diferencian las distribuciones verdaderas de los interceptos aleatorios.

A partir del análisis de los tres paneles de la Figura 3 se observó que en el caso en el que la varianza fue de 1.0 no se encontraron diferencias en el comportamiento de la distancia relativa promedio cuando se generaron los interceptos aleatorios a partir de las distribuciones normal, uniforme y mezcla de normales. Sin embargo, cuando la varianza asumió un valor de τ^2 =1.5, se observó que la distancia relativa promedio fue mayor para el caso de intercepto aleatorio uniforme. Asimismo, cuando la varianza fue de τ^2 =2.0 y los interceptos aleatorios se generaron a partir de una mezcla de normales, los valores de la distancia relativa promedio fueron superiores a los casos en los que se generaron a partir de una normal y una uniforme. De la

Figura 3 se observa que la distancia relativa disminuye a medida que aumenta el número de observaciones por grupo n y/o a medida que aumenta la varianza τ^2. El valor máximo de la distancia relativa promedio observado fue de 2.36 y se presentó cuando los interceptos aleatorios se generaron a partir de la distribución mezcla de normales, $n = 3$ y $\tau^2 = 1.0$. El valor mínimo fue de 0.87 y se presentó cuando los interceptos aleatorios se generaron a partir de la distribución normal, $n = 40$ y $\tau^2 = 2.0$.

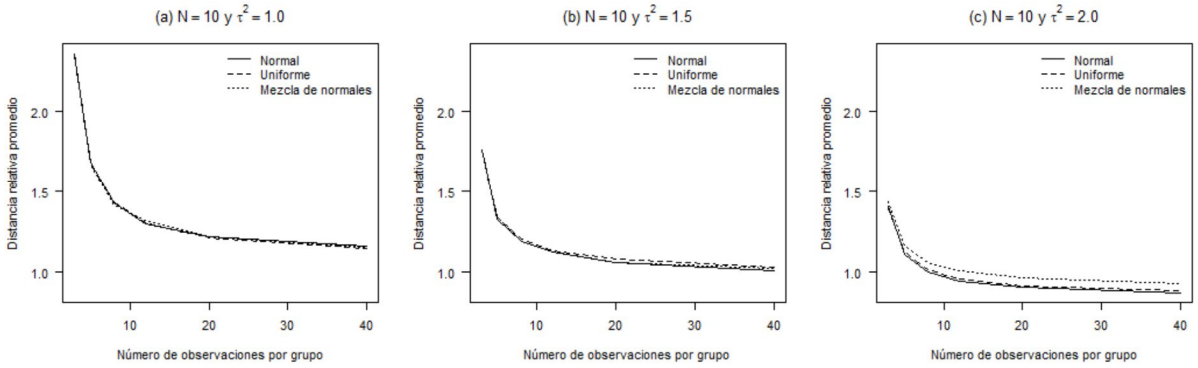

Figura 3: DR promedio versus número de observaciones por grupo (n) para $N = 10$, varianzas $\tau^2 = 1.0, 1.5$ y 2.0 y tres distribuciones para los efectos aleatorios. Fuente: Elaboración propia.

La Figura 4 muestra el comportamiento de la distancia relativa promedio para $N = 15$ grupos. Cuando la varianza de los interceptos aleatorios fue $\tau^2 = 1.0$ o $\tau^2 = 2.0$ y los interceptos aleatorios se generaron a partir de la distribución mezcla de normales, las distancias relativas promedio que se obtuvieron fueron superiores a las de los casos en los que las distribuciones fueron la uniforme y la normal. En el caso en el que la varianza tomó el valor de $\tau^2 = 1.5$, se notó un leve aumento de la distancia relativa promedio cuando la distribución de los interceptos fue la uniforme. En este caso, con $N = 15$, la máxima distancia relativa promedio que se encontró fue de 1.94 bajo el caso de la distribución uniforme con $n = 3$ y $\tau^2 = 1.0$. La mínima distancia relativa promedio fue de 0.86 bajo el caso de la distribución normal con $n = 40$ y $\tau^2 = 2.0$. Al comparar con el caso de $N = 10$ se nota que las distancias relativas disminuyeron.

El comportamiento de la distancia relativa cuando se consideraron $N = 20$ grupos se observa en la Figura 5. Cuando la varianza tomó el valor de $\tau^2 = 1.0$, las distancias relativas promedio obtenidas con la generación de las tres diferentes distribuciones fue similar. Sin embargo, cuando las varianzas aumentaron a valores de 1.5 y 2.0, las distancias relativas promedio mayores se obtuvieron cuando los interceptos aleatorios se generaron a partir de una distribución de mezcla de normales. Como en el caso anterior, las distancias relativas promedio disminuyeron cuando el número de grupos, el número de observaciones por grupo y la varianza aumentaron. El valor máximo obtenido fue de 1.73 bajo el caso de la distribución mezcla de normales, $n = 3$ y $\tau^2 = 1.0$ y el mínimo obtenido fue de 0.85 bajo el caso de la distribución normal, $n = 40$ y $\tau^2 = 2.0$.

Las Figuras 3, 4 y 5 muestran la consistencia de los estimadores de máxima verosimilitud cuando aumenta

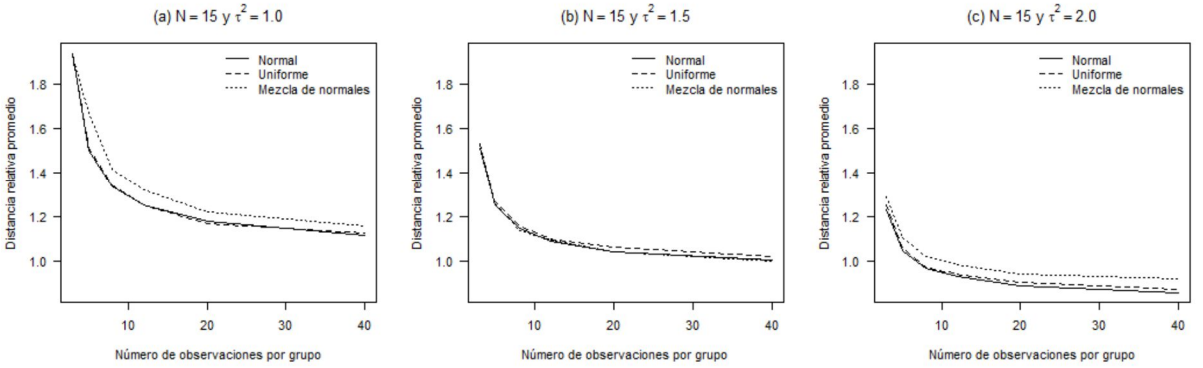

Figura 4: DR promedio versus número de observaciones por grupo (n) para $N = 15$, varianzas τ^2 =1.0,1.5 y 2.0 y tres distribuciones para los efectos aleatorios. Fuente: Elaboración propia.

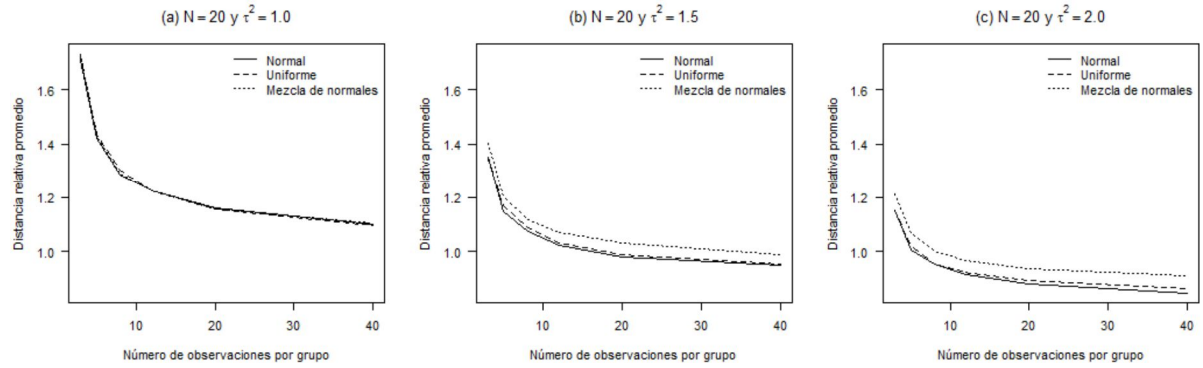

Figura 5: DR promedio versus número de observaciones por grupo (n) para $N = 20$, varianzas τ^2 =1.0,1.5 y 2.0 y tres distribuciones para los efectos aleatorios. Fuente: Elaboración propia.

la información por grupo y cuando aumenta el número de grupos. Figuras para $N = 25, 35, 50, 65, 85$ y 100 fueron construídas (no mostradas aquí) y de ellas se obtuvieron patrones similares a los mostrados en las Figuras 3, 4 y 5.

La Tabla 1 muestra las distribuciones de los interceptos aleatorios en los cuales se observaron las mayores distancias relativas promedio para un número de grupos N y varianza τ^2 fijos. De la tabla se observa que, para tamaños de grupo de $N = 10, 20$ y 25 y varianza de τ^2 =1.0, el comportamiento de la distancia relativa promedio fue similar para las tres distribuciones consideradas. Además, se observa que cuando τ^2 =1.0 la distribución de los interceptos aleatorios que generó mayor distancia relativa promedio fue la uniforme en un 33 % de los casos, mientras que cuando τ^2 =1.5 y τ^2 =2.0 la distribución que generó mayores distancias relativas promedio fue la de mezcla de normales en un 55 % y en un 100 %, respectivamente.

Tabla 1: Distribución de los interceptos aleatorios que generó mayor distancia relativa promedio con un número de grupos N y varianza τ^2 específicos.

N	τ^2 =1.0	τ^2 =1.5	τ^2 =2.0
10	Todas	Uniforme	Mezcla de normales
15	Mezcla de normales	Uniforme	Mezcla de normales
20	Todas	Mezcla de normales	Mezcla de normales
25	Todas	Mezcla de normales	Mezcla de normales
35	Mezcla de normales	Mezcla de normales	Mezcla de normales
50	Normal	Uniforme	Mezcla de normales
65	Uniforme	Uniforme	Mezcla de normales
85	Uniforme	Mezcla de normales	Mezcla de normales
100	Uniforme	Mezcla de normales	Mezcla de normales

4. CONCLUSIONES

El estudio de simulación fue llevado a cabo para analizar el impacto de la especificación incorrecta de la verdadera distribución de los efectos aleatorios en un modelo de regresión beta mixto. Los datos del estudio fueron obtenidos a partir de la generación de interceptos aleatorios con distribuciones normal, uniforme y mezcla de normales. En el proceso de estimación de los parámetros del modelo se asumió que los interceptos aleatorios tenían distribución normal y se analizó el desempeño del proceso a partir de la distancia relativa.

A partir de los resultados obtenidos en las simulaciones se encontró que el efecto de la especificación incorrecta de la distribución de los efectos aleatorios tiende a disminuir cuando el número de grupos N y el número de observaciones por grupo n aumenta. El resultado anterior, similar al encontrado por Rizopoulos *et al.* (2008) en modelos de parámetros compartidos con aplicación en estudios longitudinales, muestra la importancia del número de observaciones por grupo en el proceso de estimación de los parámetros.

Los resultados de las distancias relativas considerando todos los valores de N mostraron un efecto de la especificación incorrecta de la distribución de los efectos aleatorios cuando la varianza tomó el valor de τ^2 =2.0 y los interceptos aleatorios se generaron a partir de mezclas de normales. La Tabla 1 mostró que existe un impacto en la estimación de los parámetros del modelo al especificar de forma incorrecta la distribución de los efectos aleatorios, en particular cuando la distribución no es simétrica.

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.; Litiere, 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.; Litiere, S. & Molenberghs, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostatistics,* 11(**4**), 771-786.

Bartolucci, F.; Bacci, S. & Pigini, C. (2017). Misspecification test for random effects in generalized linear finite-mixture models for clustered binary and ordered data. *Econometrics and Statistics,* 3, 112-131.

Broström, G. & Holmberg, H. (2011). R: glmmML: Generalized linear models with clustering. *R package version 0.82-1.* Recuperado de http://CRAN.R-project.org/package=glmmML.

Drikvandi, R.; Verbeke, G. & Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics,* 73, 63-71.

Efendi, A.; Drikvandi, R.; Verbeke, G. & Molenberghs, G. (2014). A goodness-of-fit test for the random-effects distribution in mixed models. *Statistical Methods in Medical Research,* 26(**2**), 970-983.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models.* New York: Springer Science & Business Media.

Ferrari, S. & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics,* 31(**7**), 799-815.

Galvis, D. M.; Bandyopadhyay, D. & Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in medicine,* 33(**21**), 3759-3771.

Gay, D.M. (1990). Usage summary for selected optimization routines. *Computing Science Technical Report,* 153, 1-21.

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

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

Litiere, S.; Alonso, A. & Molenbergs, G. (2007). Type I and Type II Error under Random Effects Misspecification in Generalized Linear Mixed Models. *Biometrics,* 63(**4**), 1038-1044.

Litiere, 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.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models.* London: Chapman and Hall.

McCulloch, C.E. & Neuhaus, J.M. (2011a). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science,* 26, 388-402.

McCulloch, C.E. & Neuhaus, J.M. (2011b). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics,* 67(**1**), 270-279.

Neuhaus, J.M.; McCulloch, C.E. & Boylan, R. (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution or random intercepts and slopes. *Statisics in Medicine,* 32(**14**), 2419-2429.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Recuperado de https://www.R-project.org/

Rizopoulos, D.; Verbeke, G. & Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika,* 95(**1**), 63-74.

Song, P.; Qiu, Z. & Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal,* 46(**5**), 540-553.

Usuga, O.C. (2013). Modelos de regressão beta com efeitos aleatórios normais e nao normais para dados longitudinais. Tese de doutorado. Instituto de Matemática e Estatística da USP. São Paulo.

Verbeke, G. & Lessafre, 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. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

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**), 477-490.

Wu, L. (2010). *Mixed effects models for complex data.* Boca Raton: Chapman and Hall.