UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2019, Vol. 12, No. 2, pp. 172–192

Aplicación de un Modelo Lineal Generalizado Mixto para datos anidados

Application of a Generalized Linear Mixed Model for nested data

Eduardo Patricio Ceriani Rodriguez[a]
eduardo.ceriani@gmail.com

Resumen

El rendimiento académico es un fenómeno multidimensional, determinado por el estudiante, la escuela y el contexto socio-económico-cultural en que éstos agentes se desenvuelven. Con el objeto de medir el rendimiento de los establecimientos educacionales, el Ministerio de Educación aplica una batería de pruebas estandarizadas que componen el Sistema de Medición de la Calidad de la Educación, SIMCE. Existen varios factores que determinan el rendimiento en el SIMCE. En este sentido, diversos autores sostienen que el nivel socio-económico de los estudiantes es el más importante para explicar el rendimiento escolar. Por otra parte, los datos indican que tanto la dependencia escolar como la zona geográfica donde está ubicado el establecimiento educacional influyen en el rendimiento. Este estudio tiene por objetivo determinar cuáles son los mejores predictores del puntaje SIMCE de matemáticas en establecimientos educacionales de la región de la Araucanía. Los datos de carácter anidado serían modelados a través de un Modelo Lineal Generalizado Mixto (GLMM).

***Palabras clave*:** Modelo lineal generalizado mixto, datos anidados, rendimiento académico.

Abstract

The school performance is a multidimensional phenomenon determined by the student, the school and the socio-economical context in which they are immerse. In order to measure the performance of the educational institutions, the Ministry of Education applies a set of standardized tests which form the Measurement System of Quality in Education, SIMCE for its initials in Spanish. There are some factors which determine the performance in the SIMCE. Along this line, many authors hold that the socio-economical level of the students is the most important one to explain the school performance. However, the data indicates that, not only

[a]eduardo.ceriani@gmail.com

the dependence of the educational institution but also the geographical location where it is located affect the school performance. This research aims to determine which are the best predictors in the SIMCE score of mathematics in educational institutions of the Araucania region. The nested data will be modelling by using a Generalized Linear Mixed Model

***Keywords*:** Generalized linear mixed model, nested data, academic performance.

1. Introducción

El rendimiento escolar o académico es un fenómeno multidimensional que está determinado por el estudiante, el establecimiento educacional y el contexto socioeconómico cultural en que estos dos agentes se desenvuelven (Aguirre, Castro & Adasme, 2009; Torrecilla & Javier, 2008). En Chile, con el objeto de medir el rendimiento escolar de los estudiantes y ver si éstos han alcanzado los objetivos fundamentales y los contenidos mínimos obligatorios del currículo nacional, el Ministerio de Educación aplica una batería de pruebas estandarizadas que componen el Sistema de Medición de la Calidad de la Educación, SIMCE. Esta medición incluye una batería de pruebas estandarizas en dos subsectores de aprendizaje: lenguaje y matemática.

Diversos autores (Aguirre *et al.*, 2009; Cerón & Lara, 2010 y Vergara, 2009), sostienen que los antecedentes socioeconómicos familiares son los más importantes para explicar el rendimiento escolar en todos sus niveles para dichas evaluaciones. Estudios realizados en diversos países en latinoamérica (Torrecilla & Carrasco, 2011), apoyan que el contexto familiar, sus características socioeconómicas y el nivel de educación de los padres , determinan en gran medida el rendimiento académico. En particular, se ha demostrado que el nivel educacional materno tiene mayor repercusión que el paterno en el rendimiento de los estudiantes, poseer libros es más relevante que tener un computador; y el efecto de la conexión a internet es semejante a la existencia de un computador en el hogar (Aguirre *et al.*, 2009; Cerón & Lara 2010 y Vergara, 2009). En consecuencia, los peores resultados lo registran los estudiantes que provienen de estratos socioeconómicos bajos. Cabe destacar que estos factores (nivel socioeconómico de la familia, dependencia escolar y calidad del profesorado) tendrían mayor influencia en el rendimiento a medida que el nivel socioeconómico familiar del estudiante sea menor (Cerón & Lara, 2010).

En la actualidad, Chile presenta un sistema educacional que incorpora establecimientos municipales, particulares pagados y particulares subvencionados (Tokman, 2002). Los establecimientos municipales son de propiedad pública, administrados por las municipalidades (comunas), que reciben financiamiento estatal a través de subvención, entre los que se cuentan las corporaciones municipales y las direcciones de administración de educación municipal (DAEM). Por otro parte, los establecimiento particulares subvencionados son financiados parcial o totalmente en parte por el estado y finalmente, los establecimientos particulares pagados son financiadas integramente por el cobro de matrícula y/o aportes privados. Como

consecuencia, este sistema ha determinado una alta segregación de estudiantes a partir de su condición socioeconómica (García-Huidobro, 2007). La Tabla 1 muestra el Índice de Vulnerabilidad Escolar (IVE) en la región de la Araucanía para los Establecimientos Educacionales (EE) según dependencia y zona geográfica en la que están ubicados. Es claro que los EE municipales (públicos) concentran la mayor cantidad de estudiantes de características socioeconómicas vulnerables[1]. A su vez, es posible clasificar a los EE según su localidad geográfica, como urbano o rural. Actualmente, cerca del 90 % de los estudiantes están matriculados en un EE urbano (a nivel nacional) y en las últimas décadas se ha observado una tendencia progresiva a disminuir el número de estudiantes en EE rurales. Con base a los antecedentes citados, el presente estudio tiene por objetivo determinar cuáles son los mejores predictores del puntaje SIMCE en el año 2016 de matemáticas para estudiantes de cuarto básico (primaria) en la región de la Araucanía. Los datos serán modelados a través de un Modelo Lineal Generalizado Mixto (GLMM, por sus siglas en inglés).

Dependencia	IVE
EE público	90.45 %
EE particular subvencionado	87.65 %
Zona geográfica	IVE
Urbana	80.06 %
Rural	94.29 %

Tabla 1: IVE para los EE según dependencia y zona geográfica para el año 2016.

2. Metodología

Esta sección describe una extensión de los modelos lineales llamada Modelo Lineales Generalizados (GLM), los cuales permiten utilizar distribuciones pertenecientes a la familia exponencial, por ejemplo, Normal, Binomial, Poisson, Gamma, entre otras (Agresti, 2015); y además una extensión de los GLM denominados Modelos Lineales Generalizados Mixtos (GLMM). Estos últimos, serían utilizados para modelar los datos asociados a educación propuestos para la presente investigación. En primer lugar, la literatura (Agresti, 2015 y Jiang, 2007) describe los modelos lineales clásicos generales de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde $\boldsymbol{\epsilon}_{(n\times 1)}$ es un vector de errores aleatorios del modelo independientemente e idénticamente distribuidos $N(0, \sigma^2)$, $\mathbf{X}_{n\times p}$ es la matriz de diseño, $\boldsymbol{\beta}_{p\times 1}$ es un vector de efectos fijos, finalmente $\mathbf{Y}_{n\times 1}$ es el vector de respuesta, donde se cumple que $\mathbb{E}(Y_i) = \mu_i = \mathbf{X}_\mathbf{i}^\top \boldsymbol{\beta}$ e $Y_i \sim N(\mu_i, \sigma^2)$. Con el tiempo, este modelo se ha extendido considerando otras distribuciones para la variable respuesta Y. Además, la relación entre ésta y las variables explicativas no es necesariamente de la forma lineal descrita en el modelo anterior (Agresti, 2015). Este modelo es definido en términos de un conjunto

[1]Los datos son proporcionados por la Junta Nacional de Auxilio Escolar y Becas (JUNAEB) para el año 2016.

de variables independientes $Y_1, ..., Y_n$, donde la distribución de cada una de estas variables pertenece a la familia exponencial y posee las siguientes propiedades :

1. La función de densidad de cada Y_i depende del parámetro θ, y además, puede ser representado mediante la forma:

$$f(y_i; \theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \tag{1}$$

   donde a, b, c y d son funciones conocidas.

2. Las variables explicatorias proporcionan un conjunto de predictores lineales dados por la siguiente expresión.

$$\eta_i = \mathbf{X_i}^\top \boldsymbol{\beta} \tag{2}$$

donde $\boldsymbol{\beta}_{p\times 1}$ es un vector de efectos fijos y $\mathbf{X_i}^\top{}_{n\times p}$ es la matriz de diseño

3. El enlace entre 1. y 2. está dado por

$$g(\mu_i) = \eta_i = \mathbf{X_i}^\top \boldsymbol{\beta} \tag{3}$$

donde g es una función monótona y diferenciable llamada función de enlace, tal que $\mathbb{E}(Y_i) = \mu_i$. En los GLM, todos los parámetros presentados $\boldsymbol{\beta}$ se consideran regresores fijos y además, las observaciones son independientes entre si. No obstante, en ocasiones las observaciones están correlacionadas; como ocurre por ejemplo, en los estudios con datos anidados, donde usualmente las unidades están naturalmente agrupadas, por ejemplo EE pertenecientes a una misma comuna, de manera que los datos puedan presentar variabilidad correlacionada y no constante (Jian, 2007). Una solución a este problemática es considerar una extensión a los GLM que permiten asumir aleatoriedad en algunos de estos regresores, este modelo se denomina Modelo Lineal Generalizado Mixto (GLMM) (Jiang, 2007). Un GLMM, considera un vector de efectos aleatorios $\boldsymbol{\gamma}$ y la respuesta $y_1, ..., y_n$, donde la distribución de y_i es condicional al vector de efectos aleatorios y además es un miembro de la familia exponencial con distribución de densidad:

$$f_i(y_i|\gamma) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\} \tag{4}$$

donde $b(\cdot)$, $a_i(\cdot)$, $c_i(\cdot, \cdot)$ son funciones conocidas, ϕ es el parámetro de escala que puede ser o no conocido y el valor de θ_i que representa el parámetro natural, que asocia la media condicional de las observaciones con el predictor lineal presentado en la ecuación 3.

Entre las propiedades de la distribución exponencial, se destaca que la esperanza de las observaciones condicionales tiene la forma,

$$\mathbb{E}(y_i|\gamma) = \mu_i = b'(\theta_i) \tag{5}$$

y su función de varianza se define como,

$$V(\mu_i) = \frac{\delta \mu_i}{\delta \theta_i} = b''(\theta_i) \tag{6}$$

Los GLMM incorporan efectos fijos y aleatorios en las variables regresoras (Jiang, 2007). Estas se relacionan con las variables explicativas a través de un predictor lineal y una función de enlace de la siguiente forma:

$$\eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \tag{7}$$

donde $\mathbf{X}$ y $\mathbf{Z}$ son las matrices de diseño del modelo de dimensión $n \times p$ y $n \times r$, respectivamente, donde n corresponde al número de observaciones, p al número de parámetros y r al número de efectos aleatorios del modelo. Además, $\boldsymbol{\beta}$ es el vector de efectos fijos de dimensión $p \times 1$ y $\boldsymbol{\gamma}$ es el vector de efectos aleatorios de dimensión $r \times 1$ y posee una distribución $N(0, \phi^2)$. Finalmente η es un predictor lineal que se relaciona con la variable respuesta a través de una función de enlace $g(\cdot)$ y cumple la siguiente relación:

$$g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = g^{-1}(\eta) \tag{8}$$

La inversa de la función de enlace $g^{-1}(\cdot)$ es usada para modelar la relación entre el predictor lineal y la media condicional observada μ. Así, el modelo se presenta de la siguiente manera:

$$\mathbb{E}(y_i|\gamma) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = g^{-1}(\eta) \tag{9}$$

La Tabla 2 muestra las funciones de enlace y sus inversas, correspondientes para algunas distribuciones. El tipo de función de enlace $g(\cdot)$ y la correspondiente familia de distribución es elegida en base si la respuesta es binaria, discreta o continua (Agresti, 2015).

Distribuciones	Función de enlace	Función inversa
Normal	Identidad	η
Binomial	Logit/probit	$\frac{e^{\eta}}{1+e^{\eta}}$ / $\Phi(\eta)$
Poisson	Log	e^{η}
Gamma	Inversa/log	$\frac{1}{\eta}$ / e^{η}

Tabla 2: Inversa de la función de enlace

El proceso de estimación de los parámetros de un modelo estadístico es un paso clave en los análisis estadísticos. Tradicionalmente, para los GLM (que solo poseen

parámetros de efectos fijos) se puede realizar el proceso de estimación de parámetros mediante la construcción de la función de máxima verosimilitud (ML). No obstante, en el caso de los GLMM, la estimación de los parámetros corresponden a efectos fijos y efectos aleatorios, por ende, la expresión de ML no es sencilla y computacionalmente muy costosa (Jiang, 2007), por lo cual se han propuesto varias formas de aproximar la verosimilitud para estimar los parámetros para un GLMM, entre las cuales están: Método de Monte Carlo, Cadenas de Markov, Método MCMC, Métodos de Inferencia Aproximada, Aproximaciones de Laplace, Estimaciones de Cuasi-verosimilitud Penalizada y el Método de Cuadratura Gauss-Hermite (para más detalles acerca de estos métodos de estimación, consultar Jiang, 2007). En esta linea, el método por máxima verosimilitud restringida (REML, por sus siglas en inglés) es una alternativa a las estimaciones de ML, que estima los parámetros de efectos aleatorios promedio sobre los valores de los parámetros de efectos fijos. Al respecto, cabe destacar que las estimaciones de las desviaciones estándar de los efectos fijos utilizando REML son generalmente menos sesgadas que las estimaciones de ML (para más detalles acerca del REML, consultar Jiang (2007) y Pinheiro & Bates (2000). En la presente investigación se utiliza la aproximación REML para la estimación de los parámetros (fijos y aleatorios) en los modelos GLMM.

Por otro lado, existe una amplia gama de métodos para la selección de modelos estadísticos. Sin duda, los criterios de información de Akaike (AIC, por sus siglas en inglés) y de información de Schwarz (BIC, por sus siglas en inglés) son dos métodos ampliamente difundidos. El criterio de información de Akaike es una solución flexible, de enfoque de probabilidad basado en el modelo de selección (Akaike, 1973). El AIC posee la siguiente expresión $AIC = 2k - 2\ln(L)$ donde k es el número de parámetros en el modelo estadístico y L es el máximo valor de la función de verosimilitud para el modelo estimado. Por otra parte el criterio de información de Schwarz (1978) (BIC) presenta la siguiente expresión $BIC = -2\ln(L) + k\ln(n)$ donde n corresponde al tamaño muestral, k es el número de parámetros en el modelo estadístico y L es el máximo valor de la función de verosimilitud para el modelo estimado.

3. Aplicación

A continuación se presenta el proceso de aplicación de análisis y posterior modelación de los datos. Dicho proceso se haría a través del software *R* (R Development Core Team, 2015), librería `lme4` (Bates, Maechler, Bolker *et al.*, 2014). La base de datos que se utilizará para la aplicación de este estudio corresponde a los resultados de las pruebas SIMCE de matemáticas y lenguaje de estudiantes de cuarto básico del año 2016 agrupados en 32 comunas de la Región de la Araucanía (XIV). Del total observado para cada año solo 702 EE rindieron el SIMCE de lenguaje y matemáticas[2]. Por otra parte, de los 702 EE, 288 corresponde a dependencia

[2]Esta investigación utiliza como fuente de información las bases de datos de la Agencia de Calidad de la Educación. El autor agradece a la Agencia de Calidad de la Educación el acceso

pública y 414 a particular subvencionado, además, 431 a zonas geográficas rurales y 271 urbanas. En este punto es importante descatacar que la región de la Araucanía se caracteriza por un elevado número de áreas rurales, producto de la actividad económica propia de la zona, pero que no representa en este aspecto a la realidad nacional descrita anteriormente. Finalmente, las variables contenidas en la base de datos son:

1. Puntaje Matemáticas (PM): puntaje promedio en la prueba de matemáticas obtenidos por un establecimiento educacional (EE) en el año 2016.

2. Id: Factor indicando los establecimientos educacionales en los cuales forman parte de la base de datos. Los niveles son etiquetados desde 01 al 702.

3. Comuna: Factor indicando las comunas donde se encuentran los EE. Los niveles son etiquetados desde 01 al 32. Por tanto, los 702 EE están anidados en 32 grupos (comunas)

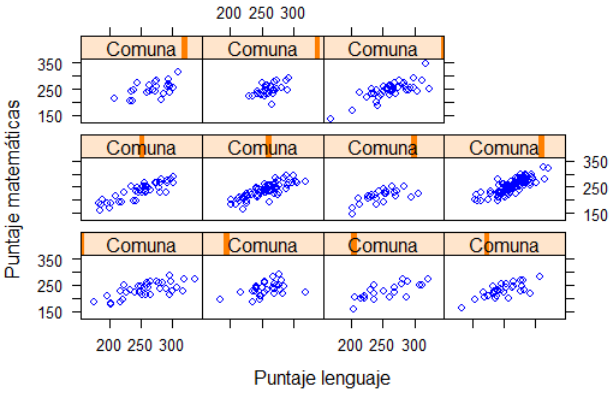
4. Puntaje Lenguaje (PL): puntaje promedio en la prueba de Lenguaje obtenidos por un EE en el año 2016.

5. Zona geográfica: Efecto del tipo de área en que se encuentra ubicado el EE, urbano o rural.

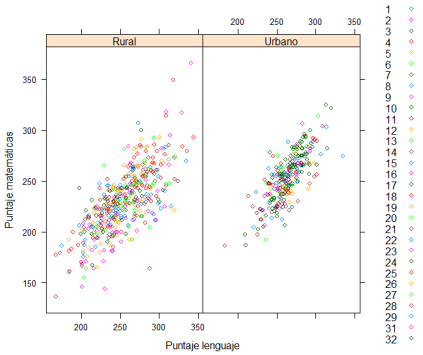
6. Dependencia: Efecto del tipo de administración, pública o subvencionada particular, que tiene cada EE.

En un análisis exploratorio de los datos para la variable respuesta PM (Figura 1). Se observa, que los EE públicos obtienen un puntaje promedio de 237.97 puntos, menor a los 239.82 puntos obtenidos por los EE subvencionados para el mismo año (1.85 puntos de diferencia). En cuanto al rendimiento en la PL, el promedio de las dependencias pública es de 257.39 puntos, menor a los 258.62 puntos para los subvencionados (1.23 puntos de diferencia). De igual forma el año 2016, la diferencia entre los urbano y rurales es de 13.47 puntos a favor de los primeros (253.47 y 230.00 puntos, respectivamente) en la PM, y de 9.22 puntos en la PL, también a favor de los establecimientos urbanos (263.78 y 254.55 puntos, respectivamente). En un análisis gráfico (Figura 1a) entre PM y PL de los EE anidados en las comunas se observa en rasgos generales un patrón lineal (por motivos gráficos solo se integran 11 de las 32 comunas en la Figura). Por otro lado, en la Figura 1b y 1c se pueden ver las curvas de los EE separados por zona geográfica (urbano y rural) y por dependencia (pública y subvencionada particular), en la cual también se observa una tendencia lineal entre los factores mencionados.

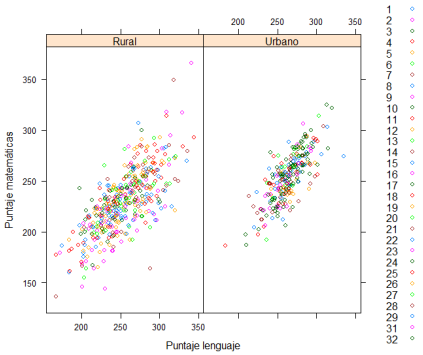
a la información. Todos los resultados del estudio son de responsabilidad del autor y en nada comprometen a dicha Institución.



(a) Gráfico exploratorio de PM y PL por comuna



(b) Gráfico exploratorio PM y PL separados por zona geográfica



(c) Gráfico exploratorio PM y PL separados por dependencia

Figura 1: Gráfico exploratorio para PM y PL

De acuerdo con el primer análisis exploratorio y basado en la posible relación lineal entre las variables, se propondría un modelo lineal general, no obstante, se debe considerar que los datos están agrupados en comunas (Figura 1a), por ende, un modelo lineal no es óptimo para modelarlos. Como se mencionó anteriormente la presente investigación modelará datos a partir de un modelo GLMM descrito en la ecuación 7, que como se describió son más adecuados para modelar este tipo de datos. Por lo tanto, la variable respuesta del modelo, Y_{ij} PM por el $i-ésimo$ EE de la $j-ésima$ comuna, donde $i = 1, ..., n$, $j = 1, ..., n_i$, posee una distribución que pertenece a la familia exponencial.

Por otra parte, se propone un predictor lineal η que se relaciona con la variable respuesta Y_{ij} a través de una función de enlace $g(\cdot)$ que corresponde a la función

identidad (Tabla 2), por tanto, se cumple que:

$$g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = g^{-1}(\eta) \tag{10}$$

donde $\mathbf{X}$ y $\mathbf{Z}$ son las matrices de diseño del modelo, $\boldsymbol{\beta}$ es el vector de efectos fijos y $\boldsymbol{\gamma}$ es el vector de efectos aleatorios. Si se establece como supuesto que la variable de respuesta es continua y se distribuye normal, entonces se asigna la variable $\boldsymbol{\epsilon_{ij}}$ como el error aleatorio no observable de distribución normal($N(0, \sigma^2)$). Luego, η se relaciona con la variable respuesta Y_{ij} a través de:

$$\mathbf{Y_{ij}} = \mathbf{X_{ij}}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon_{ij}} \tag{11}$$

Donde $\mathbf{Y_{ij}}$ es el PM en el $i - ésimo$ EE de la $j - ésima$, $\mathbf{X_{ij}}$ es la matriz que contiene las covariables del modelo, además, $\boldsymbol{\beta}$ es el vector de efectos fijos para los p parámetros del modelo tal que $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^\top$. Por otra parte, $\boldsymbol{\gamma}$ es el vector de efectos aleatorios que explica la variabilidad entre los grupos y se distribuye $N(0, \phi^2)$, $\boldsymbol{\epsilon_{ij}}$ es el vector para el error de la observación i de la $j-ésima$ comuna y se distribuye $N(0, \sigma^2)$. Además, el modelo asume que $\boldsymbol{\gamma}$ y $\boldsymbol{\epsilon}$ no están correlacionados entre sí.

Luego, para decidir la estructura del vector de efectos aleatorios $\boldsymbol{\gamma}$, se grafican los intervalos de confianza de las pendientes e interceptos (Figura 1a) obtenidos de un ajuste de regresión lineal simple por comuna, tomando PM como variable respuesta y PL como covariable, la ecuación de la regresión se muestra en la ecuación 12,

$$\mathbf{Y_{ij}} = \beta_0 + \beta_1 \boldsymbol{x_{ij}} + \boldsymbol{\epsilon_{ij}} \tag{12}$$

Donde $\boldsymbol{Y_{ij}}$ es el vector respuesta (PM), β_0 intercepto del modelo, β_1 es la pendiente del modelo para cada comuna y $\boldsymbol{x_{ij}}$ la matriz de diseño que contiene a la covariable PL de cada EE. Parte de los resultados del proceso de ajuste para cada comuna se muestran en la Tabla 3.

Comuna	Intercepto (β_0)	Pendiente (β_1)
01	99.58	0.51
02	-8.75	0.90
⋮	⋮	⋮
31	-35.65	1.09
32	72.72	0.63

Tabla 3: Estimación con la covariable PL

Una vez ajustado los modelos para cada comuna, se obtiene el gráfico de pendientes e interceptos (Figura 2), donde es posible observar que los coeficientes del modelo están correlacionados de forma negativa. Esto en el GLMM repercute en que la

matriz de varianza-covarianza de los efectos aleatorios no es diagonal (Agresti, 2015).

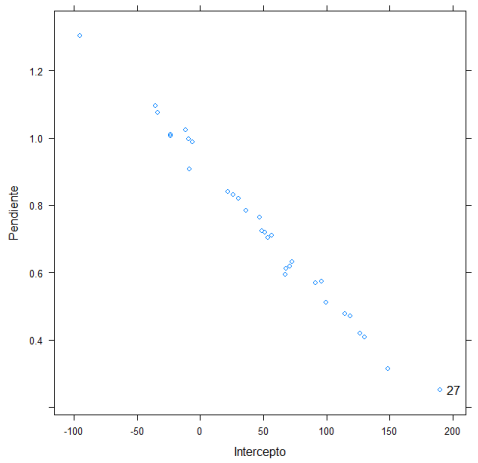


Figura 2: Coeficientes de la regresión simple, usando la covariable PL (no centralizada)

Para eliminar la correlación negativa se centralizan los valores de la covariable (PL) respeto a su media, y se vuelve ajustar el modelo definido en la ecuación 12. Para este efecto se calcula la media de PL, la cual tiene un valor de 258.11; luego la nueva covariable sería: PL-258.11. Los valores de los interceptos y pendientes del proceso de estimación de este nuevo modelo se encuentran en la Tabla 4.

Comuna	Intercepto (β_0)	Pendiente (β_1) (PL-258.11)
01	218.19	0.51
02	191.67	0.90
⋮	⋮	⋮
31	238.22	1.09
32	219.49	0.63

Tabla 4: Estimación con la covariable PL centralizada para cada comuna

Posteriormente, se construye un nuevo gráfico de interceptos y pendientes, obteniendo la Figura 3. Al modelar los efectos de la pendiente en cada comuna, en función de la covariable centrada, ahora se observan que los coeficientes no están correlacionados.

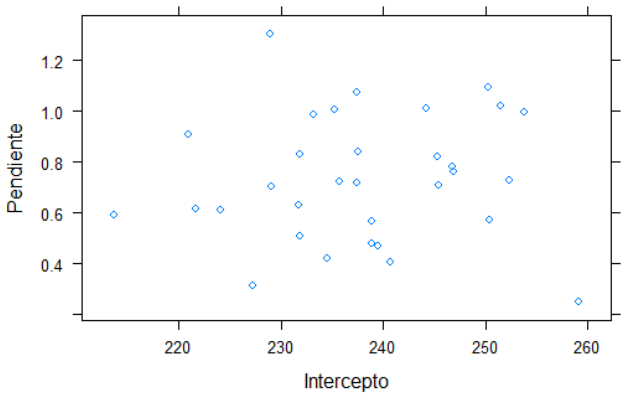


Figura 3: Coeficientes de la regresión simple, usando la covariable PL centralizada

Además, para decidir la estructura de los efectos aleatorios del modelo, ya sea, efectos aleatorios en el intercepto, en la pendiente o en ambos, se deben graficar los intervalos de confianza (IC) de las estimaciones en la Tabla 4 y ver cuál es el comportamiento de éstos. En relación a esto se obtiene la Figura 4. Luego, al observar la Figura 4, es posible notar que los IC de la pendiente (lado derecho de la figura) tienden a yuxtaponerse unos con otros, en contraste con los IC del intercepto (lado izquierdo de la figura) donde se observa mayor variabilidad en las rectas, por tanto, los IC de la Figura 4 entregan una indicación que un efecto aleatorio es necesaria para explicar la variabilidad comuna a comuna en el intercepto, no así un efecto aleatorio en la pendiente del modelo.

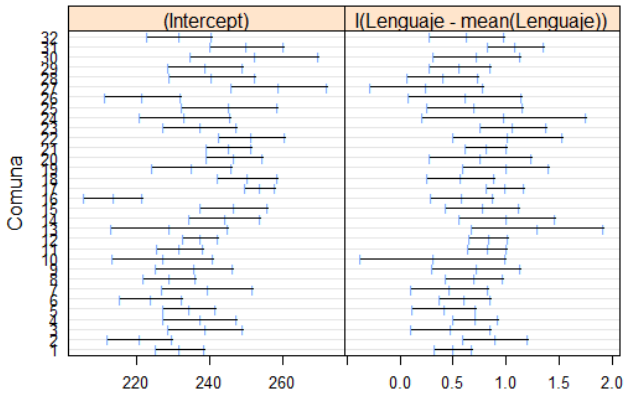


Figura 4: Gráfico de intervalos de confianza de los coeficientes paara el modelo de variable centralizada

Con base a esto se propone un modelo con intercepto aleatorio y pendiente fija. Además, como se ha mencionado, la base de datos considera tres potenciales covariables para explicar el rendimiento en PM y que podrían ser integradas a la matriz de diseño $\mathbf{X_{ij}}$. Ellas son, PL (centralizado en su media), zona geográfica del EE (urbano o rurales) y dependencia (público o subvencionado), por ende, el vector de efectos fijos del modelo quedaría, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^\top$. Por otra parte, $\boldsymbol{\gamma} = [u_1, ..., u_{32},]^\top$ es el vector de efectos aleatorios que explica la variabilidad entre los grupos (las desviaciones respecto al intercepto de la comuna) y $\boldsymbol{\epsilon} = [\epsilon_{11}, ..., \epsilon_{ij}]^\top$ representa la variabilidad entre los EE dentro de las comunas. Finalmente, y con base a lo anteriormente descrito, el modelo puede ser expresado mediante la siguiente ecuación:

$$\mathbf{Y_{ij}} = \beta_0 + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \beta_2 \boldsymbol{z_{ij}} + \beta_3 \boldsymbol{L_{ij}} + \boldsymbol{u_j} + \boldsymbol{\epsilon_{ij}} \tag{13}$$

Donde, β_0 es el intercepto fijo, β_1 es el parámetro para la covariable PL centrada en su media, $\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}$, β_2 es el parámetro para la covariable zona geográfica, $\boldsymbol{z_{ij}}$, que toma el valor 0 para EE urbanos y 1 para EE rurales. Por otra parte, β_3 es el parámetro para la covariable dependencia $\boldsymbol{L_{ij}}$ que toma el valor 0 cuando el EE sea de dependencia pública y 1 para subvencionado particular, finalmente, $\boldsymbol{u_j}$ es el intercepto aleatorio del modelo

Luego, considerando el número de covariables consideradas en la base de datos, se plantean 3 modelos usando como base el propuesto en la ecuación 13, donde cada uno será ajustado mediante el método REML. El primer modelo (Modelo (1)) para la predicción del PM, corresponde a un modelo que incorpora el intercepto del modelo β_0, el PL centralizada en su media $(x_{ij} - \bar{x})$ como covariable y el efecto aleatorio en las comunas u_j. El modelo se expresa en la ecuación 14 y los resultados del proceso de estimación de los efectos fijos se muestran a continuación en la Tabla 5.

$$Y_{ij} = \beta_0 + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \boldsymbol{u_j} + \boldsymbol{\epsilon_{ij}} \tag{14}$$

	Estimación	Error Estándar	t-valor	p-valor
Intercepto $\hat{\beta}_0$	238.08	1.31	181.11	0.00
(PSL−258.11) $\hat{\beta}_1$	0.75	0.02	28.52	0.00

Tabla 5: Estimación de efectos fijos Modelo (1)

Como se observa en la Tabla 5 las estimaciones son significativos, por ende, la covariable PL centralizada es significativa para el modelo. Luego, los componentes de varianza estimados son: $\sigma_u = 5.58$ y $\sigma_\epsilon = 20.27$, para el intercepto aleatorio y los residuos, respectivamente. Por otra parte, la Figura 5 analiza la distribución de residuos del modelo (1) a través de un gráfico de cajas con el objeto de observar el supuesto sobre los errores y ver si están centrados en cero y tienen varianza

constante a través de los niveles del grupo. En esta Figura se observa que si bien los errores están efectivamente centrados en 0, la variabilidad cambia entre comunas, sugiriendo que una nueva covariable debe integrada en el modelo.

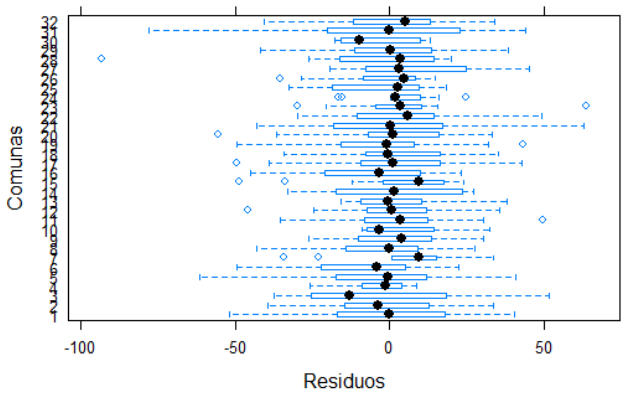


Figura 5: Gráfico de caja para los residuos por comuna del modelo (1)

Con base a lo anterior, el Modelo (2) incorpora la covariable zona geográfica, $\boldsymbol{z_{ij}}$, que es de carácter dicotómico que toma el valor 1 cuando el EE está en la zona urbana y 0 cuando éste pertenece a la zona rural y se expresa en la ecuación 15.

$$Y_{ij} = \beta_0 + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \beta_2 \boldsymbol{z_{ij}} + \boldsymbol{u_j} + \boldsymbol{\epsilon}_{ij} \tag{15}$$

	Estimación	Error Estándar	t-valor	p-valor
Intercepto $\hat{\beta}_0$	232.52	1.16	200.41	0.00
(PSL−258.3) $\hat{\beta}_1$	0.72	1.60	9.70	0.00
Zona $\hat{\beta}_2$	15.57	0.02	2.71	0.00

Tabla 6: Estimación de efectos fijos modelo (2)

Como se observa en la Tabla 6 las estimaciones son significativos, por ende, las covariable PL centralizada y zona geográfica son significativas para el modelo. Luego, la estimación de los componentes de varianza son: $\sigma_u = 3.43$ y $\sigma_\epsilon = 19.26$, para el intercepto aleatorio y los residuos, respectivamente. Además, para comprobar si existe variabilidad entre los EE públicos y subvencionados particulares se incluye esta variable a nuevo modelo, el Modelo (3); en este sentido, la variable dependencia es de carácter dicotómico, tomando valores 1 para EE subvencionados particulares y 0 para EE municipales, el modelo se expresa en la siguiente ecuación:

$$Y_{ij} = \beta_0 + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \beta_2 \boldsymbol{z_{ij}} + \beta_3 \boldsymbol{L_{ij}} + \boldsymbol{u_j} + \boldsymbol{\epsilon}_{ij} \tag{16}$$

	Estimación	Error Estándar	t-valor	p-valor
Intercepto $\hat{\beta}_0$	231.91	1.44	160.70	0.00
(PSL−258.3) $\hat{\beta}_1$	0.72	0.02	28.68	0.00
Zona $\hat{\beta}_2$	15.62	1.60	9.73	0.00
Dependencia $\hat{\beta}_3$	1.08	1.52	0.71	0.47

Tabla 7: Estimación de efectos fijos modelo (3)

Como se observa en la Tabla 7, las estimaciones $\hat{\beta}_1$ y $\hat{\beta}_2$ son estadísticamente significativas, no obstante, se observa que la estimación $\hat{\beta}_3$ no es significativa para el modelo. En consecuencia, esta covariable no será considerada para el modelo final, por tanto se mantendría el Modelo (2) que incorpora la covariable PL centralizada y zona geográfica. Volviendo a este modelo, para observar gráficamente la variabilidad de los residuos entre EE rurales y urbanos se obtiene la Figura 6 que representa los residuos estandarizados versus los valores ajustados del Modelo (2). Se puede observar diferencias entre la variabilidad entre EE urbanos y rurales, por ende, para corregir la heterocedasticidad del Modelo (2) se realizará una nueva estimación de los parámetros de este modelo, pero esta vez con una estructura de varianza constante.

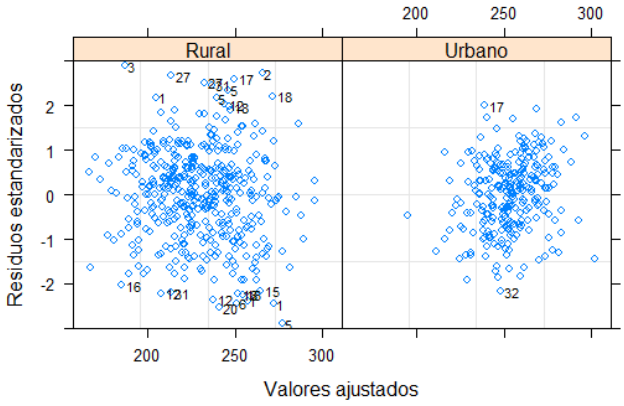


Figura 6: Gráfico de valores ajustado v/s residuos estandarizados del Modelo (4) separado por comunas

Por lo tanto, el modelo (4) tendría por tanto las mismas covariables que el modelo (2), pero con una estructura de varianza constante para los EE urbanos y rurales. Bajo esta concepción se estiman nuevamente los parámetros del Modelo (2), los

resultados se muestran en la Tabla 8. Se observa que al igual que para el Modelo (2) las estimaciones son significativas, por ende, la covariable PL centralizada y zona son significativas para el Modelo (4). Luego, las estimaciones de los componentes de varianza son: $\sigma_u = 3.38$ y $\sigma_\epsilon = 15.38$, para el intercepto aleatorio y los residuos, respectivamente.

	Estimación	Error Estándar	t-valor	p-valor
Intercepto $\hat{\beta}_0$	232.62	1.23	188.08	0.00
(PSL−258.3) $\hat{\beta}_1$	0.74	0.02	8.73	0.00
Zona $\hat{\beta}_2$	15.35	1.49	29.04	0.00

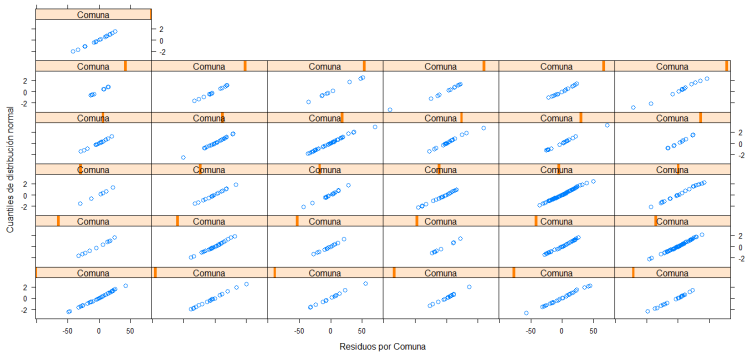
Tabla 8: Estimación de efecto fijos Modelo (4)

Hasta el momento se han obtenido 3 potenciales modelos con los cuales se podría modelar los datos: (1), (2) y (4). La Tabla 9 muestra un cuadro resumen para los criterios de selección Akaike (AIC) y Criterio de información de Schwarz (BIC). En ella es posible apreciar que a pesar de que el Modelo (4) tiene similares valores para los parámetros estimados del Modelo (2), el primero posee un menor valor tanto en AIC como en BIC, lo que justifica el haber realizado el proceso de estimación del modelo pero corrigiendo la heterocedasticidad. Por tanto, basado en los criterios AIC y BIC el modelo óptimo para modelar los datos correspondientes al PM en estudiantes de cuarto año básico en la región de la Araucanía es el Modelo (4).

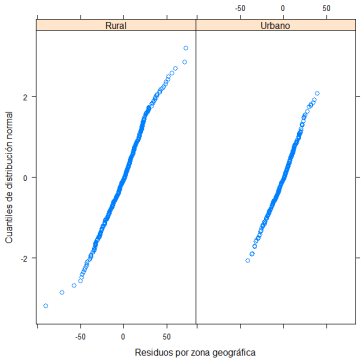
Modelo	AIC	BIC
Modelo (1)	6254.946	6273.15
Modelo (2)	6169.42	6192.169
Modelo (4)	6139.531	6166.829

Tabla 9: Valores de AIC y BIC para modelos (1), (2) y (4).

Para finalizar el análisis del modelo, resta realizar el proceso de diagnóstico de los supuestos para el Modelo (4). Los supuestos del modelo corresponden a que el efecto aleatorio entre los grupos (comunas) u_j se distribuye $N(0, \phi^2)$, además ϵ_{ij} que corresponde al error para la $i-ésima$ observación de la $j-ésimo$ comuna, se distribuye $N(0, \sigma^2)$), y finalmente, los efectos aleatorios u_j y ϵ_{ij} no están correlacionados entre si, por lo tanto, la $cov(u_j, \epsilon_{ij}) = 0$. Al respecto, en la Figura 7a se puede apreciar que los residuos dentro de los grupos no se apartan del supuesto de normalidad, pues éstos mantienen la tendencia lineal. También, al observar la Figura 7b, se observa que es razonable suponer normalidad de los residuos por zona geográfica, al observarse el mismo patrón anterior.



(a) Residuos por comuna v/s cuantiles de distribución normal



(b) Residuos por zona geográfica v/s cuantiles de distribución normal

Figura 7: Residuos dentro de las comunas y zona geográfica

Por otra parte, la Figura 8 muestra un gráfico para los efectos aleatorios comparados con cuantiles de la distribución normal, donde es posible observar una tendencia lineal, dicho patrón se corrobora mediante el test de normalidad de Shapiro-Wilk, con $W = 0.97981$ y $p - value = 0.8075$, por tanto, no se rechaza la hipótesis de normalidad para los efectos aleatorios en las comunas.

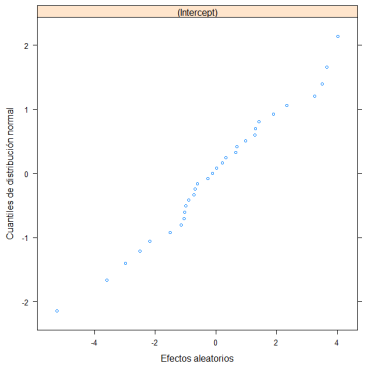
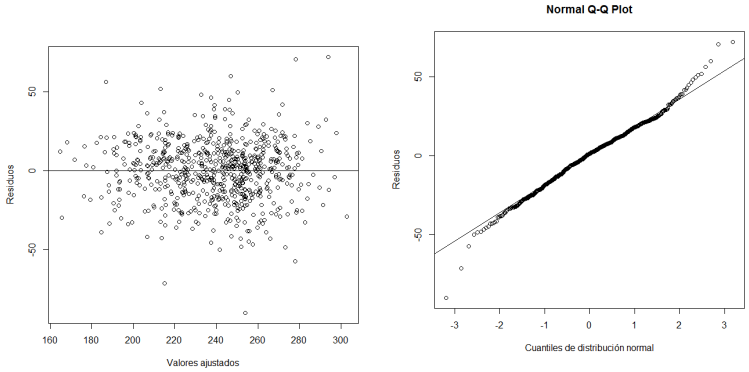


Figura 8: Efectos aleatorios v/s cuantiles de distribución normal

Finalmente, la Figura 9a muestra que no se observa tendencia alguna entre los residuos y los valores ajustados del modelo (4), además la nube de puntos se centra en torno al cero con solo unos pocos valores atípicos que se alejan de esta tendencia, por otra parte, la Figura 9b contrasta los residuos del modelo y los cuantiles de la distribución normal, observándose un clara tendencia lineal entre ambos.



(a) Residuos v/s valores de ajustados del modelo (4)

(b) Distribución de los residuos del modelo (4) v/s cuantiles de distribución normal

Figura 9: Diagnóstico del modelo

Una vez ya seleccionado el modelo y realizado el proceso de diagnóstico de éste, se realizará un proceso de predicciones sobre el modelo, para detectar diferencia entre los diferentes grupos (comunas) y niveles de un factor (diferencia entre EE de zonas urbanas y rurales en diferentes comunas). En primer lugar, la Tabla 10 muestra el valor para los interceptos aleatorios de las diferentes comunas.

Comuna	Efectos aleatorios por comuna u_j
Comuna 1	-2.44
Comuna 2	-3.51
Comuna 3	-0.10
⋮	⋮
Comuna 31	3.42
Comuna 32	-2.09

Tabla 10: Coeficientes aleatorios para las comunas del modelo (4)

Con dichos valores es posible obtener una predicción del PM para un EE de una comuna en particular. Al respecto, se analizarán dos casos. Cabe destacar que el modelo final (Modelo (4)) quedó de la siguiente manera y los resultados del proceso de estimación están en la Tabla 11:

$$Y_{ij} = \beta_0 + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \beta_2 \boldsymbol{z_{ij}} + \boldsymbol{u_j} + \boldsymbol{\epsilon}_{ij} \tag{17}$$

	Estimación
Intercepto $\hat{\beta}_0$	232.62
(PSL−258.11) $\hat{\beta}_1$	0.74
Zona $\hat{\beta}_2$	15.35

Tabla 11: Estimaciï¿½n de efecto fijos Modelo (4)

Donde $(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}})$ corresponde a la covariable para el PL centralizada en su media, $\boldsymbol{z_{ij}}$ en la covariable para la zona geográfica del EE, toma el valor 1 para establecimientos urbanos y 0 para rurales, finalmente, $\boldsymbol{u_j}$ corresponde a la variabilidad entre las comunas. Luego, esto puede re-ordenarse de la siguiente manera:

$$Y_{ij} = (\beta_0 + \boldsymbol{u_j}) + \beta_1(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}) + \beta_2 \boldsymbol{z_{ij}} + \boldsymbol{\epsilon_{ij}} \tag{18}$$

- Caso 1. EE urbano de la comuna 01

Para el siguiente caso $j = 1$, la covariable $z_{ij} = 1$ y $\beta_0 + u_j = 232.62 + (-2.44)$ por tanto, la ecuación de regresión quedaría de la siguiente manera:

$$\hat{Y}_{i1} = (232.62 - 2.44) + 0.74 * (\boldsymbol{x_{ij}} - 258.11) + 15.57 * 1 \tag{19}$$

Para este caso, la comuna 01 está 2.44 punto en promedio por debajo de la media regional, por tanto el modelo castigará en 2.44 puntos el puntaje final, por otro lado, el modelo bonificará con 15.57 por ser un EE urbano. Finalmente, si el PL

de un EE urbano está sobre el promedio, el modelo bonificará en 0.74 puntos por cada punto extra que este EE tenga sobre el promedio regional, si está por debajo de esta media producirá el fenómeno opuesto, restándole 0.74 puntos por cada uno bajo la media.

- Caso 2. EE rural de la comuna 31

Para el caso 2, $j = 31$, la covariable $z_{ij} = 0$ y $\beta_0 + u_j = 232.62 + 3.42$ por tanto, la ecuación de regresión quedará de la siguiente manera:

$$\hat{Y}_{i31} = (232.62 + 3.42) + 0.74 * (\boldsymbol{x_{ij}} - 258.11) + 15.57 * 0 \tag{20}$$

Para este caso, la comuna 31 está 3.42 punto en promedio por sobre la media regional, por tanto el modelo bonificará en 2.44 puntos el puntaje final, por otro lado, el modelo no bonificará con 15.57 por ser un EE rural. Luego, el efecto del PL será similar al caso anterior. Finalmente, se estimará el coeficiente de correlación intraclase (ICC), la que representa la magnitud de asociación entre dos EE de una misma comuna (Agresti, 2015). Puesto que las estimaciones de los componentes de varianza son: $\sigma_u = 3.38$ y $\sigma_\epsilon = 15.38$, para el intercepto aleatorio y los residuos, respectivamente, el ICC estimado entre dos mediciones de un mismo EE es:

$$\frac{\hat{\sigma_u^2}}{\hat{\sigma_u^2} + \hat{\sigma_e^2}} = \frac{3.38^2}{3.38^2 + 15.38^2} = 0.046 \tag{21}$$

4. Conclusiones

En este trabajo se ha presentado la aplicación un Modelo Lineal Generalizado Mixto (GLMM) para datos anidados. En particular esta técnica integra efectos aleatorios a los ya conocidos Modelos lineales clásicos y Modelos Lineales Generalizados (GLM), que permiten modelar más eficientemente la naturaleza de los datos agrupados. Esta investigación modelo datos asociados al rendimiento escolar de establecimientos educacionales (EE) donde estos últimos están naturalmente agrupados en comunas. Al respecto, se plantean 4 modelos usando las GLMM que integran diferentes covariables y estructuras de varianza. Los resultados del proceso de estimación de los modelos planteados (1), (2), (3) y (4), arrojan a este último como el óptimo para el proceso de modelado de datos que tiene por objetivo determinar cuales son los mejores predictores del puntaje SIMCE en el año 2016 de matemáticas (PM) para estudiantes de cuarto básico en la región de la Araucanía.

En particular, este modelo incorpora un efecto aleatorio para las comunas y efectos fijos para las covariables, PL (Puntaje SIMCE en la prueba de Lenguaje) centrado en su media y la zona geográfica en la que se encuentra el EE (Establecimiento educacional), además dentro de este factor, el modelo incorpora una estructura de

varianza igual para ambos grupos con el fin de corregir la heterocedasticidad. Al respecto el modelo, muestra que todos los efectos mencionados son significativos para el modelo, donde, la estimación para PL tiene un valor de 0.74, para el caso de efecto fijo zona geográfica, la estimación tiene un valor de 15.35; cabe destacar que ésta es una variable dicotómica, que toma valores 0 para los EE urbanos y 1 para los EE rurales, por ende, se espera que un EE urbano tenga 15.35 puntos por sobre su par rural. Finalmente, mediante el uso del Criterio de Akaike (AIC) y Schwarz (BIC) se llegó a la conclusión ya mencionada.

Un interesante resultado se obtuvo, al considerar en el modelo la covariable dependencia, que buscaba diferencia entre EE públicos y subvencionados particulares, en ese sentido, se determina que esta covariable no es significativa para el modelo, por ende, el modelo concluye que no hay diferencias significativas en los rendimientos en PM entre estos dos tipos de EE. Cabe recordar, que estos resultados son válidos solamente para EE que forman parte de la muestra para el año 2016, y se necesitan futuros estudios para obtener mayor información respecto a estos datos. También, es importante destacar, que según lo apreciado en la Tabla 1 los EE de carácter público y rurales, tienen mayor índice de vulnerabilidad escolar que aquellos EE de dependencia subvencionado particular y urbanos; esto a nivel nacional. Es conformidad de los resultados, es razonable suponer que si existe una diferencia en el índice de vulnerabilidad escolar (IVE) entre los EE urbanos y rurales para los EE de la Araucanía, y además, cabría investigar si existen diferencias en el IVE entre EE públicos y urbanos significativas en la Región de la Araucanía, y si estas diferencias (si es que existen) están incidiendo en el rendimiento de PM. En este sentido, sería interesante observar en futuras investigaciones ajustar un modelo similar con la covariable IVE.

Recibido: Agosto de 2019
Aceptado: Febrero de 2020

Referencias

Agresti, A. (2015), *Foundations of linear and generalized linear models*, John Wiley & Sons.

Aguirre, M., Castro, M. & Adasme, A. (2009), 'Factores que inciden en el rendimiento escolar en chile', *Estudios de Economía Regional, Talca, CEOC: Universidad de Talca* .

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Bates, D., Maechler, M., Bolker, B., Walker, S. et al. (2014), 'lme4: Linear mixed-effects models using eigen and s4', *R package version* **1**(7), 1–23.

CERÓN, F. & LARA, M. (2011), 'Factores asociados con el rendimiento escolar', *Santiago de Chile: MINEDUC/SIMCE* .

García-Huidobro, J. E. (2007), 'Desigualdad educativa y segmentación del sistema escolar. consideraciones a partir del caso chileno', *Revista pensamiento educativo* **40**(1), 65–85.

Jiang, J. (2007), *Linear and generalized linear mixed models and their applications*, Springer Science & Business Media.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & Team, R. C. (2007), 'Linear and nonlinear mixed effects models', *R package version* **3**(57), 1–89.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Team, R. C. (2015), 'R: A language and environment for statistical computing. r foundation for statistical computing. 2014'.

Tokman, A. (2002), Is private education better? evidence from chile, Technical report, Central Bank of Chile.

Torrecilla, F. J. M. & Carrasco, M. R. (2011), '¿ la escuela o la cuna? evidencias sobre su aportación al rendimiento de los estudiantes de américa latina. estudio multinivel sobre la estimación de los efectos escolares', *Profesorado. Revista de Curriculum y Formación de Profesorado* **15**(3), 27–50.

Torrecilla, M. & Javier, F. (2008), 'Los modelos multinivel como herramienta para la investigación educativa'.

Vergara, V. (2009), 'Redes sociales y efecto de los pares como predictores del rendimiento escolar en alumnos de cuarto año básico de la comuna de concepción', *REXE: Revista de estudios y experiencias en educación* **8**(16), 39–50.