PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

ANÁLISIS DE INFLUENCIA BAJO INFERENCIA BAYESIANA EN EVALUACIONES ESCOLARES DE ALTAS CONSECUENCIAS

Tesis para optar el grado de Magíster en Estadística

AUTOR

Andrés Guillermo Christiansen Trujillo

ASESOR

Dr. Cristian Luis Bayes Rodríguez

JURADO

Dr. Luis Hilmar Valdivieso Serrano
Dr. Víctor Giancarlo Sal y Rosas Celi

LIMA - PERÚ
2018

Dedicatoria

Al Dr. Fortunato Trujillo Ramírez, abuelo y maestro.

A mi abuela Teresa, la incombustible "Ita", por su amor incondicional y por encontrar la felicidad en las pequeñas cosas.

A mis padres, Cecilia y Hernán, que me enseñaron a querer, a respetar a los demás, a defender mis derechos y a pensar por mí mismo.

A Rodrigo Christiansen, que redefine lo superlativo. Como hermano, amigo y persona, él es del más alto tipo, calidad y orden; supremo.

Agradecimientos

Al Dr. Cristian Bayes con absoluta admiración y respeto. Su constante guía y paciencia han hecho posible esta investigación.

A Paula Cruzado, por transformarlo todo.

A Rogger Anaya, por estar en las buenas y en las malas.

A mis amigos Sadith Ramos y Wilmer Hernández con los que pasé incontables aventuras en esta maestría. Sin ustedes esto sería imposible.

A Giovanna Moreano, por haberme dado su cariño, respeto y la oportunidad de mi vida.

A todos mis compañeros de la UMC, especialmente a Pedro, Milagros, Manuel, Maricris, Yuriko, Judith, Mapu, José Carlos y Liliana.

A otros amigos, que sabiéndolo o no, ayudaron en la culminación de este proyecto: Carolina, Marcio, Brian, Alberto, Ady, Renato, Evelyn, Patuska, Chaska, Karina, Alejandra y Alejandro.

Y a todos los que me olvido.

Resumen

La presente investigación estudia una metodología para la detección de observaciones atípicas mediante un análisis de influencia bajo la perspectiva de la inferencia bayesiana. Se utiliza la medida de ϕ-divergencia y el estimador de Monte Carlo, derivado de ésta, trabajados previamente por Peng y Dey (1995), para el cálculo de las divergencias Kullback-Leibler, distancia rectilínea y ji-cuadrado. Además, en el presente trabajo se busca realizar este análisis de influencia en evaluaciones de altas consecuencias (evaluaciones cuyos resultados tienen un alto impacto en la vida de los estudiantes o docentes). El estudio de simulación revela que es posible recuperar observaciones previamente distorsionadas como atípicas. Finalmente, se aplica la metodología a una evaluación realizada por el Ministerio de Educación. Esta aplicación revela que la metodología estudiada es capaz de identificar escuelas con resultados no esperados dadas sus condiciones y resultados anteriores.

Palabras-clave: análisis de influencia, divergencia, evaluaciones de altas consecuencias, beta inflacionada.

Abstract

The present research studies a methodology for the detection of atypical observations using influential analysis under the perspective of Bayesian inference. The measure of ϕ-divergence and the Monte Carlo estimator, derived from it, previously worked by Peng y Dey (1995), is used to calculate the Kullback-Leibler, rectilinear distance and chi-squared divergences. In addition, this paper seeks to perform this influence analysis in high consequence evaluations (evaluations whose results have a high impact on the lives of students or teachers). The simulation study reveals that it is possible to recover previously distorted observations as atypical. Finally, the methodology is applied to an evaluation carried out by the Ministry of Education. This application reveals that the methodology studied is able to identify schools with unexpected results given their previous conditions and results.

Keywords: influential analysis, divergence, high-stake testing,beta inflated.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En el año 2014 el Ministerio de Educación fue autorizado a crear el Bono de Incentivo al Desempeño Escolar (BDE) como reconocimiento a la mejora en el aprendizaje de los estudiantes de las instituciones públicas (Decreto de urgencia N 002-2014). Esta medida ha generado la entrega de incentivos monetarios al personal docente y administrativo de las instituciones educativas que han obtenido mejores resultados y mayor crecimiento en las pruebas de Lectura y Matemática de la Evaluación Censal de Estudiantes (ECE) en 2do grado de primaria, 4to grado de primaria y 2do grado de secundaria (Decreto supremo N° 287-2014-EF; Decreto supremo N° 203-2015-EF).

La disposición de incentivos puede traer una serie de consecuencias negativas. Como señala Whitley (1998) un porcentaje importante de escolares y universitarios tienden a intentar hacer trampa en las pruebas que realizan, sobre todo si existen consecuencias asociadas a estas (Cizek & Wollack, 2017). En un contexto de incentivos monetarios asociados al desempeño, esta práctica puede institucionalizarse llevando a directivos y docentes a participar en el engaño (Jacob & Levitt, 2003).

Por ello, es de vital importancia determinar qué escuelas podrían estar incurriendo en estas prácticas. En el caso peruano, dado que la ECE evalúa un millón y medio de estudiantes en más de 30 000 escuelas en el Perú (Ministerio de Educación, 2017; Ministerio de Educación, 2017), una aproximación válida sería aplicar métodos de detección de valores atípicos (Kim, Woo & Dickison, 2017; Skorupski, Fitzpatrick & Egan, 2017) que permitan identificar escuelas con resultados extraños. Esto, a su vez, podría usarse como insumo para un análisis más detallado de los patrones de respuesta de los estudiantes de estas escuelas.

1.1. Organización del trabajo

En este trabajo se propone la utilización de un análisis de influencia bajo la perspectiva de la inferencia bayesiana que permita identificar escuelas con resultados atípicos en un año dados sus resultados en años anteriores.

En el capítulo 2 se hará una breve exposición de la medida de ϕ-divergencia que servirá para determinar la influencia de las observaciones. Este capítulo concluirá con la presentación de un estimador de Monte Carlo (Peng & Dey, 1995) que permita calcular de forma sencilla diferentes tipos de divergencias.

En el capítulo 3 se expondrá el modelo de regresión beta inflacionado. En este trabajo se utilizará el rendimiento previo de la escuela para crear un modelo explicativo de la proporción de estudiantes en bajos y altos niveles de logro. Es por ello que se propone utilizar el modelo

beta inflacionado, que permite asignar probabilidades a los valores entre 0 y 1, así como a estos extremos.

En el capítulo 4 se harán una aplicación en datos simulados que consideren diferentes tipos de datos atípicos. Luego, en el capítulo 5 se aplicará esta técnica a una evaluación realizadas por el Ministerio de Educación. Finalmente en el Anexo se incluyen los códigos utilizados para la estimación de modelo beta inflacionado y el cálculo de las divergencias.

Capítulo 2

Análisis de influencia bajo inferencia bayesiana

En la presente investigación se busca realizar un análisis de influencia que permita la detección de casos atípicos, en especial, resultados no esperado de escuelas en pruebas de rendimiento. Para ello, se propone la utilización de medidas de ϕ-divergencia que recojan la discrepancia entre dos distribuciones *a posteriori* ocasionadas por la inclusión de un caso con alguna perturbación. Por ello, en el presente capítulo se desarrollan estas medidas de divergencia y se concluye en un estimador de Monte Carlo para su cálculo.

2.1. Medidas de divergencia y ϕ-divergencia

En estadística se hace necesario saber elaborar una medida que permita determinar cuánto dos distribuciones se parecen entre sí. Como señalan Ali y Silvey (1966) estas medidas han sido llamadas de separación, de información discriminatoria y de variación de la distancia. Además, estos coeficientes comparten la propiedad de incrementarse cuando dos distribuciones son menos parecidas entre sí. Para Kullback (1978) esta medida debe ser llamada de *divergencia*.

Las medidas de divergencia son funciones similares a la distancia. Sin embargo, no cumplen con los axiomas de simetría y desigualdad triangular (Amari & Nagaoka, 2000). De forma general, se define la medida de divergencia mediante una función continua D en un espacio S de todas las funciones de probabilidad con soporte común, como $D = D(\cdot,\cdot)\colon S \times S \to \mathbb{R}$ tal que para cualquier $P, Q \in S$:

$$D(P,Q) \geq 0, \qquad \text{y} \qquad D(P,Q) = 0 \quad \text{si y solo si} \quad P = Q.$$

Desde la teoría de la probabilidad se proponen las medidas de ϕ-divergencia de una distribución P con respecto a una distribución Q como (Csiszár & Shields, 2004):

$$D_\phi(P,Q) = \int \phi\left(\frac{Q(x)}{P(x)}\right) P(x)dx \tag{2.1}$$

Se pueden realizar diferentes selecciones de ϕ, lo que permite obtener distintas medidas de divergencia (Dey & Birmiwal, 1993; Peng & Dey, 1995). Por ejemplo:

- $\phi(x) = -\log x$: define la divergencia de Kullback-Leibler,
- $\phi(x) = \frac{1}{2}|x-1|$: define la distancia rectilínea o norma ℓ_1,
- $\phi(x) = (x-1)^2$: define la divergencia ji-cuadrado,

entre otras.

Siguiendo (2.1) Peng y Dey (1995) definen una medida general de ϕ-divergencia para medir la discrepancia entre dos distribuciones *a posteriori*. Esto supone un modelo estadístico de regresión $\{f(y|\theta,x);\theta\in\Theta\}$ para un vector $y=(y_1,\ldots,y_n)$ donde θ es un vector p-dimensional de parámetros y x una matriz de covariables. Sea $\pi(\theta)$ una distribución *a priori*, entonces la distribución *a posteriori* en este modelo es dada por:

$$\pi(\theta|y)=\frac{\pi(\theta)f(y|\theta,x)}{f(y|x)}$$

donde $f(y|x)=\int\pi(\theta)f(y|\theta,x)d\theta$. Luego, definiendo una perturbación δ, se puede definir la distribución *a posteriori* perturbada como:

$$\pi_\delta(\theta|y)=\frac{\pi_\delta(\theta)f_\delta(y|\theta,x)}{f_\delta(y|x)}$$

donde $\pi_\delta(\theta)$ es la distribución *a priori* perturbada, $f_\delta(y|\theta,x)$ es la función de verosimilitud perturbada y $f_\delta(y|x)=\int\pi_\delta(\theta)f_\delta(y|\theta,x)d\theta$.

Peng y Dey (1995) recogen la definición (2.1) y definen:

$$D_\phi=D\left(\pi(\theta|y),\pi_\delta(\theta|y)\right)=\int\phi\left(\frac{\pi_\delta(\theta|y)}{\pi(\theta|y)}\right)\pi(\theta|y)d\theta, \tag{2.2}$$

donde δ indica una perturbación en la distribución *a posteriori*. De esta manera, se entiende a $\pi(\theta|y)$ como la distribución *a posteriori* obtenida con los datos originales y a $\pi_\delta(\theta|y)$ como la distribución *a posteriori* considerando una perturbación en los datos. Además, se asume ϕ como una función convexa con $\phi(1)=0$.

En el presente trabajo se utilizará un modelo de regresión bajo el enfoque bayesiano. Para ello Peng y Dey (1995) definen un factor de perturbación $\delta(\theta,y,x)$ que puede depender tanto de los parámetros contenidos en θ, de las covariables x como de la variable respuesta y. Este marco permite considerar una perturbación en la distribución *a priori* y en la función de verosimilitud de y. Se define como:

$$\delta(\theta,y,x)=\frac{f_\delta(y|\theta,x)\pi_\delta(\theta)}{f(y|\theta,x)\pi(\theta)} \tag{2.3}$$

donde $f_\delta(y|\theta,x)$ es la función de verosimilitud de y y π_δ la distribución *a priori* bajo algún tipo de perturbación. Además, $f(y|\theta,x)$ y $\pi(\theta)$ representan la verosimilitud y la distribución *a priori* sin considerar la perturbación, respectivamente. Debido a que la presente investigación se enfoca en la determinación de la existencia de una perturbación en los datos, se asume que la distribución *a priori* es constante. Por lo tanto, $\pi_\delta(\theta)=\pi(\theta)$, de esta manera, la ecuación (2.3) se simplifica en:

$$\delta(\theta,y,x)=\frac{f_\delta(y|\theta,x)}{f(y|\theta,x)}. \tag{2.4}$$

2.2. Medidas de influencia para la exclusión de un caso

La definición general de Peng y Dey (1995) puede ser utilizada para determinar la influencia que puede tener una observación en los parámetros a estimar. Es decir, puede usarse para medir el impacto que tiene la exclusión de esa observación sobre la estimación de la distribución *a*

posteriori. De esta manera, si la observación tiene un impacto desproporcionado, puede ser considerada como un dato influyente (Everitt & A., 2010; Rawlings, Sastry & Dickey, 2001).

Determinada la ecuación (2.2) y el factor de perturbación (2.4) se puede construir medidas que determinen la divergencia entre las distribuciones originales y las que consideren alguna perturbación. En el presente trabajo, se busca encontrar cómo la exclusión de algunas observaciones puede generar distribuciones *a posteriori* que presenten una divergencia significativa con respecto a la distribución sin considerar la perturbación.

Bajo esta perspectiva, sólo se considerará la perturbación ocasionada por la exclusión de un caso completo. La medida de la influencia de una observación y_r depende de la discrepancia entre $\pi(\theta|y)$ y $\pi(\theta|y_{(r)})$ que son las distribuciones *a posteriori* de θ dadas por y e $y_{(r)}$, donde $y_{(r)} = (y_1, \ldots, y_{r-1}, y_{r+1}, \ldots, y_n)$. Por lo tanto, se considera una perturbación en la verosimilitud pero no en la distribución *a priori*, que contemple la influencia de la remoción de un caso sobre el total de observaciones. Utilizando (2.4) se tiene que:

$$\delta(\theta, y, x) = \frac{f(y_{(r)}|\theta, x)}{f(y|\theta, x)}. \tag{2.5}$$

Es fácil notar que la expresión (2.5) es el inverso de la distribución de los datos de y_r dado $y_{(r)}$:

$$\delta(\theta, y, x) = \frac{f(y_{(r)}|\theta, x)}{f(y_{(r)}|\theta, x) f(y_r|\theta, x_r, y_{(r)})} = \frac{1}{f(y_r|\theta, x_r, y_{(r)})}. \tag{2.6}$$

En (2.6), el factor de perturbación δ toma en cuenta la exclusión del r-ésima observación. Este factor servirá para medir la divergencia de la distribución *a posteriori* sin la observación con respecto a la obtenida con los datos completos.

2.3. Estimación Monte Carlo de la medida de ϕ-divergencia

Peng y Dey (1995) proponen también un estimador Monte Carlo para la medida generalizada de divergencia propuesta. Para poder obtenerlo es necesario, en primer lugar, reemplazar la razón en (2.4). Utilizando la definición de la distribución *a posteriori* (Hoff, 2009) para la distribuciones con y sin perturbación de y, se tiene:

$$\frac{\pi_\delta(\theta|y)}{\pi(\theta|y)} = \frac{f_\delta(y|\theta, x)\pi_\delta(\theta)/f_\delta(y|x)}{f(y|\theta, x)\pi(\theta)/f(y|x)}. \tag{2.7}$$

Luego, considerando que ambas distribuciones a priori son iguales, es decir, $\pi(\theta) = \pi_\delta(\theta)$:

$$\frac{\pi_\delta(\theta|y)}{\pi(\theta|y)} = \frac{f_\delta(y|\theta, x)/f_\delta(y|x)}{f(y|\theta, x)/f(y|x)}. \tag{2.8}$$

Esta ecuación puede ponerse términos de la razón de las distribuciones *a posteriori* para poder ser reemplazada en (2.2). De esta manera:

$$\frac{\pi_\delta(\theta|y)}{\pi(\theta|y)} = \delta(\theta, y, x)\frac{f(y|x)}{f_\delta(y|x)} \tag{2.9}$$

Por lo tanto, combinando (2.2) y (2.9), se obtiene una nueva expresión para la forma generalizada de la ϕ-divergencia:

$$D_\phi = \int \phi\left(\delta(\theta, y, x)\frac{f(y|x)}{f_\delta(y|x)}\right)\pi(\theta|y)d\theta, \tag{2.10}$$

A su vez, la expresión $f(y|x)/f_\delta(y|x)$ puede ser simplificada utilizando la definición de la distribución marginal y de la distribución *a posteriori*. Así, en primer lugar, el término $f_\delta(y|x)$ puede ser expresado como la distribución marginal de $y|x$ proveniente de una distribución $y|\theta, x$:

$$f_\delta(y|x) = \int f_\delta(y|\theta, x)\pi(\theta)d\theta,$$

obteniéndose:

$$\frac{f(y|x)}{f_\delta(y|x)} = \frac{f(y|x)}{\int f_\delta(y|\theta, x)\pi(\theta)d\theta} \tag{2.11}$$

Además para poder luego obtener la forma de la distribución *a posteriori* de θ se multiplica el numerador de (2.11) por una constante igual a 1:

$$\frac{f_\delta(y|x)}{f(y|x)} = \frac{\int f_\delta(y|\theta, x)\frac{f(y|\theta,x)}{f(y|\theta,x)}\pi(\theta)d\theta}{f(y|x)}$$

En primer lugar, se puede reemplazar la expresión $f_\delta(y|\theta, x)/f(y|\theta, x)$ por $\delta(\theta, y, x)$ dada la ecuación (2.4), quedando:

$$\frac{f_\delta(y|x)}{f(y|x)} = \frac{\int \delta(\theta, y, x)f(y|\theta, x)\pi(\theta)d\theta}{f(y|x)} = \int \delta(\theta, y, x)\left(\frac{f(y|\theta, x)\pi(\theta)}{f(y|x)}\right)d\theta \tag{2.12}$$

La expresión $f(y|\theta, x)\pi(\theta)/f(y|x)$ es igual a la distribución *a posteriori* de $\theta|y$, por lo que (2.12) queda de la siguiente manera:

$$\frac{f(y|x)}{f_\delta(y|x)} = \frac{1}{\int \delta(\theta, y, x)\pi(\theta|y)d\theta}.$$

Este término puede ser reemplazado en (2.10), obteniéndose la expresión

$$D_\phi = \int \phi\left(\frac{\delta(\theta, y, x)}{\int \delta(\theta, y, x)\pi(\theta|y)d\theta}\right)\pi(\theta|y)d\theta \tag{2.13}$$

De esta manera, suponiendo que se tiene una cadena $\{\theta^S\}_{s=1}^B$ de simulaciones, donde B es el número total de muestras de la distribución *a posteriori* $\pi(\theta|y)$, se tendría un estimador Monte Carlo D_ϕ con la siguiente forma:

$$\hat{D}_\phi = \frac{1}{B}\sum_{s=1}^{B}\phi\left(\frac{\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}\right). \tag{2.14}$$

2.4. Medidas de divergencia

2.4.1. Divergencia de Kullback-Leibler

Como se señaló antes, se puede obtener la divergencia Kullback-Leibler (KL) reemplazando $\phi(x)$ por $-\log(x)$. Para el caso del estimador de Monte Carlo en (2.14):

$$\hat{D}_{KL} = -\frac{1}{B}\sum_{s=1}^{B}\log\left(\frac{\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}\right)$$

Debido a las propiedades del logaritmo la ecuación anterior es equivalente a:

$$\begin{aligned}\hat{D}_{KL} =& \frac{1}{B}\sum_{s=1}^{B}\log\left(\frac{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}{\delta(\theta^s)}\right)\\ =& \log\frac{\left(\prod_{s=1}^{B}B^{-1}\sum_{s=1}^{B}\delta(\theta^s)\right)^{1/B}}{\left(\prod_{s=1}^{B}\delta(\theta^s)\right)^{1/B}}\\ =& \log\frac{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}{\left(\prod_{s=1}^{B}\delta(\theta^s)\right)^{1/B}}\end{aligned}$$

Lo que resulta en:

$$\hat{D}_{KL} = \log\frac{B^{-1}\sum_{s=1}^{B}(1/f(y_r|\theta^s,x_r,y_{(r)}))}{\left(\prod_{s=1}^{B}(1/f(y_r|\theta^s,x_r,y_{(r)}))\right)^{1/B}}. \tag{2.15}$$

2.4.2. Distancia rectilínea o norma ℓ_1

Para el caso de la distancia rectilínea o norma ℓ_1, se reemplaza $\phi(x)$ por $|x-1|/2$ en el estimador de Monte Carlo (2.14):

$$\hat{D}_{\ell_1} = \frac{1}{B}\sum_{s=1}^{B}\frac{1}{2}\left|\frac{\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)} - 1\right|$$

lo cual puede reformularse como:

$$\begin{aligned}\hat{D}_{\ell_1} =& \frac{1}{2B}\sum_{s=1}^{B}\left|\frac{\delta(\theta^s) - B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}\right|\\ =& \frac{\sum_{s=1}^{B}\left|\delta(\theta^s) - B^{-1}\sum_{s=1}^{B}\delta(\theta^s)\right|}{2\sum_{s=1}^{B}\delta(\theta^s)}\end{aligned}$$

Utilizando la expresión (2.6):

$$\hat{D}_{\ell_1} = \frac{\sum_{s=1}^{B}\left|(1/f(y_r|\theta^s,x_r,y_{(r)})) - B^{-1}\sum_{s=1}^{B}(1/f(y_r|\theta^s,x_r,y_{(r)}))\right|}{2\sum_{s=1}^{B}(1/f(y_r|\theta^s,x_r,y_{(r)}))}. \tag{2.16}$$

2.4.3. Divergencia ji-cuadrado

En el caso de la divergencia ji-cuadrado, reemplazando $(x-1)^2$ en (2.14), se obtiene:

$$\hat{D}_{\chi^2} = \frac{1}{B}\sum_{s=1}^{B}\left(\frac{\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)} - 1\right)^2$$

Lo que resulta en:

$$\begin{aligned}
\hat{D}_{\chi^2} =& \frac{1}{B}\sum_{s=1}^{B}\left(\frac{(\delta(\theta^s))^2}{\left(B^{-1}\sum_{s=1}^{B}\delta(\theta^s)\right)^2} - \frac{2\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)} + 1\right)\\
\hat{D}_{\chi^2} =& \frac{B^{-1}\sum_{s=1}^{B}(\delta(\theta^s))^2}{\left(B^{-1}\sum_{s=1}^{B}\delta(\theta^s)\right)^2} - \frac{2B^{-1}\sum_{s=1}^{B}\delta(\theta^s)}{B^{-1}\sum_{s=1}^{B}\delta(\theta^s)} + 1\\
\hat{D}_{\chi^2} =& \frac{B^{-1}\sum_{s=1}^{B}(\delta(\theta^s))^2}{\left(B^{-1}\sum_{s=1}^{B}\delta(\theta^s)\right)^2} - 2 + 1.
\end{aligned}$$

Finalmente, considerando (2.6), se obtiene:

$$\hat{D}_{\chi^2} = \frac{B^{-1}\sum_{s=1}^{B}\left((1/f(y_r|\theta^s, x_r, y_{(r)}))\right)^2}{\left(B^{-1}\sum_{s=1}^{B}(1/f(y_r|\theta^s, x_r, y_{(r)}))\right)^2} - 1. \tag{2.17}$$

Capítulo 3

Medidas de influencia en el modelo beta inflacionado

En el presente trabajo se busca realizar un análisis de influencia en evaluaciones de altas consecuencias (evaluaciones cuyos resultados tienen un alto impacto en la vida de los estudiantes o docentes) mediante las medidas de divergencia expuestas en el capítulo anterior (Cizek & Wollack, 2017). Debido a que uno de los criterios más importantes para las escuelas es el porcentaje de sus estudiantes en niveles altos y bajos de desempeño, es necesario utilizar un modelo que permita predecir la proporción de estudiantes en cada nivel. Dentro de este marco, se usará el modelo beta inflacionado.

Ospina y Ferrari (2010) proponen el modelo beta inflacionado para cuando la variable aleatoria Y que toma valores dentro del intervalo $]0,1[$ así como los valores 0 y 1 con probabilidad positiva. Esta distribución es esencialmente una mixtura entre las distibuciones beta y Bernoulli. Bayes y Valdivieso (2016) proponen una nueva parametrización, para la cual utilizan la siguiente notación $y_i \sim \mathcal{B}Im(\alpha_0, \alpha_1, \gamma_i, \phi)$, donde la función de masa, la media y la varianza están dadas por:

$$f(y|\alpha_0,\alpha_1,\gamma,\phi) = \begin{cases} \alpha_0(1-\gamma), & \text{si } y = 0, \\ (1-\alpha_0(1-\gamma))b\left(y|\dfrac{\gamma(1-\alpha_1)}{1-\alpha_0(1-\gamma)-\alpha_1\gamma},\phi\right), & \text{si } y \in (0,1), \\ \alpha_1\gamma, & \text{si } y = 1, \end{cases}$$

$$\mathrm{E}(y) = \gamma,$$

$$\mathrm{Var}(y) = \frac{1+\alpha_0\phi}{1+\phi} + \left(\frac{(1-\alpha_1)^2\,\phi}{(1-\alpha_0(1-\gamma)-\alpha_{1\gamma})\,(1+\phi)}\right)\gamma^2,$$

siendo $y \in [0,1]$, $\alpha_0 \in [0,1]$, $\alpha_1 \in [0,1]$, $\gamma_i \in [0,1]$, $\phi \geq 0$ y $b(y|\mu,\phi)$ la función de densidad de una distribución beta bajo la parametrización:

$$b(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1. \tag{3.1}$$

3.1. Modelo de regresión

Bayes y Valdivieso (2016) proponen que la proporción y pueda seguir una distribución $\mathcal{B}Im(\alpha_0, \alpha_1, \gamma, \phi)$, y ajustan una serie de ecuaciones de regresión a la media de y y a los parámetros α_0 y α_1. Este modelo se denomina como modelo de regresión a la media beta inflacionada. Sean $y_1, \ldots, y_n$ observaciones tal es que $y_i \sim \mathcal{B}Im(\alpha_{0i}, \alpha_{1i}, \gamma_i, \phi)$:

$$g_1(\alpha_{0i}) = \tilde{x}_i^\top \omega,$$
$$g_2(\alpha_{1i}) = \breve{x}_i^\top \eta,$$
$$g_3(\gamma_i) = {x_i}^\top \beta,$$

donde $\omega = [\omega_0, \omega_1, \ldots, \omega_{k_1}]^\top$, $\eta = [\eta_0, \eta_1, \ldots, \eta k_2]^\top$ y $\beta = [\beta_0, \beta_1, \ldots, \beta k_3]^\top$ son parámetros de regresión, y $\tilde{x}_i = [1, \tilde{x}_{i1}, \ldots, \tilde{x}_{ik_1}]^\top$, $\breve{x}_i = [1, \breve{x}_{i1}, \ldots, \breve{x}_{ik_2}]^\top$ y $x_i = [1, x_{i1}, \ldots, x_{ik_3}]^\top$ son las columnas de variables independientes para el sujeto i. Como función enlace, proponen la función logística para g_1, g_2 y g_3:

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

La función de masa de $y_i \sim \mathcal{BI}m(\alpha_{0i}, \alpha_{1i}, \gamma_i, \phi)$ para el sujeto i queda entonces escrita como:

$$f(y_i|\alpha_{0i}, \alpha_{1i}, \gamma_i, \phi) = \begin{cases} \alpha_{0i}(1-\gamma_i), & \text{si } y_i = 0, \\ (1-\alpha_{0i}(1-\gamma_i))b\left(y|\dfrac{\gamma_i(1-\alpha_{1i})}{1-\alpha_{0i}(1-\gamma_i)-\alpha_{1i}\gamma_i}, \phi\right), & \text{si } y_i \in (0,1), \\ \alpha_{1i}\gamma_i, & \text{si } y_i = 1. \end{cases} \quad (3.2)$$

En el caso de la presente investigación se considerará que los parámetros α_0, α_1 son constantes, y se las covariables explicarán la media de γ mediante la función de enlace logística descrita previamente. De esta manera, se trabajará con el modelo:

$$Y_i \sim \mathcal{BI}m(\alpha_0, \alpha_1, \gamma_i, \phi), \quad (3.3)$$

donde

$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right) = {x_i}^\top \beta, \quad i = 1, \ldots, n. \quad (3.4)$$

3.2. Estimación

En Bayes y Valdivieso (2016) se propone que la estimación por máxima verosimilitud del vector de parámetros $\theta = [\alpha_0, \alpha_1, \phi, \beta]^\top$. Asumiendo que las observaciones se ordenan de tal manera que primero quedan los y_i iguales a 0, luego a 1 y finalmente los continuos. Además que n_0 es el número de casos con $y_i = 0$, n_1 los casos con $y_i = 1$ y n total de casos, se tiene que la función de verosimilitud:

$$K(\theta) = \prod_{i=1}^{n_0} \alpha_0 (1-\gamma_i) \prod_{i=n_0+1}^{m} \alpha_1 \gamma_i \prod_{i=m+1}^{n} (1-\alpha_0(1-\gamma_i))b\left(y|\frac{\gamma_i(1-\alpha_1)}{1-\alpha_0(1-\gamma_i)-\alpha_{1\gamma_i}}, \phi\right). \quad (3.5)$$

En este trabajo se adoptará para la estimación un enfoque bayesiano y se utilizarán las siguientes distribuciones *a priori* $\alpha_0 \sim U(0,1)$, $\alpha_1 \sim U(0,1)$ y $\phi \sim U(0,a)$ y $\beta_j \sim U(-b,b)$ independientes entre sí. De esta manera la distribución *a priori* de los datos es dada por:

$$P(\theta) \propto 1, \quad (3.6)$$

y la distribución *a posteriori* es dada por:

$$P(\theta|y) \propto K(\theta), \tag{3.7}$$

Cabe resaltar que esta última distribución no tiene forma conocida, por lo que se utilizarán cadenas de Markov de Monte Carlo, en concreto el algoritmo de Gibbs para generar simulaciones de la distribución *a posteriori* dada en (3.7). Este algoritmo fue implementado en el programa WinBUGS (Spiegelhalter, Thomas, Best & Lunn, 2003)

Sea $\{\theta^s = (\alpha_0^s, \alpha_1^s, \phi^s, \beta^s)^\top\}_{s=1}^B$ las B simulaciones obtenidas de la distribución *a posteriori*. Las divergencia de Kullback-Leibler, la distancia rectilínea y la divergencia ji-juadrado para cada observación son estimadas utilizando los estimadores de Monte Carlo dados en (2.15), (2.16) y (2.17), obteniendo las siguientes expresiones:

$$\begin{aligned}
\hat{D}_{KLr} =& \log \frac{\left(\prod_{s=1}^B B^{-1} \sum_{s=1}^B (1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))\right)^{1/B}}{\left(\prod_{s=1}^B (1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))\right)^{1/B}} \\
\hat{D}_{\ell_1 r} =& \frac{\sum_{s=1}^B \left|(1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s)) - B^{-1} \sum_{s=1}^B (1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))\right|}{2 \sum_{s=1}^B (1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))} \\
\hat{D}_{\chi^2 r} =& \frac{B^{-1} \sum_{s=1}^B \left((1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))\right)^2}{\left(B^{-1} \sum_{s=1}^B (1/f(y_r|\alpha_0^s, \alpha_1^s, \gamma_i^s, \phi^s))\right)^2} - 1,
\end{aligned}$$

donde γ_i^s es la esperanza de la observación i en θ^s, que es calculada usándose (3.4).

Capítulo 4

Aplicación en datos simulados

Para comprobar que la metodología propuesta logra reconocer datos atípicos se realizó una aplicación con datos simulados. Esta consistió en la generación de 100 casos bajo el modelo beta inflacionado. En ellas se generaron cuatro formas de datos atípicos y se calcularon tres tipos de divergencias para cada observación.

4.1. Aplicación en datos sin distorsiones

En primer lugar, se generaron 100 muestras de una covariable x mediante una distribución $N(0,1)$, la cual se relacionara con una variable respuesta y mediante el modelo expuesto en (3.2), considerando:

$$y_i \sim \mathcal{B}Im(0.1, 0.1, \gamma_i, 20)$$
$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right) = 0.6 + 1x_i$$

Luego, se estimaron estos parámetros bajo inferencia bayesiana mediante el paquete estadístico WinBUGS y el uso del muestreador de Gibbs. Se consideran para ello las siguientes distribuciones *a priori*: $\alpha_0 \sim U(0,1)$, $\alpha_1 \sim U(0,1)$, $\beta_0 \sim U(-5,5)$, $\beta_1 \sim U(-5,5)$ y $\phi \sim U(0,100)$.

Utilizando el algoritmo de Gibbs se simularon 2 cadenas de Markov de tamaño 3000 cada una, de estas se descartaron las primeras 1500 antes de obtener convergencia. Por lo tanto, se obtuvieron en total 3000 simulaciones de la distribución *a posteriori* con las que se realizarán las estimaciones de los parámetros y divergencias.

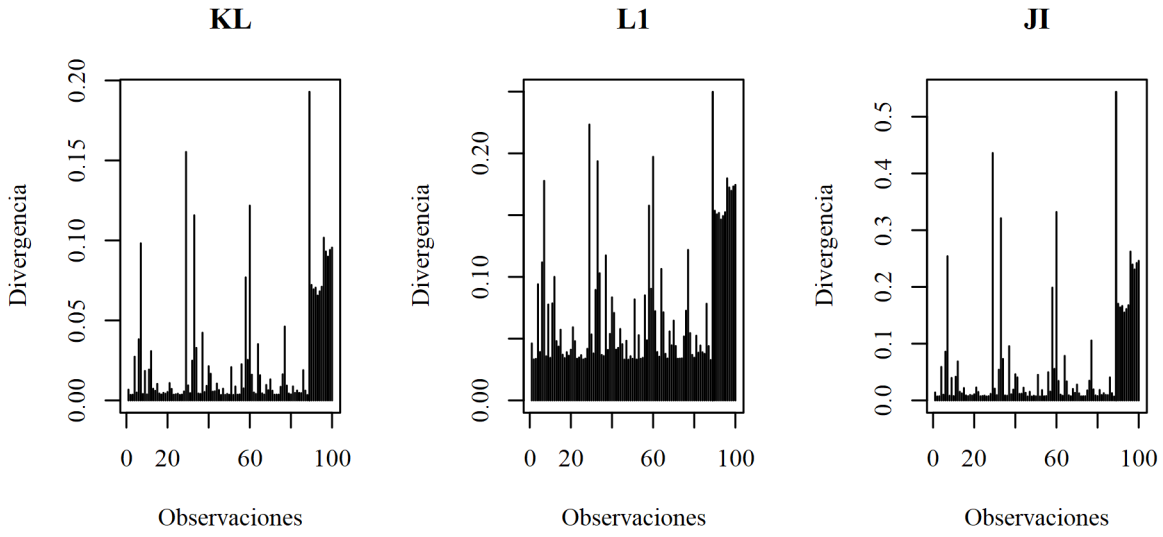
Los parámetros estimados se asemejan a los reales, como se aprecia en la Tabla 4.1. En la Tabla 4.1 se muestra la media de los parámetros estimados de las simulaciones 1501 hasta la 3000. Además, en las Figuras 7.1 y 7.2 (ver Anexo) se puede observar que los parámetros del modelo convergen en las iteraciones mostradas y presentan una baja autocorrelación.

Tabla 4.1: Medidas de tendencia central y percentiles de parámetros estimados en la aplicación en datos simulados sin considerar distorsiones

Parámetro	**Media**	**DE**	**P2.5**	**P25**	**P50**	**P75**	**P97.5**
α_0	0.16	0.05	0.07	0.13	0.16	0.20	0.28
α_1	0.10	0.04	0.04	0.07	0.09	0.12	0.18
β_0	0.46	0.09	0.28	0.41	0.47	0.52	0.64
β_1	0.91	0.06	0.79	0.86	0.91	0.95	1.03
ϕ	16.79	2.45	12.34	15.03	16.66	18.32	22.07

Una vez estimados los parámetros del modelo, se procedió a calcular las divergencias KL, ℓ_1 y ji-cuadrado para cada uno de los casos simulados dados los parámetros estimados (sección 2.4). Como se puede ver en la Figura 4.1, no existe ningún caso que se destaque de los demás. Por lo que se podría considerar la ausencia de datos atípicos.

Figura 4.1: Divergencias para la aplicación con datos simulados sin considerar distorsiones



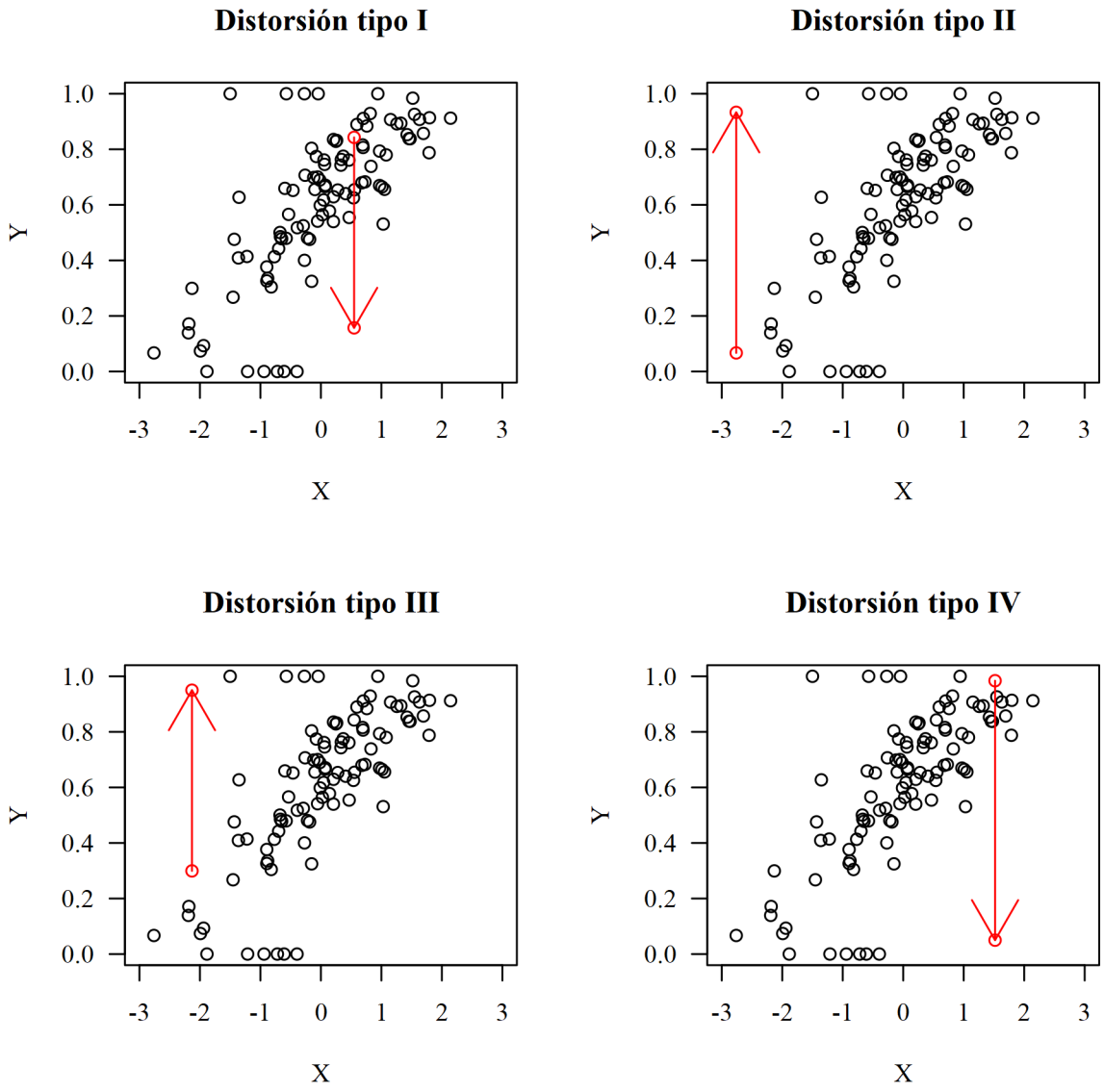
4.2. Aplicación con datos distorsionados

Dado que la presente investigación presenta y estudia una metodología para la detección de datos atípicos mediante el cálculo de divergencias en las distribuciones *a posteriori*, se proponen cuatro tipos de distorsiones que serán introducidas en las observaciones de la muestra:

- Tipo I: un caso con $0.75 < y < 0.90$ es convertido en su complemento, $1 - y$.
- Tipo II: un caso con $0.10 < y < 0.25$ es convertido en su complemento, $1 - y$.
- Tipo III: un caso con $0.10 < y < 0.25$ es convertido en 0.95.
- Tipo IV: un caso con $0.75 < y < 0.90$ es convertido en 0.05.

Estas distorsiones se presentan esquemáticamente en la Figura 4.2.

Figura 4.2: Esquematización de las distorsiones introducidas en los datos simulados



Se aplicó un análisis de regresión con los mismos supuestos mencionados previamente. En este se encuentran resultados similares; desde la iteración 1501 la estimación de los parámetros converge y estos presentan bajas autocorrelaciones. Como se aprecia en las Tablas 4.2, 4.3, 4.4 y 4.5 la estimación de los parámetros se ve afectada. Esto ocurre principalmente con el parámetro β_1. Como se puede observar la inclusión de un caso distorsionado reduce la media de las estimaciones de este parámetro, siendo esta reducción más importante en el caso de la distorsión tipo II donde es del 17.8 %. Además, el límite inferior del intervalo de credibilidad toma valores entre 0.594 y 0.732, significando una reducción entre 24.8 % y 7.3 %. Finalmente, el límite superior del intervalo de credibilidad varía entre 0.897 y 1.000, representado una reducción entre 12.9 % y 2.9 % del original. Es decir, la inclusión de casos distorsionados si afecta la estimación general de los parámetros.

Tabla 4.2: Medidas de tendencia central y percentiles de parámetros estimados en la aplicación en datos simulados considerando un distorsión de tipo I

Parámetro	**Media**	**DE**	**P2.5**	**P25**	**P50**	**P75**	**P97.5**
α_0	0.161	0.052	0.070	0.123	0.156	0.196	0.271
α_1	0.096	0.036	0.036	0.069	0.092	0.118	0.173
β_0	0.418	0.093	0.232	0.356	0.420	0.481	0.594
β_1	0.860	0.068	0.732	0.812	0.859	0.905	1.000
ϕ	12.966	1.846	9.665	11.660	12.850	14.190	16.870

Tabla 4.3: Medidas de tendencia central y percentiles de parámetros estimados en la aplicación en datos simulados considerando una distorsión de tipo II

Parámetro	**Media**	**DE**	**P2.5**	**P25**	**P50**	**P75**	**P97.5**
α_0	0.167	0.056	0.073	0.126	0.163	0.203	0.288
α_1	0.094	0.035	0.035	0.068	0.090	0.116	0.174
β_0	0.460	0.098	0.259	0.397	0.463	0.526	0.645
β_1	0.748	0.078	0.594	0.697	0.750	0.799	0.897
ϕ	9.890	1.416	7.323	8.895	9.815	10.830	12.800

Tabla 4.4: Medidas de tendencia central y percentiles de parámetros estimados en la aplicación en datos simulados considerando una distorsión de tipo III

Parámetro	**Media**	**DE**	**P2.5**	**P25**	**P50**	**P75**	**P97.5**
α_0	0.165	0.051	0.072	0.128	0.162	0.197	0.273
α_1	0.095	0.036	0.036	0.070	0.092	0.117	0.173
β_0	0.465	0.093	0.277	0.404	0.465	0.529	0.643
β_1	0.784	0.076	0.637	0.732	0.784	0.835	0.933
ϕ	10.059	1.473	7.414	8.999	9.974	11.020	13.171

Tabla 4.5: Medidas de tendencia central y percentiles de parámetros estimados en la aplicación en datos simulados considerando una distorsión de tipo IV

Parámetro	**Media**	**DE**	**P2.5**	**P25**	**P50**	**P75**	**P97.5**
α_0	0.160	0.054	0.070	0.122	0.156	0.194	0.278
α_1	0.100	0.038	0.040	0.072	0.094	0.124	0.186
β_0	0.394	0.095	0.191	0.334	0.396	0.457	0.574
β_1	0.778	0.074	0.640	0.727	0.777	0.825	0.923
ϕ	9.604	1.362	7.134	8.662	9.508	10.470	12.480

Se esperaría encontrar que los casos distorsionados presenten las mayores divergencias en los cuatro escenarios analizados. En la Tabla 4.6 se muestran los casos distorsionados y aquellos que presentaron mayores divergencias. Se puede observar que los que sufrieron las distorsiones son precisamente los que tienen mayores divergencias en todos los escenarios.

Tabla 4.6: Casos con la mayor divergencia en la aplicación en datos simulados. En todos los escenarios los casos distorsionados fueron los que presentaron mayores divergencias

Distorsión	**Caso distorsionado**	**Divergencias**		
		KL	ℓ_1	χ^2
Sin distorsión	-	89	89	89
Tipo I	75	75	75	75
Tipo II	41	41	41	41
Tipo III	12	12	12	12
Tipo IV	60	60	60	60

Por otro lado, como se muestra en las Figuras 4.3, 4.4, 4.5 y 4.6 los casos distorsionados con mayores divergencias se diferencian de forma importante de los demás, presentando divergencias mucho mayores a las encontrados en los casos sin distorsión.

Figura 4.3: Divergencias para la aplicación con datos simulados considerando una distorsión de tipo I

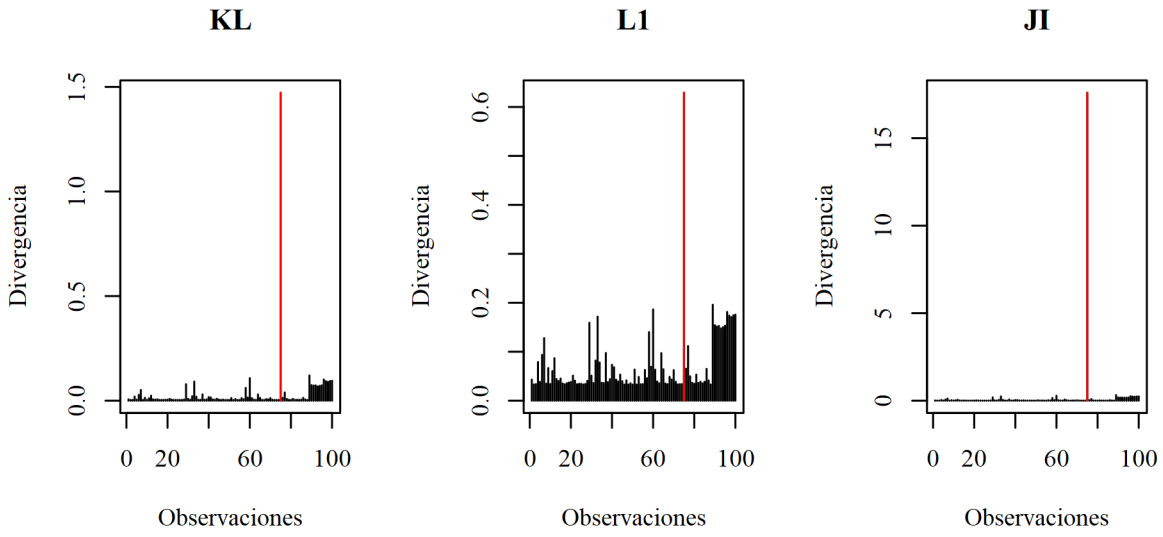


Figura 4.4: Divergencias para la aplicación con datos simulados considerando una distorsión de tipo II

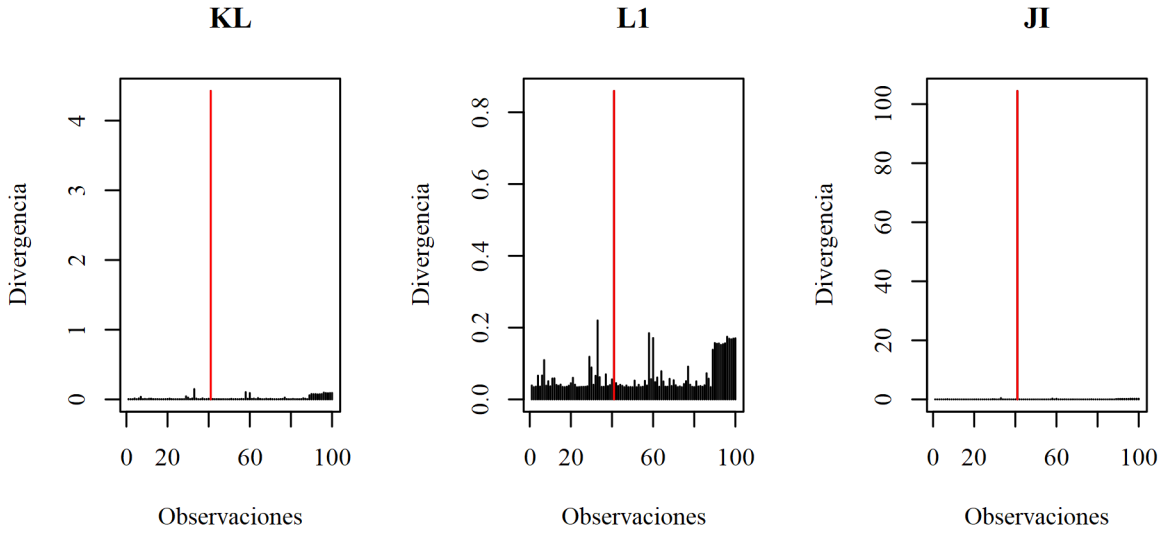


Figura 4.5: Divergencias para la aplicación con datos simulados considerando una distorsión de tipo III

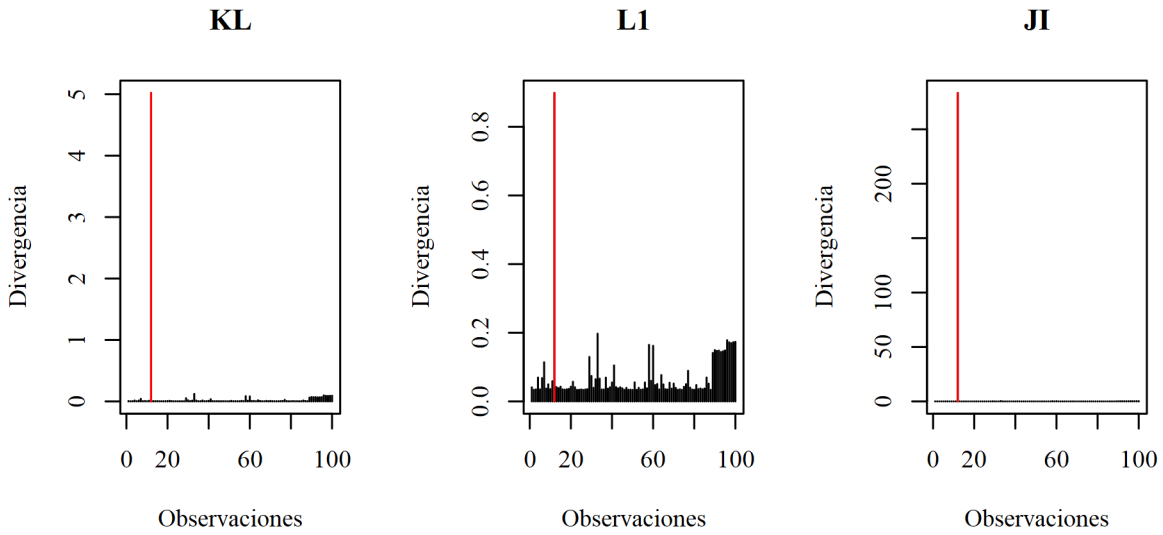
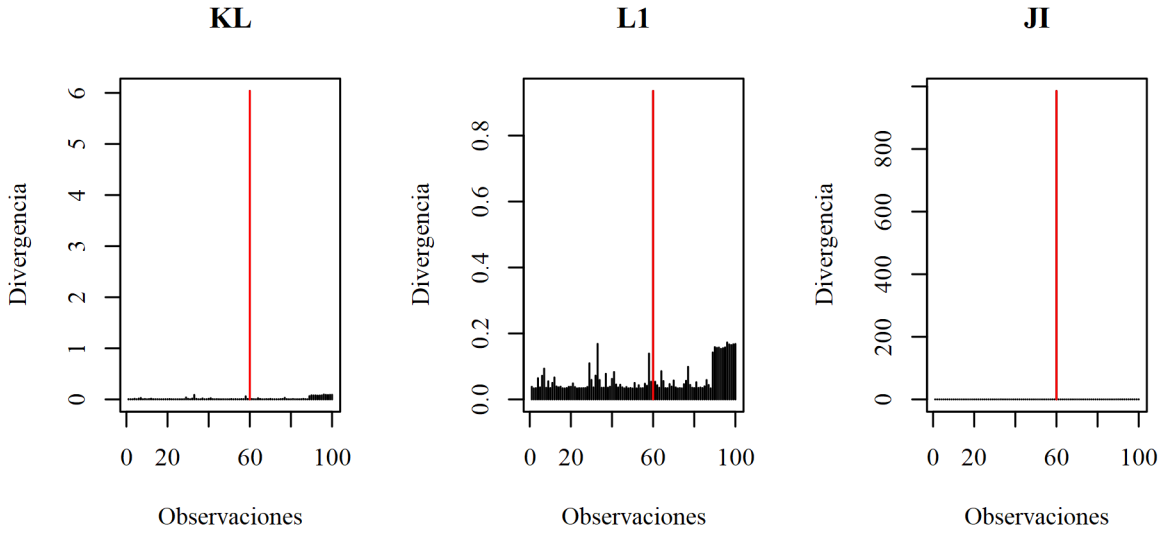


Figura 4.6: Divergencias para la aplicación con datos simulados considerando una distorsión de tipo IV



Estos resultados muestran que es posible que, dado este modelo, la metodología sea capaz de detectar un valor atípico introducido. Sin embargo, es importante comprobar esto mediante un estudio de simulación considerando distintos tamaños de muestra. A continuación se presenta dicho estudio utilizando los cuatro tipos de distorsión.

4.3. Estudio de simulación

Para comprobar lo encontrado previamente, se realizó una simulación que considerara los cuatro tipos de distorsión. Para ello se generaron 100 simulaciones con diferentes tamaños de muestra (100, 200, 500 y 1000). En cada una de ellas se aplicaron por separado los cuatro tipos de distorsiones propuestos. Luego, se realizó una estimación bayesiana utilizando los mismos parámetros iniciales que en la sección 4.1 pero con 750 iteraciones. Finalmente, se observó si el caso que presentaba mayores divergencias era el caso distorsionado.

Para todos los tamaños de muestra simulados y en los cuatro tipos de distorsión analizados se recuperó el caso atípico como aquél con mayor divergencia. Esto sucedió con los tres tipos de divergencias. Es decir, siempre el caso atípico fue identificado como el caso con mayores divergencias, sin importar el tamaño de la muestra, el tipo de distorsión o el tipo de divergencia.

En la Tabla 4.7 se muestra qué tanto las divergencias de los casos distorsionados se alejan de las divergencias promedio para cada tipo de distorsión y tamaño de muestra. Se muestra el número de desviaciones estándar en que los casos atípicos se alejan de la media, esto se reporta mediante r donde $D^\star = \overline{D} + rSD_D$, siendo $D^\star$ la divergencia del valor atípico, $\overline{D}$ la divergencia promedio y SD_D la desviación estándar de las divergencias. Además, se presenta el número de rangos intercuartílicos en que los casos atípicos de alejan del percentil 75, esto se reporta mediante q donde $D^\star = D_{75} + qRIC_D$, siendo $D^\star$ la divergencia del valor atípico, D_{75} la el valor de la divergencia en el percentil 75 y RIC el rango intercuartílico. Como se puede apreciar en esta Tabla los casos atípicos introducidos y detectados presentaron divergencias mucho mayores a aquellos sin distorsiones. De forma general, la introducción de un dato distorsionado genera

una divergencia cada vez más alejada de la media conforme crece el tamaño de muestra.

Tabla 4.7: Divergencias promedio de los casos atípicos. Número de desviaciones estándar en que los casos atípicos que se alejan de la media, esto se reporte mediante r. Número de rangos intercuartílicos en que los casos atípicos de alejan del percentil 75, esto se reporte mediante q

Distorsión	**n**	**Divergencias**			**r**			**q**		
		KL	ℓ_1	χ^2	KL	ℓ_1	χ^2	KL	ℓ_1	χ^2
Tipo I	100	6.52	0.91	139.44	9.89	8.9	9.9	797.03	31.03	8005.25
	200	6.34	0.91	152.73	14.06	12.56	14.07	1165.62	36.22	13418.8
	500	4.2	0.85	77.21	22.3	19.28	22.32	1327.88	41.22	12018.46
	1000	2.65	0.75	45.47	31.54	26.04	31.59	1340.34	44.51	11388.6
Tipo II	100	3.53	0.76	70.78	9.73	8.41	9.81	434.53	25.9	4083.4
	200	3.86	0.81	81.85	14.02	12.19	14.07	713.91	32.4	7382.15
	500	3.2	0.79	53.95	22.28	18.87	22.31	1018.09	38.31	8427.39
	1000	1.91	0.67	25.4	31.52	25.12	31.58	956.56	39.72	6181.08
Tipo III	100	5.78	0.89	124.84	9.88	8.88	9.9	702.51	30.62	7134.85
	200	5.35	0.88	119.44	14.06	12.48	14.07	985.55	35.28	10691.32
	500	3.74	0.82	66.49	22.29	19.13	22.31	1188.81	40.15	10403.34
	1000	2.19	0.7	32.71	31.54	25.59	31.59	1097.99	41.84	8012.51
Tipo IV	100	7.2	0.92	150.53	9.89	8.93	9.9	874.83	31.39	8590.38
	200	6.66	0.92	159.26	14.06	12.57	14.07	1221.27	36.48	13965.85
	500	4.28	0.85	79.12	22.3	19.3	22.32	1354.45	41.44	12317.99
	1000	2.67	0.75	46.16	31.55	26.07	31.59	1353.28	44.65	11555.11

Esto permite comprobar que la técnica propuesta es capaz de detectar valores atípicos mediante el cálculo de las divergencias en las distribuciones *a posteriori*. Por ello, a partir de esto se propone aplicarla en una evaluación nacional de altas consecuencias diseñada e implementada por el Ministerio de Educación.

Capítulo 5

Aplicación en una evaluación de altas consecuencias

Como se detalló previamente, las evaluaciones escolares de altas consecuencias son aquellas cuyos resultados derivan en algún tipo de sanción o premio para docentes, directores o estudiantes. En este tipo de evaluaciones es de suma importancia verificar que los puntajes obtenidos por estudiantes y escuelas sean lo más precisos posibles y evitar cualquier intento de trampa o copia (Cizek & Wollack, 2017).

Esto es aplicable al caso peruano y a las pruebas aplicadas por el Ministerio de Educación. Estas brindan dos tipos de resultados por estudiante: un puntaje continuo y una categoría del desarrollo de sus aprendizajes (Ministerio de Educación, 2017). Las instituciones educativas suelen establecer sus metas de desempeño alrededor del crecimiento del porcentaje de alumnos que se encuentran en altos niveles de logro y la reducción de los bajos niveles de logro. Para el caso analizado se cuenta con tres categorías o niveles de logro, de menor a mayor: En Inicio, En Proceso y Satisfactorio (Ministerio de Educación, 2017). Este tipo de caracterización es aplicable a la técnica mostrada previamente pues permite modelar la proporción de estudiantes en el nivel más bajo (En Inicio) y el nivel más alto (Satisfactorio) en Lectura. Además, se cuenta con evidencia de intentos de fraude por parte de colegios que intentan conseguir el Bono Escuela.

En el presente trabajo, esto se analizarán las divergencias que puedan encontrarse en un tipo de evaluación realizada por el Ministerio de Educación[1]. Se realizará un análisis de influencia en una evaluación que contó con 760 escuelas y alrededor de 17 000 estudiantes por año. Para estas escuelas se cuenta con sus rendimientos en 5 años consecutivos y en 5 estratos de su diseño muestral. Para esta evaluación, el Ministerio de Educación realizó previamente un análisis que identificó al estrato 2 como aquel con mejoras en su desempeño en magnitudes no esperadas.

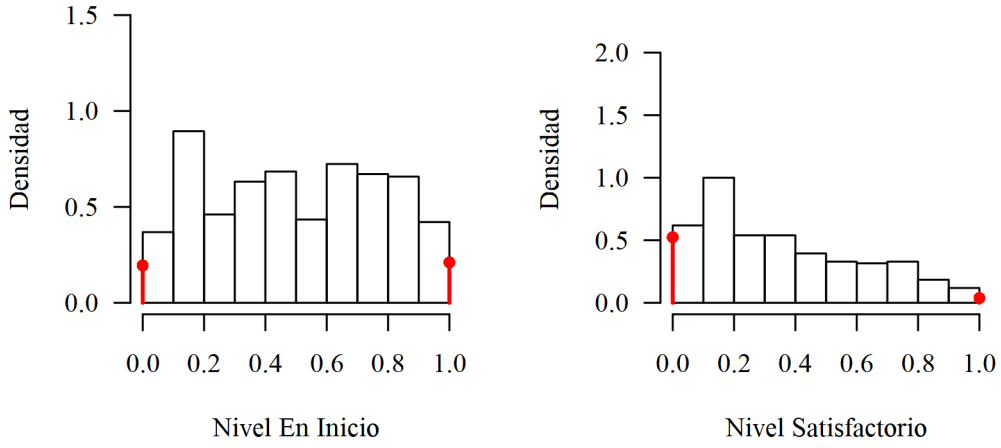
En este caso, se utilizará la proporción de estudiantes en los niveles En Inicio y Satisfactorio en el último año como variables respuesta. A su vez, el rendimiento en años previos como puntaje continuo ("medida promedio")[2] será considerado como predictor del desempeño del último año. Se utilizarán los resultados en la prueba de Lectura, omitiendo los resultados de la prueba de Matemática.

Es importante mencionar que en esta evaluación, son pocos los estudiantes que alcanzan el nivel Satisfactorio (Figura 5.1), por lo que la mayoría de progresiones año a año radica en el cambio del nivel En Inicio al nivel en Proceso.

[1] En el presente trabajo no se identifican a las evaluaciones ni a los estratos considerados por su nombre oficial para preservar la confidencialidad de los resultados de las escuelas.

[2] Esta medida se obtiene a través de un puntaje Rasch como se especifica en Ministerio de Educación (2017).

Figura 5.1: Histograma de niveles de logro



De otro lado, como se muestra en la Tabla 5.1 la mitad de las escuelas evaluadas tienen más del 50 % de estudiantes en el nivel de logro más bajo. Además, también se puede ver que la mitad de escuelas no tienen a ningún estudiante en el nivel de logro más alto. Con respecto a la medida continua de rendimiento en Lectura, se puede apreciar que si bien la media no muestra variación importante a lo largo de los años, el máximo rendimiento ha ido aumentando.

Tabla 5.1: Estadísticos descriptivos de variables utilizadas en el modelo de regresión

	Mínimo	**P25**	**P50**	**Media**	**P75**	**Máximo**	**DE**
Proporción nivel En Inicio	0.00	0.13	0.54	0.51	0.91	1.00	0.38
Proporción nivel Satisfactorio	0.00	0.00	0.00	0.20	0.33	1.00	0.30
Rendimiento año 1	-2.47	-0.95	-0.37	-0.21	0.34	3.10	0.96
Rendimiento año 2	-3.39	-0.90	-0.37	-0.21	0.37	3.34	0.98
Rendimiento año 3	-4.68	-0.88	-0.35	-0.25	0.33	4.00	0.92
Rendimiento año 4	-2.63	-0.85	-0.36	-0.22	0.32	4.37	0.90

Con el fin de determinar si es que la técnica expuesta podría detectar un comportamiento atípico en el estrato 2, se construyó un modelo de regresión beta inflacionada para los resultados de los niveles En Inicio y Satisfactorio de la prueba en el quinto año. Se consideraron como predictores los resultados del puntaje continuo de los primeros 4 años en Lectura. Como se puede apreciar en las Figuras 5.2 y 5.3 existe cierta relación entre el rendimiento de años anteriores y la proporción de estudiantes en los diferentes niveles de logro en el año 5. De forma general, rendimientos bajos en años anteriores se relacionan con altas proporciones de estudiantes en los niveles más bajos.

Figura 5.2: Diagrama de dispersión entre la proporción de estudiantes en el nivel En Inicio y el rendimiento en años anteriores

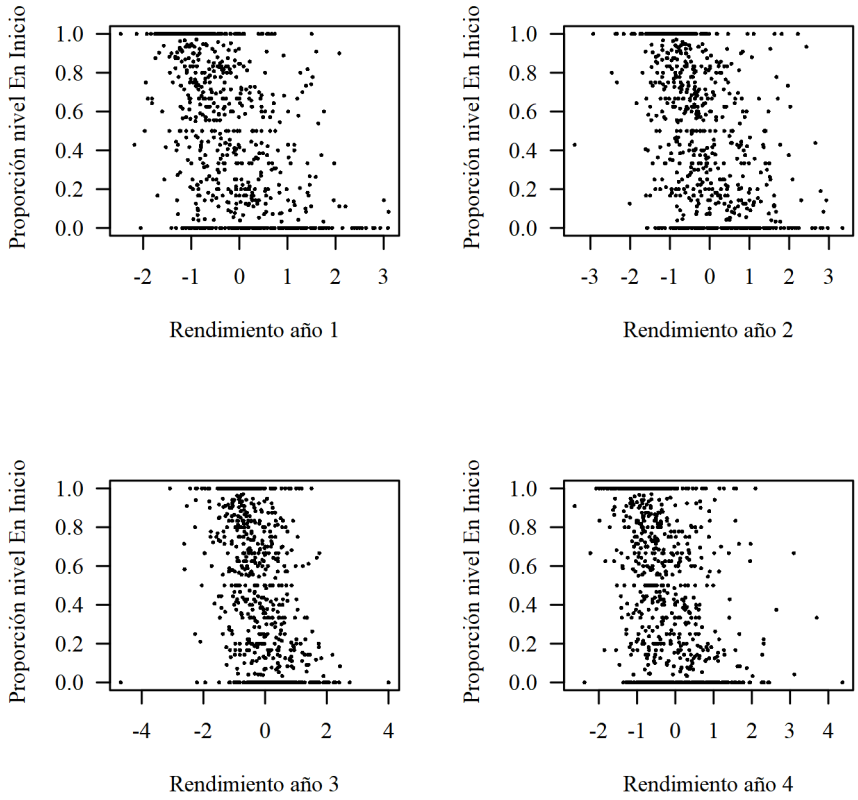
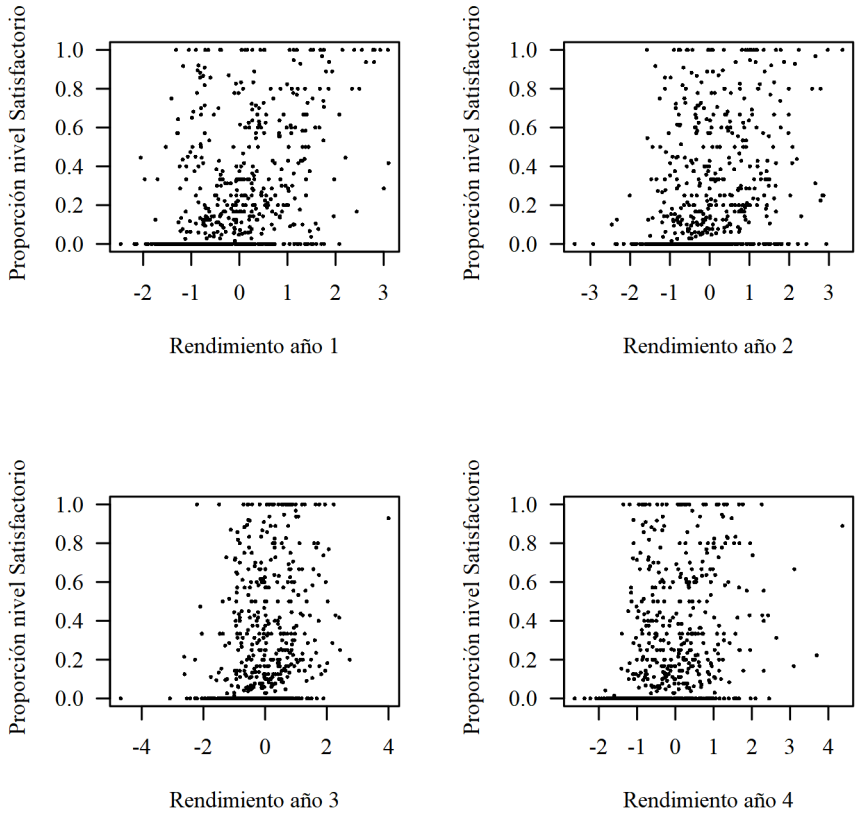


Figura 5.3: Diagrama de dispersión entre la proporción de estudiantes en el nivel Satisfactorio y el rendimiento en años anteriores



A continuación se utilizó el modelo de regresión beta inflacionada para modelar la relación entre la proporción de estudiantes en los niveles de logro En Inicio y Satisfactorio en Lectura y el rendimiento (medida promedio) de años anteriores en Lectura. En este caso se modela la proporción en el quinto año y se utiliza la información de rendimiento de los 4 años anteriores para predecirla. Se considera el modelo (3.3), denotando la proporción de estudiantes de una escuela en determinado nivel de logro como Y, la cual se distribuye como $Y_i \sim \mathcal{BI}m(\alpha_0, \alpha_1, \gamma_i, \phi)$, donde α_0, α_1 y ϕ se asumen como fijos. Mientras que la media de esta variable se modela como:

$$g(\gamma_i) = {x_i}^\top \beta, \quad i = 1, \ldots, n$$

donde $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^\top$ son los parámetros de regresión para el intercepto y el rendimiento de años anteriores y $g(\gamma_i)$ es una función de enlace logística.

Como se especificó en la sección 3.2 se utilizaron las siguientes distribuciones *a priori*: $\alpha_0 \sim U(0,1)$, $\alpha_1 \sim U(0,1)$, $\beta_0 \sim U(-5,5)$, $\beta_1 \sim U(-5,5)$ y $\phi \sim U(0,100)$. En este caso, utilizando el algoritmo de Gibbs en WinBugs (Spiegelhalter et ál., 2003) se simularon 2 cadenas de Markov de tamaño 2000 cada una, de estas se descartaron las primeras 1000 antes de obtener convergencia. Por lo tanto, se obtuvieron 2000 simulaciones de la distribución *a posteriori*, con las que se realizó la estimación de parámetros y de las divergencias para cada escuela.

Como se puede ver en la Tabla 5.2 un mayor rendimiento en los años del 1 al 4 se relaciona con

menores proporciones de estudiantes en el nivel En Inicio. Por otro lado, un mayor rendimiento se asocia principalmente con una mayor proporción de estudiantes en el nivel Satisfactorio, a excepción del rendimiento del año 3.

Tabla 5.2: Modelo de regresión beta inflacionada para estudiar los niveles En Inicio y Satisfactorio

Parámetros	**En Inicio**			**Satisfactorio**		
	Media	P2.5	P97.5	Media	P2.5	P97.5
Intercepto	-0.170	-0.283	-0.053	-1.399	-1.529	-1.272
Rendimiento año 1	-0.340	-0.466	-0.216	0.371	0.234	0.516
Rendimiento año 2	-0.213	-0.332	-0.095	0.210	0.072	0.350
Rendimiento año 3	-0.259	-0.398	-0.127	-0.058	-0.216	0.093
Rendimiento año 4	-0.153	-0.281	-0.029	0.126	-0.005	0.261
ϕ	3.368	2.981	3.745	3.186	2.766	3.635
α_0	0.402	0.358	0.449	0.648	0.615	0.680
α_1	0.409	0.361	0.455	0.203	0.148	0.272

Con las 2000 simulaciones obtenidas de la distribución *a posteriori*, utilizando el algoritmo de Gibbs, se calcularon las divergencias KL, ℓ_1 y ji-cuadrado para cada escuela (Tabla 5.3). Se puede observar que en el nivel En Inicio las escuelas del estrato 2 presentan en promedio las mayores divergencias. De otro lado, en el nivel Satisfactorio estas se ven en los estratos 1 y 3.

Tabla 5.3: Divergencias promedio ($\times 10^3$) calculadas para cada estrato

Estrato	**n**	**Porcentaje**	**En Inicio**			**Satisfactorio**		
			KL	ℓ_1	χ^2	KL	ℓ_1	χ^2
1	16	2.11 %	1.32	20.23	2.65	3.21	25.97	6.44
2	161	21.18 %	3.95	29.26	7.92	3.17	23.26	6.35
3	196	25.79 %	2.04	22.37	4.06	2.8	24.3	5.55
4	61	8.03 %	3.66	27.87	7.22	1.93	18.58	3.84
5	326	42.89 %	2.03	22.98	4.05	1.81	17.35	3.61

Las Figuras 5.4 y 5.5 muestran las divergencias de todas las escuelas de la muestra. Como se esperaba, las divergencias del estrato 2 son consistentemente mayores que las de los demás estratos. El caso del nivel Satisfactorio es menos claro pues se encuentran grandes divergencias en los estratos 3 y 5 también.

Figura 5.4: Divergencias ($\times 10^3$) en el nivel En Inicio

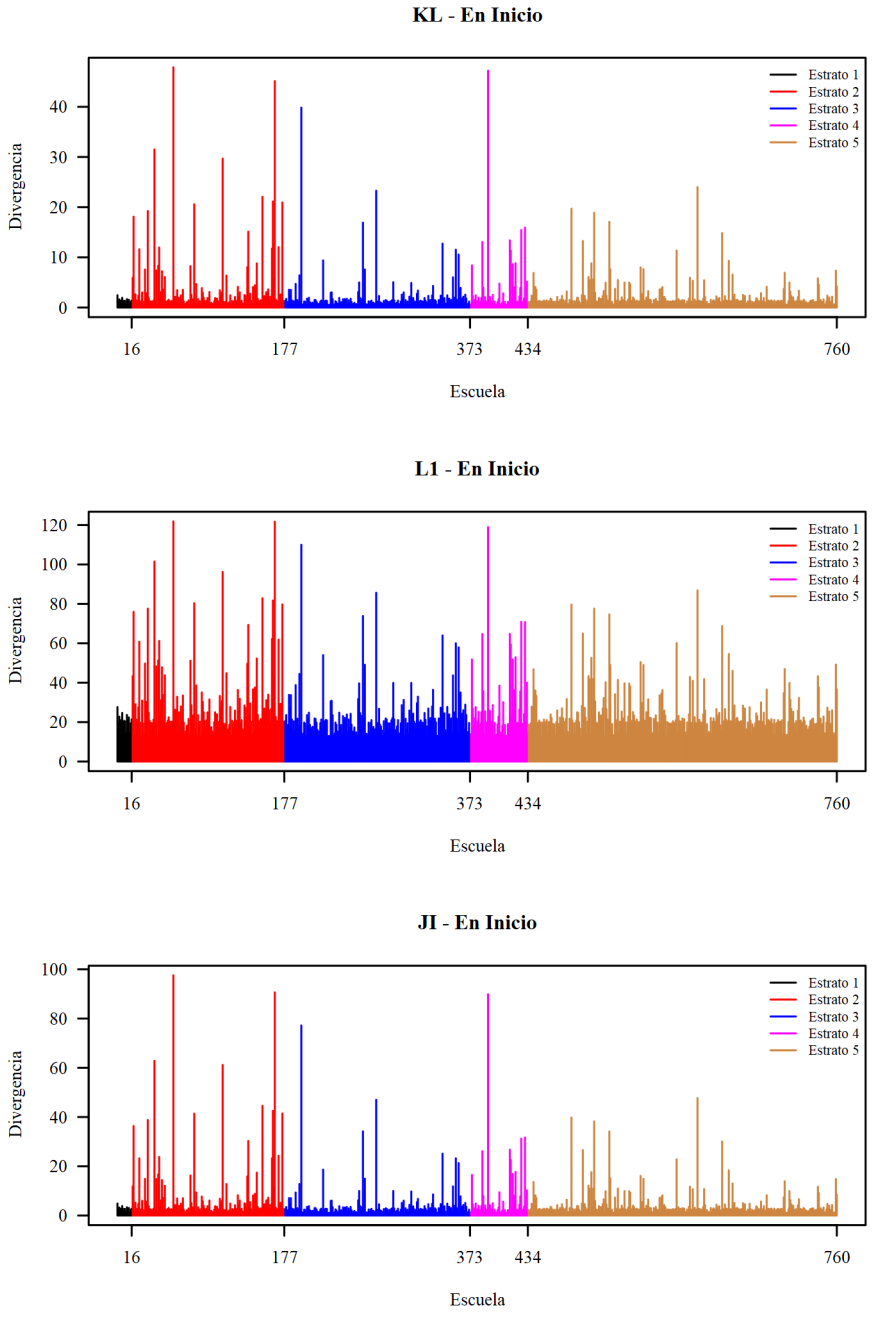
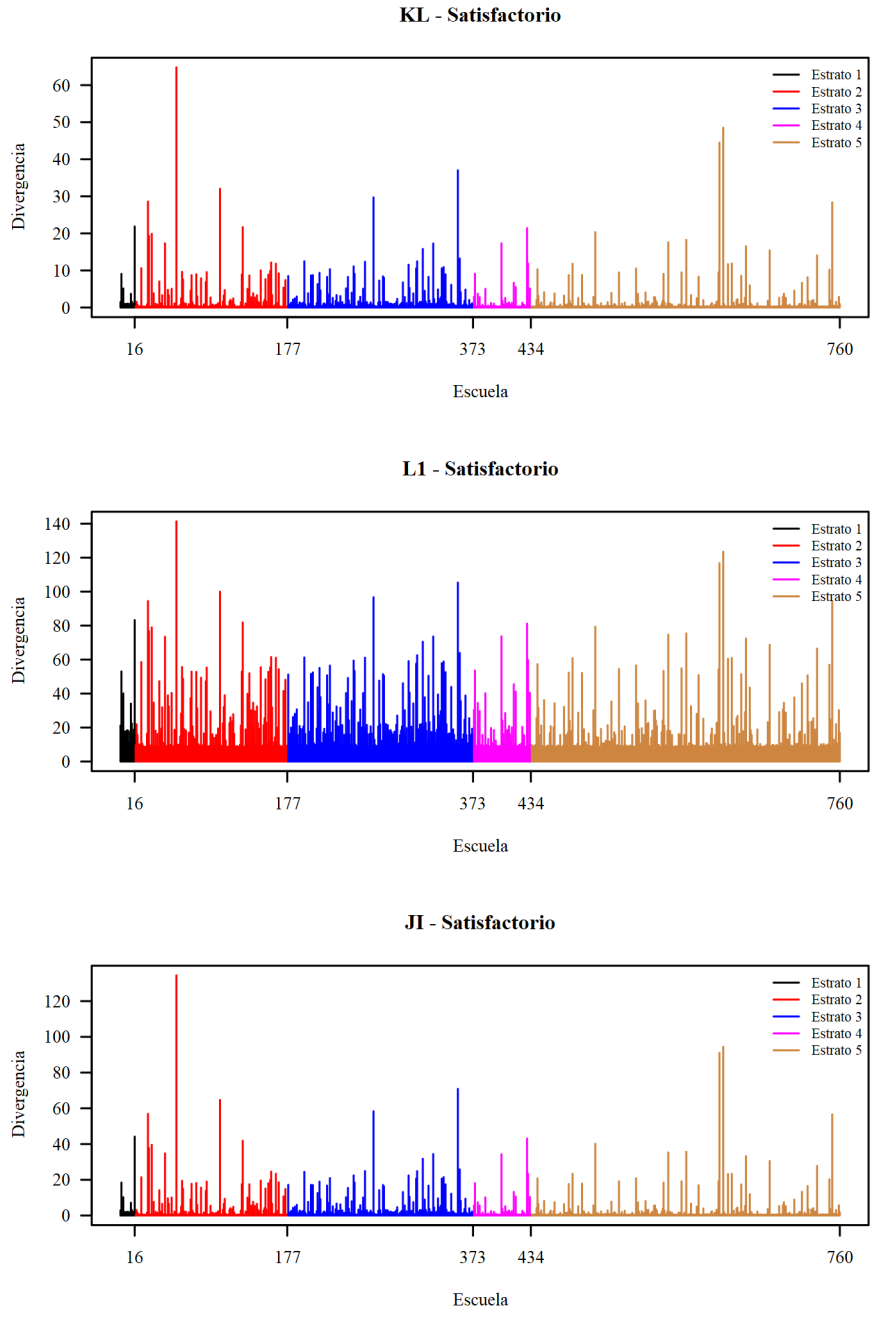


Figura 5.5: Divergencias ($\times 10^3$) en el nivel Satisfactorio



La aplicación de esta técnica para esta evaluación corrobora lo encontrado previamente por el Ministerio de Educación, sobre todo en los niveles bajos de desempeño. El estrato 2 presenta resultados consistentemente atípicos. Tanto en el nivel En Inicio como en el nivel Satisfactorio es la misma escuela (identificada como 9270604) la que presenta mayores divergencias. Esta escuela pasa de tener 89 % de sus estudiantes en el nivel En Inicio a solo 4 % y en el nivel Satisfactorio pasa de 0 % a 92 %. Otra escuela que reduce de forma considerable el porcentaje de estudiantes con bajo rendimiento es la escuela 20945644 que pasa de 100 % a 4 % de estudiantes en el nivel En Inicio. Finalmente, también existen escuelas que aumentan de forma importante el porcentaje de estudiantes con bajo rendimiento, este es el caso de la escuela 60727223, que pasa de 0 % de

estudiantes en el nivel En Inicio a 90 %.

Este método también permite detectar escuelas previamente no consideradas como atípicas por pertenecer a un estrato que en conjunto no obtiene resultados fuera de lo común. Este procedimiento puede servir como guía inicial para seleccionar escuelas en las que hacer un análisis forense de datos más exhaustivo para entender las razones de sus altos valores de divergencias.

Capítulo 6

Conclusiones

6.1. Conclusiones

En este documento se estudió la metodología para la detección de valores atípicos a través de un análisis de influencia bajo inferencia bayesiana propuesta por Peng y Dey (1995). Para ello se utilizaron medidas de ϕ-divergencia que permitieran conocer qué tanto la exclusión de un caso generaba modificaciones en la distribución *a posteriori*.

En primer lugar se presentó la aproximación de Peng y Dey (1995) para una estimación Monte Carlo de la medida de ϕ-divergencia. En el presente trabajo se estudiaron tres tipos de divergencias: Kullback-Leibler, norma ℓ_1 y ji-cuadrado. A continuación se presentó el modelo beta inflacionado que se considera pertinente para la aplicación en escuelas con estudiantes clasificados como pertenecientes a altos y bajos niveles de desempeño. Este modelo permite especificar la probabilidad de 0 y 1 dentro de una proporción.

Se realizó una aplicación en datos simulados con y sin valores atípicos. Con esta aplicación se comprobó que el método recuperaba de forma efectiva los casos atípicos como aquellos con mayores divergencias en los cuatro tipos de distorsión estudiadas. Este resultado se complementó con un estudio de simulación donde se probaron los cuatro tipos de distorsión en cuatro tipos de escenarios: con tamaños muestrales de 100, 200, 500 y 1000. En todos los escenarios, tipos de distorsiones y divergencias, los casos atípicos introducidos fueron detectados como aquellos con mayores divergencias. Esto sirvió para comprobar la pertinencia de la metodología propuesta.

Finalmente, se aplicó este análisis a un tipo de evaluación realizada por el Ministerio de Educación. En este caso, se sabía qué tipo de escuelas contaban con resultados no esperados. La metodología propuesta identificó el estrato previamente señalado como atípico con las mayores divergencias tomando en cuenta resultados anteriores.

6.2. Sugerencias para futuras investigaciones

- A partir de esta metodología se sugiere realizar un análisis más profundo sobre la información individual de aquellas escuelas que presentan mayores divergencias en esta evaluación, así como en otras pruebas del Ministerio de Educación. Esto permitiría determinar las posibles razones de los resultados atípicos.

- De otro lado, sería importante considerar el modelamiento de los parámetros α_0 y α_1 que podría permitir encontrar en qué medida las covariables impactan en la consecución de 0 % y 100 % de estudiantes en los niveles más altos y bajos de rendimiento.

- También sería importante medir otras covariables que expliquen los resultados de las es-

cuelas, como por ejemplo, condiciones en las que opera la escuela, área, tipo de gestión, conocimiento docente, etc.

- Finalmente, podría explorarse la utilización de otros modelos de regresión. Por ejemplo, podría considerarse el resultado de la medida continua de rendimiento (medida promedio) (Ministerio de Educación, 2017) como la variable respuesta y diferentes funciones de enlace. Esto podría ocasionar la detección de diferentes escuelas o incluso estudiantes con valores atípicos.

Capítulo 7

Anexo

7.1. Código en WinBUGS para la estimación de modelo beta inflacionado

```
model{
# YC = Y continuo
# YD = Y discreto
# NC = conteo de Y continuos
# N = conteo total
for(j in 1:NC){
YC[j] ~ dbeta(a[j], b[j])
}
for(i in 1:N){
a[i] <- mu[i]*fi
b[i] <- (1-mu[i])*fi
mu[i] <- ( G[i]*(1-alfa1) )/( 1-alfa0*(1-G[i])-alfa1*G[i] )
YD[i,1:3] ~ dmulti(p[i, 1:3], 1)

p[i,1] <- 1- delta0[i]-delta1[i]
p[i,2] <- delta0[i]
p[i,3] <- delta1[i]
# donde
delta0[i] <- alfa0*(1-G[i])
delta1[i] <- alfa1*G[i]
Z[i] <- beta00+beta01*X1[i]+beta02*X2[i]+beta03*X3[i]
G[i] <- exp(Z[i])/(1+exp(Z[i]))
}
fi ~ dunif(0,100)
beta00 ~ dunif(-5,5)
beta01 ~ dunif(-5,5)
beta02 ~ dunif(-5,5)
beta03 ~ dunif(-5,5)
alfa0 ~ dunif(0,1.0)
alfa1 ~ dunif(0,1.0)
}
```

7.2. Gráficos de convergencia y autocorrelación

Figura 7.1: Estimación de la aplicación en datos simulados sin considerar distorsiones

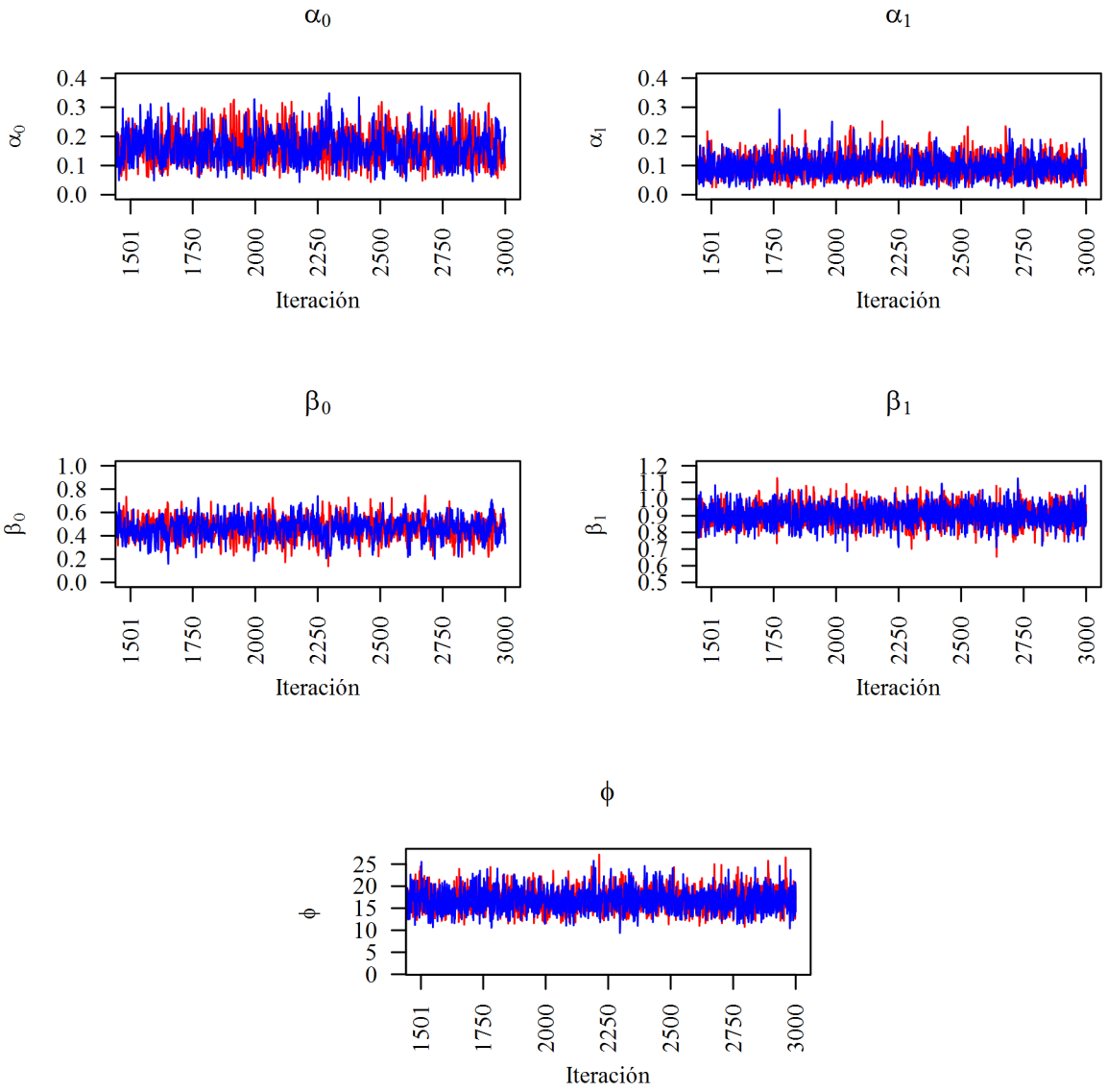
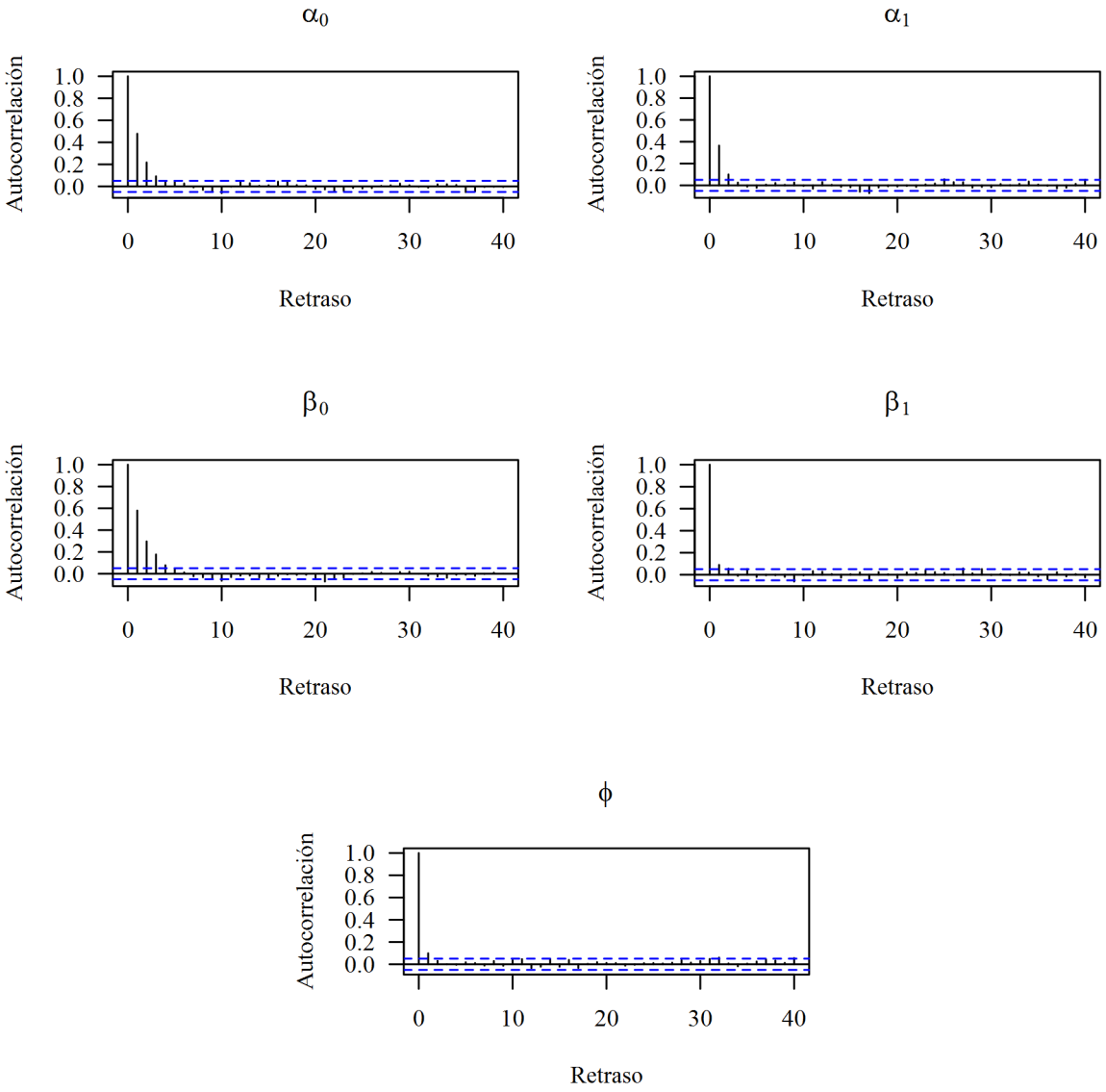


Figura 7.2: Gráficos de autocorrelación para la estimación de la aplicación en datos simulados sin considerar distorsiones



7.3. Código en R para cálculo de divergencias

```
dbetavec<-function(x,mu,fi){
dbeta(x,(mu*fi),(fi*(1-mu)))
}
meancuad<-function(x){mean(x**2)}
sumNO<-function(x){
sum(abs(x-(1/length(x))*sum(x)))  /(2*sum(x))
}
diKL<-function(x){
log(apply(x,2,mean)/apply(x,2,geometric.mean))
}
diJI<-function(x){
(apply(x,2,meancuad)/(apply(x,2,mean))**2)-1
}
diNO<-function(x){
apply(x,2,sumNO)
}


MSS<-matrix(NA,nrow=ene,length(Y))
for(i in 1:length(Y)){
Z<-beta0+beta1*X1[i]+beta2*X2[i]+beta3*X3[i]
G<-1/(1+exp(-Z))
U<-(G*(1-alfa1))/(1-alfa0*(1-G)-alfa1*G)
P<-fi
d0<-alfa0*(1-G)
d1<-alfa1*G
ciT<-(1-d0-d1)*mapply(dbetavec,x=rep(Y[i],ene),mu=U,fi=P)
if(Y[i]==0){MSS[,i]<-d0}
if(Y[i]==1){MSS[,i]<-d1}
if((Y[i]!=0)&(zY[i]!=1)){MSS[,i]<-ciT}
}
MSS<-1/MSS

div<-matrix(NA,length(Y),1*3)
div[,1]<-diKL(MSS)
div[,3]<-diJI(MSS)
div[,2]<-diNO(MSS)
```

Referencias

Ali, S. & Silvey, S. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, *28*(1), 131-142.

Amari, S. & Nagaoka, H. (2000). *Methods of information theory.* Providence, RI: American Mathematical Society.

Bayes, C. & Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables. *Journal of Applied Statistics*, *43*(10), 1814-1830.

Cizek, G. J. & Wollack, J. A. (2017). Handbook of Quantitative Methods for Detecting Cheating on Tests. En G. J. Cizek & J. A. Wollack (Eds.), (caps. Exploring Cheating on Tests. The Context, the Concern, and the Challenges). New York: Routledge.

Csiszár, I. & Shields, P. (2004). *Information theory and statistics: A tutorial.* Hanover, MA: now Publishers.

Dey, D. & Birmiwal, L. (1993). Robust bayesian analysis using entropy and divergence measures. *Statistics & Probability Letters*, *20*, 287-294.

Everitt, B. & A., S. (2010). *The cambridge dictionary of statistics.* Cambridge: Cambridge University Press.

Hoff, P. (2009). *A first course in bayesian statistical methods.* New York: Springer.

Jacob, B. A. & Levitt, S. D. (2003). *Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.* Cambridge: National Bureau of Economic Research.

Kim, D., Woo, A. & Dickison, P. (2017). Handbook of Quantitative Methods for Detecting Cheating on Tests. En G. J. Cizek & J. A. Wollack (Eds.), (cap. Identifying and Investigating Aberrant Responses using Psychometrics-Based and Machine Learning-Based Approaches). New York: Routledge.

Kullback, S. (1978). *Information theory and statistics.* Glouchester, MA: now Publishers.

Ministerio de Educación. (2017). *Reporte técnico de la Evaluación Censal de Estudiantes (ECE 2016). Segundo y cuarto de primaria, segundo de secundaria.* Lima: Oficina de Medición de Calidad de los Aprendizajes.

Ministerio de Educación. (2017). *Reporte técnico de la evaluación censal de estudiantes (ece 2016). segundo y cuarto de primaria, segundo de secundaria.* Lima: Oficina de Medición de la Calidad de los Aprendizajes.

Ospina, R. & Ferrari, S. (2010). Inflated beta distributions. *Statistical papers*, *51*, 111-126.

Peng, F. & Dey, D. (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, *23*(2), 199-213.

Rawlings, J., Sastry, G. & Dickey, D. (2001). *Applied regression analysis: A research tool.* New York: Springer-Verlag.

Skorupski, W., Fitzpatrick, J. & Egan, K. (2017). Handbook of Quantitative Methods for Detecting Cheating on Tests. En G. J. Cizek & J. A. Wollack (Eds.), (cap. A Bayesian Hierarchical Model for Detecting Aberrant Growth at the Group Level). New York: Routledge.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). *WinBUGS User Manual*.

Whitley, B. E. (1998). Factors Associated with Cheating Among College Students: A Review. *Research in Higher Education*, *39*(3), 235-274.