Modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos

Juan Camilo Díaz Zapata

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2017

Modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos

Juan Camilo Díaz Zapata

Tesis presentada como requisito parcial para optar al título de:
Magister en Ciencias Estadísticas

Director:
Ph.D. Freddy Hernández Barajas

Línea de Investigación:
Modelos de Regresión Mixtos

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2017

(Dedicatoria)

Este trabajo es dedicado a mi madre, Blanca y a mi futura esposa Zaret.

Agradecimientos

A Dios, por regalarme salud, tiempo, sabiduría e inteligencia para poder realizar este trabajo, que toda la gloria sea para él.

A mi tutor, Freddy Hernández, por el constante apoyo, paciencia, orientación y confianza depositada a lo largo de estos dos años.

A mis padres, Blanca y Guillermo por la paciencia, tiempo, cariño y apoyo día tras día.

A mi prometida, Zaret, por su paciencia, apoyo incondicional y comprensión; Por ayudarme, escucharme y alentarme en los momentos más difíciles y así poder ver realizado este trabajo.

A mis sobrinos, José Manuel y Julieta, por la comprensión y el tiempo regalado, a su corta edad respetaron cada uno de los espacios necesarios para elaborar este trabajo.

A mi jefe, Andrés Ibarra, por haberme brindado el tiempo para realizar este trabajo y los datos que me permitieron realizar las aplicaciones a datos reales.

A la universidad de Antioquia, por brindarme sus instalaciones, que me permitieron trabajar por buen tiempo en este trabajo.

Resumen

El modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos, es un modelo de regresión donde la variable respuesta se encuentra definida a partir de una distribución para datos proporcionales, como la distribución beta o la distribución simplex, que dan resultados en el intervalo cero uno, más dos valores dados por cero y/o uno, representando la ausencia o presencia total de cierta característica. Diferentes autores han trabajado en el desarrollo de diferentes modelos y metodologías de estimación, sin embargo, no se ha desarrollado un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos, que reúna los principales modelos de regresión de este tipo y que la estimación de los parámetros sea vía máxima verosimilitud y la cuadratura de Gauss-Hermite. En este trabajo se presenta el paquete **ZOIP** del sistema computacional R, en el que se implementa la distribución ZOIP (Zeros Ones Inflated Proporcional), el modelo de regresión para efectos fijos y mixtos, ZOIP, que reúne las distribuciones y los modelos de regresión de efectos fijos y mixtos para las distribuciones beta y simplex inflada con ceros y/o unos, la estimación de los parámetros se hace vía máxima verosimilitud y la cuadratura de Gauss-Hermite. Se realizan diferentes estudios de simulación que muestran la convergencia de los parámetros y la alternativa de estimación que presenta mejor desempeño. Además, se presenta el ajuste de diferentes modelos de regresión ZOIP a datos reales.

Palabras clave: modelos lineales mixtos, datos proporcionales inflados, cuadratura de Gauss-Hermite, máxima verosimilitud.

Abstract

The mixed regression model for proportional data inflated with zeros and/or ones is a regression model where the response variable is defined by a proportional data distribution. A proportional data distribution could be the beta distribution or the simplex distribution, which its results are given in the interval zero one, and in other two values given by zero and/or one that represent the absence or total presence of a given characteristic. Although several authors have worked on the development of different models and estimation methodologies, a mixed regression model for proportional data inflated with zeros and/or ones that aggregates the main regression models and the estimation of the parameters through the maximum likelihood and Gauss-Hermite quadrature has not been developed. In this paper, it is presented the ZOIP package for the computational system R, in which the ZOIP (Zeros Ones Inflated Proportional) distribution is implemented. In the ZOIP package is used a regression model for fixed and mixed effects, which unifies the distributions and these types of regressions for beta and simplex distributions inflated with zeros and/or ones, and the estimation of the parameters solved by the maximum likelihood and Gauss-Hermite quadrature. Several simulation studies are performed showing the convergence of the parameters and the

estimation alternative that presents the best performance. Furthermore, it is presented the adjustment of different ZOIP regression models to real data.

Keywords: linear mixed models, proportional inflated data, Gauss-Hermite quadrature, maximum likelihood.

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

En el análisis de datos es usual estudiar variables aleatorias que den como respuesta datos proporcionales, derivados de cálculos de tasas, porcentajes y razones; esta variable aleatoria se encuentra descrita en el intervalo (0,1) y es caracterizada por lo general por la distribución beta o la distribución simplex (Jørgensen, 1997), sobre la distribución beta se han encontrado diferentes parametrizaciones, de acuerdo a la modificación de sus parámetros, existe la parametrización original, la parametrización de Ferrari y Cribari-Neto (2004) y la de Rigby y Stasinopoulos (2005), sin embargo, las anteriores distribuciones no tienen en cuentan que existen ocasiones donde este tipo de variables aleatorias dan respuestas de cero o de uno, es por esto que otros autores como Ospina y Ferrari (2010) proponen una nueva clase de distribución, por ejemplo la beta, que permiten incluir los valores de cero o uno, llamada distribución beta inflada por ceros o unos.

El análisis de regresión sobre variables aleatorias infladas con ceros y/o unos, considera diferentes desafíos, uno de ellos es la inclusión en la modelación de los parámetros que permiten describir la inflación de los datos en cero y uno, por lo que autores como Ospina y Ferrari (2012) y Kosmidis et al. (2015) han trabajado sobre el desarrollo de modelos de regresión para datos proporcionales inflados en ceros y/o unos. El segundo desafío es la estimación de los parámetros regresores asociados a los cuatro parámetros de la distribución que describe una variable aleatoria para datos proporcionales inflada con cero y/o unos, dicha estimación puede ser realizada vía máxima verosimilitud (Ospina y Ferrari, 2012) o mediante metodologías bayesianas como la propuesta por Galvis y Lachos (2014), estas metodologías deben ser resueltas computacionalmente.

Los modelos de regresión mixtos para datos proporcionales inflados con ceros y/o unos, representan un desafío mayor, entorno a la estimación de los parámetros, es por eso que algunos autores como Usuga y Giampaoli (2013), Bonat et al. (2015), Song et al. (2004) y Rigby y Stasinopoulos (2005) presentan modelos de regresión para datos proporcionales mixtos, sin incluir las inflaciones en cero o uno, la inclusión de los parámetros de inflación más los efectos aleatorios hacen que la estimación por máxima verosimilitud no se resuelve computacionalmente fácil, ni mucho menos analíticamente, ya que requiere la solución de integrales no cerradas asociadas a los efectos aleatorios, dichas integrales deben ser aproximadas por diferentes metodologías, como por ejemplo la cuadratura de Gauss-Hermite, introducido en los modelos lineales generalizados por Fahrmeir y Tutz (2001) o el método de reducción se-

cuencial (Ogden, 2015), entre otros. El hecho de que la estimación vía máxima verosimilitud sea complicada, hace que otros autores como Galvis y Lachos (2014) busquen una solución alternativa dentro de las metodologías bayesianas, MCMC.

A nivel computacional es posible mencionar que existen diferentes paquetes en el software R que describen los datos proporcionales, uno de ellos y el más común es el paquete **betareg** (Zeileis et al., 2010), (Ferrari y Pinheiro, 2010) y (Simas et al., 2010), el cual incluye la distribución beta y la estimación de modelos de regresión beta fijos y mixtos. Qiu et al. (2008) implementan el paquete **simplexreg** el cual incluye la distribución y el modelo de regresión con efectos fijos de la distribución simplex.

En este trabajo se incluye en una sola distribución y en un solo modelo de regresión de efectos fijos y mixtos, las principales distribuciones y modelos de regresión para datos proporcionales inflados con ceros y/o unos, por lo que es creada la distribución ZOIP (Zeros Ones Inflated Proporcional), los modelos de regresión ZOIP para efectos fijos y mixtos. La estimación de los parámetros regresores de los modelos de regresión ZOIP, son realizados vía máxima verosimilitud y la cuadratura de Gauss-Hermite. Adicional como no existe un paquete en R que reúna las principales distribuciones y modelos de regresión de efectos fijos y mixtos para modelar los datos proporcionales inflados con ceros y/o unos, se implementa el paquete **ZOIP**, que permite generar y ajustar distribuciones y modelos de regresión para efectos fijos y mixtos para datos proporcionales inflados con ceros y/o unos.

1.1. Organización del trabajo

La estructura de este trabajo es considerada de la siguiente manera, en el capítulo 2 se presenta las principales distribuciones descritas en la literatura para datos proporcionales, en diferentes parametrizaciones, además se implementa la distribución ZOIP para datos proporcionales inflados con ceros y/o unos, propuesta en este trabajo, luego se muestra cómo se utiliza la distribución ZOIP en el paquete propuesto **ZOIP**, por último se realizan diferentes estudios de simulación para argumentar la convergencia satisfactoria del ajuste de la distribución ZOIP y una aplicación a datos reales.

Un modelo de regresión de efectos fijos para datos proporcionales es planteado en el capítulo 3, con base a la distribución ZOIP propuesta en el capítulo anterior, se muestra la inferencia estadística para la estimación de los parámetros vía máxima verosimilitud, luego se observa cómo utilizar el modelo de regresión para efectos fijos en el paquete **ZOIP** y las salidas de las funciones de dicho paquete asociadas a la estimación de un modelo de regresión ZOIP, por último se realiza un estudio de simulación donde se demuestra la convergencia de la estimación de los parámetros regresores y se muestra el ajuste de un modelo de regresión ZOIP a datos reales.

En el capítulo 4 se muestra la implementación y el desarrollo analítico de un modelo de regresión para datos proporcionales inflados con ceros y/o unos, teniendo en cuenta efectos fijos e interceptos aleatorios en los parámetros de la media y la varianza, se muestra la derivación de la función de verosimilitud, necesaria para estimar los parámetros regresores asociados al modelo, para dicha estimación es necesario utilizar una aproximación, la aproximación utilizada en este trabajo es la cuadratura de Gauss-Hermite, en este capítulo también se muestran las diferentes variaciones que existen en la cuadratura de Gauss-Hermite, seguido de la aproximación de la función de verosimilitud por medio de la cuadratura de Gauss-Hermite adaptativa. Se observa la forma de utilizar el paquete **ZOIP** para el ajuste de un modelo de regresión ZOIP mixto. Luego se muestra el ajuste de un modelo de regresión ZOIP mixto a datos reales y un estudio de simulación con varios escenarios que nos permite concluir, cual es la mejor metodología de estimación de los parámetros regresores, en un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos.

Finalmente, en el capítulo 5 se presentan las conclusiones, recomendaciones y trabajos futuros derivados de este trabajo.

2 Capítulo 2: Distribución ZOIP

En modelación estadística es posible encontrarnos con variables respuesta como proporciones, porcentajes o tasas que se encuentran en el intevalo (0, 1). La distribución más utilizada en la literatura para caracterizar este tipo de variables es la distribución beta con soporte en el intervalo (0,1), la cual ha sido reparametrizada por autores como Ferrari y Cribari-Neto (2004) y Rigby y Stasinopoulos (2005); otras distribuciones no tan comunes en la literatura pero que caracterizan este tipo de variables son la distribución simplex (Jørgensen, 1997), beta-rectangular (Hahn, 2008) y la distribución LogitSep (Hossain, 2015). Por otra parte, es común que los porcentajes o proporciones puedan dar valores iguales a cero o uno, representando la ausencia o presencia total de la característica de interés, respectivamente, Las distribuciones descritas anteriormente no pueden ser admisibles para este tipos de variables, es por esto que se han desarrollado distribuciones infladas con ceros y/o unos, para tratar estos casos, como lo hizo Ospina y Ferrari (2010) quienes presentan una distribución beta inflada en la que hacen una combinación entre una distribución discreta para la parte de los valores que pueden tomar cero o uno y una parte continua para los valores continuos entre cero y uno. Rigby y Stasinopoulos (2005) incluyen dentro de sus modelos gamlss la distribución beta inflada con ceros y/o unos según su parametrización.

Esto ha dado pie para que diferentes autores hayan empezado a desarrollar diferentes modelos de regresión para tratar este tipo variables, Ospina y Ferrari (2012) propusieran una clase general de modelos de regresión beta inflados con cero o uno, además Kosmidis et al. (2015) han estudiado dichos modelos inflados recientemente, pero con una distribución distinta a la presentada por Ospina y Ferrari (2012). Galvis y Lachos (2014) presentan modelos de regresión para diferentes distribuciones para datos proporcionales inflados con ceros y/o unos mediante metodologías de estimación bayesianas.

Muchos autores han implementado distribuciones para datos proporcionales en el software estadístico R, Zeileis et al. (2010) implementan el paquete **betareg** donde se encuentran los modelos de regresión beta propuestos por Ferrari y Cribari-Neto (2004), Qiu et al. (2008) implementan el paquete **simplexreg** para realizar análisis de distribución y regresión sobre una distribución simplex, para datos proporcionales no inflados, otros autores como (Rigby y Stasinopoulos, 2008) incluyen en el paquete **gamlss** la distribución beta inflada con ceros y/o unos y la posibilidad de realizar modelos de regresión sobre ellos.

Aunque muchos autores han implementado las distribuciones para datos proporcionales inflados con ceros y/o unos, ninguno ha presentado una propuesta como la de reunir en una sola distribución las diferentes distribuciones para datos proporcionales y sus diferentes parametrizaciones, además de implementarla en un solo paquete, como se presenta en el paquete **ZOIP** en R Core Team (2017) disponible en el repositorio web `GitHub`.

El capítulo se encuentra organizado de la siguiente manera: primero se presentan las distribuciones más representativas para datos proporcionales, en la siguiente sección se presenta la distribución para datos proporcionales inflados con ceros y/o unos ZOIP (Zeros Ones Inflated Proportional), seguido por el desarrollo analítico de la estimación de los parámetros de la distribución ZOIP vía máxima verosimilitud, luego se presenta la forma de utilizar el paquete **ZOIP** para ajustar una distribución ZOIP, por último se aplica el ajuste de una distribución ZOIP en un estudio de simulación y para datos reales.

2.1. Distribución para datos proporcionales

Para los casos de modelación donde la variable de interés es una proporción, un porcentaje o una tasa. Este tipo de variables no pueden ser analizadas con la distribución normal, debido a que el soporte de la normal es la recta real $\mathbb{R}$, además en este tipo de variables es común la asimetría e incluso la bimodalidad, por esta razón en la literatura estadística se han propuesto distribuciones para este tipo de comportamientos, como la distribución beta, que cuenta con diferentes parametrizaciones (Ferrari y Cribari-Neto (2004) y Rigby y Stasinopoulos (2005)) y la distribución simplex propuesta por Jørgensen y Barndorff-Nielsen (1991), de igual manera otras distribuciones más particulares como la beta-rectangular (Hahn, 2008) y LogitSep (Hossain, 2015) se acoplan a este comportamiento, a continuación se mostraran las funciones de densidad de probabilidad, la media, la varianza y dependencias de algunas de las distribuciones mencionadas anteriormente.

2.1.1. Distribución beta original

Si una variable aleatoria y definida entre cero y uno, tiene distribución beta con parámetros p y q se acostumbra a denotarla por $y \sim \text{Be}(p, q)$ y la función de densidad de probabilidad de la distribución es dada por:

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{2-1}$$

donde los parámetros $p > 0$, $q > 0$ y $\Gamma(\cdot)$ es la función gamma. El valor esperado y la varianza de y están dadas por:

$$E(y) = \frac{p}{p+q} \tag{2-2}$$

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{2-3}$$

2.1.2. Distribución beta parametrización Ferrari y Cribari-Neto (2004)

Ferrari y Cribari-Neto (2004) propusieron otra parametrización para la distribución beta en función de los parámetos μ y ϕ donde μ corresponde a la media de la distribución y ϕ es interpretado como un parámetro de precisión. Si $0 < y < 1$ y $Y \sim \text{Be}(\mu, \phi)$ la función de densidad de probabilidad de la distribución está dada por:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{2-4}$$

donde $0 < \mu < 1$ y $\phi > 0$. El valor esperado y la varianza de y están dados por:

$$E(y) = \mu \tag{2-5}$$

$$Var(y) = \frac{\mu(1-\mu)}{1+\phi} \tag{2-6}$$

Además note que la parametrizacion de la distribución beta original es equivalente a la de Ferrari y Cribari-Neto (2004) cuando:

$$p = \mu\phi \tag{2-7}$$

$$q = (1-\mu)\phi \tag{2-8}$$

2.1.3. Distribución beta parametrización Rigby y Stasinopoulos (2005)

Rigby y Stasinopoulos (2005) propusieron una nueva parametrización para la distribución beta con parámetros μ y σ donde μ es la media de la distribución y σ es interpretado como un parámetro de dispersión, se dice que $Y \sim \text{Be}(\mu, \sigma)$ con $0 < y < 1$, si la función de densidad de probabilidad de la distribución está dada por:

$$f(y;\mu,\sigma) = B(\mu,\sigma) y^{\mu((1-\sigma^2)/\sigma^2)-1}(1-y)^{(1-\mu)((1-\sigma^2)/\sigma^2)-1} \tag{2-9}$$

donde $B(\mu,\sigma) = \frac{\Gamma((1-\sigma^2)/\sigma^2)}{\Gamma(\mu((1-\sigma^2)/\sigma^2))\Gamma((1-\mu)((1-\sigma^2)/\sigma^2))}$,

donde $0 < \mu < 1$ y $0 < \sigma < 1$. La media y la varianza de y están dadas por:

$$E(y) = \mu \tag{2-10}$$
$$Var(y) = \sigma^2\mu(1-\mu) \tag{2-11}$$

Además note que la parametrizacion de la distribucion beta original es equivalente a la de Rigby y Stasinopoulos (2005) cuando:

$$p = \frac{\mu(1-\sigma^2)}{\sigma^2} \tag{2-12}$$
$$q = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \tag{2-13}$$

2.1.4. Distribución simplex

La distribución simplex fue introducida por Jørgensen y Barndorff-Nielsen (1991) y es un caso particular de los modelos de dispersión propuestos por Jørgensen (1997), dicha distribución depende de los parámetros μ que es la media de la distribución y σ^2 que es un parámetro de dispersión. Si $0 < y < 1$ y $Y \sim S^{-}(\mu, \sigma^2)$ la función de densidad de probabilidad esta dada por:

$$f(y;\mu,\sigma^2) = \left\{2\pi\sigma^2[y(1-y)]^3\right\}^{-1/2} \exp\left\{-\frac{y(1-y)\mu^2(1-\mu)^2}{2\sigma^2(y-\mu)^2}\right\} \tag{2-14}$$

donde $0 < \mu < 1$ y $\sigma > 0$. Además el valor esperado y la varianza están dadas por:

$$E(y) = \mu \tag{2-15}$$
$$Var(y) = \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\} \tag{2-16}$$

donde $\Gamma(a,b)$ está dado por la función$-\Gamma$ incompleta definido como $\Gamma(a,b) = \int_b^\infty t^{a-1}b^t dt$. ver más en Zhang et al. (2016).

2.2. Distribución ZOIP (Zeros Ones Inflated Proporcional)

En las distribuciones vistas en la sección 2.1, se evidenciaron ciertas distribuciones que se ajustan al comportamiento de datos proporcionales, porcentajes o tasas que están en el intervalo (0,1), sin embargo, es común que estos datos tomen valores en cero y/o uno que representarían la ausencia o presencia total de cierta característica, por lo que no sería posible ajustar los datos a las distribuciones vistas anteriormente y es por eso que en este trabajo se propone la distribución ZOIP, como un conjunto de distribuciones para datos proporcionales inflados con ceros y/o unos.

La distribución para datos proporcionales inflados con ceros y/o unos se compone de la mezcla de tres distribuciones, dos de ellas discretas, que son distribuciones degeneradas en cero y uno, y una tercera distribución continua que además es una función de densidad de probabilidad para datos proporcionales, como las presentadas anteriormente, si la variable aleatoria y tiene distribución ZOIP con parámetros μ, σ, p_0 y p_1, se denotará como $Y \sim ZOIP(\mu, \sigma, p_0, p_1)$, la función de densidad de probabilidad está dado por:

$$g(y; \mu, \sigma, p_0, p_1) = \begin{cases} p_0 & \text{si } y = 0, \\ p_1 & \text{si } y = 1, \\ (1 - p_0 - p_1) f(y; \mu, \sigma) & \text{si } y \in (0, 1) \end{cases} \tag{2-17}$$

donde $p_0 \geq 0$ representa la probabilidad que $y = 0$ y $p_1 \geq 0$ representa la probabilidad de que $y = 1$, además $0 \leq p_0 + p_1 \leq 1$ y $f(y; \mu, \sigma)$ representa alguna de las funciones de densidad de probabilidad para datos proporcionales, descritas en la sección anterior. La media y varianza de y, están dadas por

$$E(y) = p_1 + (1 - p_0 - p_1) E^*(y) \tag{2-18}$$

$$Var(y) = p_1(1 - p1) + (1 - p_0 - p_1) \left[Var^*(y) + (p_0 + p_1)[E^*(y)]^2 - 2E^*(y) p_1 \right] \tag{2-19}$$

donde $E^*(y)$ es el valor esperado de una distribución para datos proporcionales como las descritas en (2-2), (2-5), (2-10) y (2-15). Además la $Var^*(y)$ es la varianza de una distribución para datos proporcionales como se muestra en (2-3), (2-6), (2-11) y (2-16).

Si para la distribución $ZOIP(\mu, \sigma, p_0, p_1)$ se elige la distribución beta con parametrización Ferrari y Cribari-Neto (2004) entonces el parámetro σ tomará el rol del parámetro ϕ de la

distribución, si la parametrización es beta original μ y σ tomarán el rol de p y q respectivamente. En las demás distribuciones y parametrizaciones μ y σ tomaran los valores y dominios correspondientes a su distribución.

La distribución ZOIP se encuentra inflada con ceros y unos, es decir bilateralmente, pero existe la posibilidad que hayan casos de estudio en que se encuentren datos inflados con unos únicamente, por lo que $p_0 = 0$ y por lo tanto se estará llamando no una distribución ZOIP, sino una distribución OIP (Ones Inflated Proporcional) y si los datos se encuentran inflados con ceros únicamente, es decir $p_1 = 0$ se tendrá una distribución ZIP (Zeros Inflated Proporcional), Si los datos no se encuentran inflados, entonces $p_0 = p_1 = 0$ y la distribución ZOIP será una distribución para datos proporcionales clásica.

En la Figura **2-1** se muestran las densidades para varias de las distribuciones ZOIP-beta en sus diferentes parametrizaciones y ZOIP-simplex, es de aclarar que en las Figuras 1a. 1b. 1c. los valores de los parámetros son diferentes pero dan como resultado la misma distribución gracias a las ecuaciones descritas en (2-7), (2-8) para el caso Ferrari y Cribari-Neto (2004) y (2-12), (2-13) para el caso Rigby y Stasinopoulos (2005). Además se puede observar en la Figura 1d. como la distribución ZOIP-simplex hereda el comportamiento bimodal de la distribución clásica simplex, con valores mínimo y máximo en cero y uno respectivamente.

2.3. Inferencia estadística

Para estimar los parámetros de la distribución ZOIP se usa el método de máxima verosimilitud. La función de verosimilitud para $\boldsymbol{\theta} = (\mu, \sigma, p_0, p_1)^\top$, basado en una muestra de $\boldsymbol{y}_i$ observaciones independientes, es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\mathbf{y}_i; \mu, \sigma, p_0, p_1) \tag{2-20}$$

Para encontrar los estimadores de máxima verosimilitud (MLE) de la distribución ZOIP, se consideraran dos casos:

1. **ZOIP-beta original**
   Considera la parametrización de la distribución beta original y la ecuación definida en (3-2) se tiene que:

   $\boldsymbol{\theta} = (p, q, p_0, p_1)^\top$

   $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{\theta}|y_i) = L_1(p_0) \cdot L_2(p_1) \cdot L_3(p, q)$

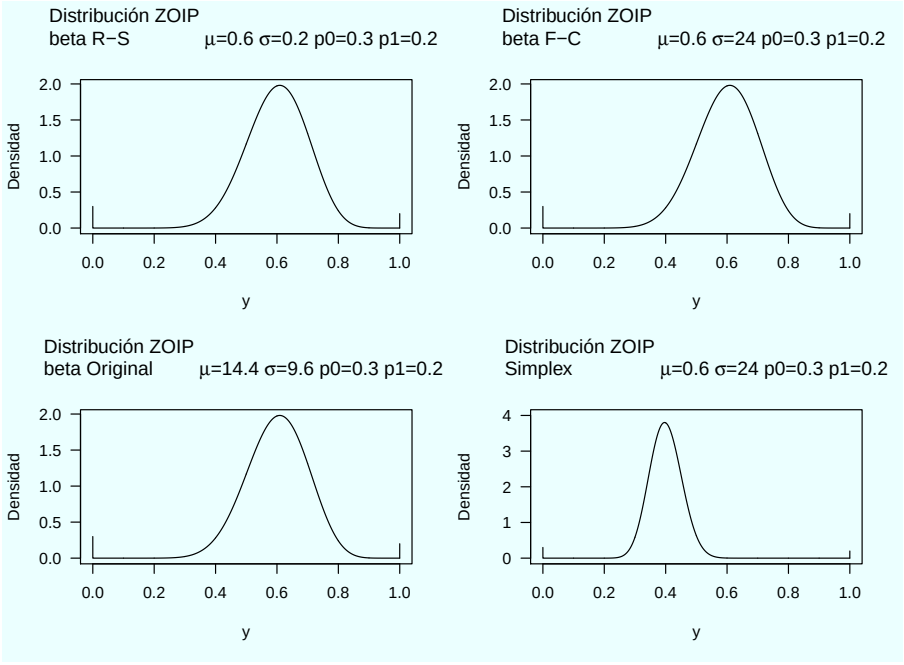

Figura 2-1: Densidades para la distribución ZOIP para algunos valores de los parámetros, donde R-S se refiere a Rigby & Stasinopoulos (2005) y F-C es Ferrari & Cribari-Neto (2004).

Note que la función de verosimilitud es factorizada en tres términos, dos de ellos del componente discreto y uno compuesto por p y q del componente continuo, por tanto los parámetros son separables (Pace y Salvan, 1997), así la máxima verosimilitud puede ser tratada por separado.

$$L_1(p_0) = \prod_{i=1}^{n} p_0^{S_0(y_i)}(1-p_0)^{1-S_0(y_i)} = p_0^{\sum_{i=1}^{n} S_0(y_i)}(1-p_0)^{n-\sum_{i=1}^{n} S_0(y_i)}$$

donde:

$$S_j(y_i) = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{si } y_i \neq j \end{cases} \quad ; \quad j = 1, 2 \tag{2-21}$$

Ahora sacando logaritmo natural a la función de verosimilitud.

$$\ell_1(p_0) = \sum_{i=1}^{n} S_0(y_i) log(p_0) + (n - \sum_{i=1}^{n} S_0(y_i)) log(1-p_0)$$

$$\frac{\delta \ell_1(p_0)}{\delta p_0} = \frac{\sum_{i=1}^{n} S_0(y_i)}{p_0} - \frac{n - \sum_{i=1}^{n} S_0(y_i)}{1 - p_0} = \sum_{i=1}^{n} S_0(y_i) - p_0 n = 0$$

$$\hat{p}_0 = \frac{1}{n} \sum_{i=1}^{n} S_0(y_i)$$

$$\therefore \hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} S_1(y_i)$$

Ahora se halla MLE para los parámetros del componente continuo de la función.

$$\ell_3(p,q) = \sum_{i=1:y_i \in (0,1)}^{n} log(f(p,q|y_i)) = nlog(\Gamma(p+q)) - nlog(\Gamma(p)) - nlog(\Gamma(q))$$

$$+(p-1) \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) + (q-1) \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i)$$

entonces

$$\frac{\delta \ell_3(p,q)}{\delta p} = \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) + \frac{n \cdot \delta log(\Gamma(p+q))}{\delta p} - \frac{n \cdot \delta log(\Gamma(p))}{\delta p} - \frac{n \cdot \delta log(\Gamma(q))}{\delta p} = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta q} = \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i) + \frac{n \cdot \delta log(\Gamma(p+q))}{\delta q} - \frac{n \cdot \delta log(\Gamma(p))}{\delta q} - \frac{n \cdot \delta log(\Gamma(q))}{\delta q} = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta p} = \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) - n(-\psi(p+q) + \psi(p)) = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta q} = \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i) - n(-\psi(p+q) + \psi(q)) = 0$$

donde $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$

Este sistema de ecuaciones no tiene una solución de forma cerrada, por lo que para encontrar los MLE de p y q es necesario utilizar algoritmos iterativos, por ejemplo el método de Newton Raphson, mínimos cuadrados ponderados y en el paquete **ZOIP** se utiliza optimizadores a la función de verosimilitud mediante la función `nlminb` de R, sin embargo, se puede garantizar que los puntos críticos encontrados serán máximos de la función de verosimilitud, ya que si hallamos la segunda derivada de la función se

tiene que:

$$\frac{\delta^2 \ell_3(p,q)}{\delta p^2} = -n(\psi^{'}(p) - \psi^{'}(p+q)) < 0$$

$$\frac{\delta^2 \ell_3(p,q)}{\delta q^2} = -n(\psi^{'}(q) - \psi^{'}(p+q)) < 0$$

debido que la varianza de la transformación logarítmica de la variable es:

$$var(log(y)) = E[log^2(y)] - (E[log(y)])^2 = \psi^{'}(p) - \psi^{'}(p+q) > 0$$

$$var(log(1-y)) = E[log^2(1-y)] - (E[log(1-y)])^2 = \psi^{'}(q) - \psi^{'}(p+q) > 0$$

ver más en Owen (2008).

Para encontrar las estimaciones de los parámetros de beta en parametrizaciones de Ferrari y Cribari-Neto (2004) y Rigby y Stasinopoulos (2005), basta con encontrar los estimadores MLE anteriores de la parametrización original y utilizar las ecuaciones definidas en (2-7), (2-8) para el caso de Ferrari y Cribari-Neto (2004) y (2-12), (2-13) para el caso de Rigby y Stasinopoulos (2005).

2. **ZOIP-simplex**

Para este caso lo único que varía con respecto al anterior es la estimación en el componente continuo.

$$L_3(\mu,\sigma) = \prod_{i=1:y_i\in(0,1)}^{n} \left[2\pi\sigma^2[y_i(1-y_i)]^3\right]^{-1/2} exp\left(-\frac{1}{2\sigma^2}d(y_i;\mu)\right)$$

donde $d(y_i;\mu) = \frac{y_i(1-y_i)\mu^2(1-\mu)^2}{(y_i-\mu)^2}$

$$\ell_3(\mu,\sigma) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{3}{2}\sum_{i=1:y_i\in(0,1)}^{n} log(y_i(1-y_i)) - \sum_{i=1:y_i\in(0,1)}^{n}\frac{1}{2\sigma^2}d(y_i;\mu)$$

$$\frac{\delta \ell_3(\mu,\sigma)}{\delta\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu) = \sigma(-n\sigma^2 + \sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu)) = 0$$

no es admisible que $\sigma = 0$ entonces:

$$-n\sigma^2 + \sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu) = 0$$

$$\therefore \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu)$$

El estimador MLE de σ^2 depende del valor estimado en μ, entonces:

$$\frac{\delta\ell_3(\mu,\sigma)}{\delta\sigma} = -\frac{1}{2\sigma^2}\sum_{i=1:y_i\in(0,1)}^{n}\frac{\delta d(y_i;\mu)}{\delta\mu} = 0$$

$$\frac{\delta d(y_i;\mu)}{\delta\mu} = \sum_{i=1:y_i\in(0,1)}^{n}\frac{y_i(1-y_i)\mu^2(1-\mu)^2}{2(y_i-\mu)^3} + \frac{2y_i(1-y_i)\mu(1-\mu)^2 - 2y_i(1-y_i)\mu^2(1-\mu)}{(y_i-\mu)^2} = 0$$

No tiene una solución cerrada analíticamente, entonces se deben utilizar algoritmos iterativos tal como Newton Raphson o mínimos cuadrados ponderados, en el paquete **ZOIP** se utiliza optimizadores para la función de verosimilitud mediante la función `nlminb` de R, para encontrar puntos críticos donde $\delta d(y_i;\mu)/\delta\mu = 0$.

2.4. Distribución ZOIP en el paquete ZOIP

En esta sección se presenta el paquete **ZOIP** de R alojado en `GitHub` y creado por los autores para analizar datos proporcionales inflados con ceros y/o unos y ajustar una distribución ZOIP.

2.4.1. Instalación

Para acceder a la última versión del paquete **ZOIP**, se encuentra ubicada en `GitHub`, el cual es un alojamiento de repositorios Git, para obtener dicha versión es necesario ejecutar el siguiente código que instala el paquete **devtools**, que es necesario para descargar el paquete **ZOIP** y otros paquetes complementarios, para el correcto funcionamiento del paquete.

```
if (!require("devtools")) install.packages("devtools")
if (!require("rmutil")) install.packages("rmutil")
if (!require("boot")) install.packages("boot")
if (!require("numDeriv")) install.packages("numDeriv")
if (!require("GHQp")) install.packages("GHQp")
devtools::install_github("jucdiaz/ZOIP", force = TRUE)
library(ZOIP) # Carga el paquete
```

2.4.2. Funciones sobre distribución ZOIP

En el paquete **ZOIP** existen cuatro funciones llamadas `dZOIP`, `pZOIP`, `qZOIP` y `rZOIP` el cual corresponden a las funciones de densidad de probabilidad, la función de distribución acumulada, la función cuantil y la función generadora de números aleatorios de la distribución ZOIP, respectivamente; en el siguiente código se observa como se halla la densidad de probabilidad en el punto 0.5 de una distribución ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) descrita como ZOIP($\mu = 0.2$, $\sigma = 0.5$, $p_0 = 0.2$, $p_1 = 0.2$)

```
dZOIP(x = 0.5, mu = 0.2, sigma = 0.5, p0 = 0.2, p1 = 0.2, family = "R-S")
## [1] 0.3243543
```

Además se halla la probabilidad acumulada hasta el punto 0.5 de una distribución OIP-beta con parametrización Ferrari y Cribari-Neto (2004) dada por $ZOIP(\mu = 0.2, \sigma = 3, p_0 = 0, p_1 = 0.2)$

```
pZOIP(q = 0.5, mu = 0.2, sigma = 3, p0 = 0, p1 = 0.2, family = "F-C")
## [1] 0.7181223
```

Se calcula el percentil en el punto 0.7 de una distribución ZIP-beta original dada por $ZOIP(\mu = 0.6, \sigma = 2.4, p_0 = 0.2, p_1 = 0)$

```
qZOIP(p = 0.7, mu = 0.6, sigma = 2.4, p0 = 0.2, p1 = 0,
    family = "Original")
## [1] 0.2061418
```

Por último se generaron 8 valores aleatorios de una distribución ZOIP-simplex descrita como $ZOIP(\mu = 0.6, \sigma = 3, p_0 = 0.2, p_1 = 0.2)$. La función `set.seed` sirve para garantizar la repetición de los valores aleatorios generados en el ejemplo.

```
set.seed(12345)
rZOIP(n = 8, mu = 0.2, sigma = 3, p0 = 0.2, p1 = 0.2, family = "Simplex")
## [1] 0.3185479 1.0000000 0.3765073 1.0000000 0.1626598
## [6] 0.0000000 0.1138673 0.1840670
```

2.4.3. Función RM.ZOIP

La función `RM.ZOIP` estima los parámetros de una distribución ZOIP, vía máxima verosimilitud utilizando el optimizador deseado (`nlminb`, `optim`). La estructura de la función `RM.ZOIP` es la siguiente:

```
RM.ZOIP(formula.mu, formula.sigma = ~1, formula.p0 = ~1,
    formula.p1 = ~1, data, link = c("identity", "identity",
        "identity", "identity"), family = "R-S", optimizer = "nlminb")
```

Los argumentos de la función `RM.ZOIP` son:

- `formula.mu`: Formula que define la función de regresión para el parámetro μ, Para ajustar una distibución ZOIP debe tomar el valor de `y` $\sim$ `1`, donde y es la variable a ajustar.
- `formula.sigma`: Formula que define la función de regresión para el parámetro σ, Para ajustar una distibución ZOIP debe tomar el valor de $\sim$ `1`.
- `formula.p0`: Formula que define la función de regresión para el parámetro p_0, Para ajustar una distibución ZOIP debe tomar el valor de $\sim$`1`.
- `formula.p1`: Formula que define la función de regresión para el parámetro p_1,Para ajustar una distibución ZOIP debe tomar el valor de $\sim$`1`.
- `data`: Es el conjunto de datos en formato `data.frame` donde debe contener los datos de la variable a ajustar y el nombre debe ser la tal cual como está en las fórmula para el parámetro μ.
- `family`: Elección de la distribución ZOIP deseada para ajustar, si toma el valor de `''R-S''` se utilizará la distribución ZOIP-beta con parametrización Rigby y Stasinopoulos (2005), si toma el valor de `''F-C''` se utilizará la distribución ZOIP-beta parametrización Ferrari y Cribari-Neto (2004), el valor de `''Original''` se utilizará la distribución ZOIP-beta con parametrización original, `''Simplex''` utilizará la distribución ZOIP-simplex.
- `link`: Es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y formula. Para ajustar una distribuci ión ZOIP se debe utilizar como función enlace la opción `identity` en sus cuatro parámetros, independientemente de la distribuci ión ZOIP escogida, en familia, por defecto `link=c(''identity'',''identity'',''identity'',''identity'')`.

- `optimizer`: Elección del optimizador, utilizado para encontrar la convergencia de la máxima verosimilitud. se puede elegir el valor de ‘‘`nlminb`’’ o ‘‘`optim`’’, por defecto ‘‘`nlminb`’’

En el siguiente ejemplo se mostrara el ajuste de una distribución ZOIP, para ello mostraremos la salida de la función `RM.ZOIP` de 1000 observaciones simuladas para la distribución ZOIP-beta parametrización Rigby y Stasinopoulos (2005).

```
yi <- data.frame(yi = rZOIP(n = 1000, mu = 0.6, sigma = 0.2,
    p0 = 0.03, p1 = 0.05, family = "R-S"))

mod <- RM.ZOIP(formula.mu = yi ~ 1, formula.sigma = ~1,
    formula.p0 = ~1, formula.p1 = ~1, data = yi, family = "R-S")
mod
```

```
## Call:
## RM.ZOIP(formula.mu = yi ~ 1, formula.sigma = ~1, formula.p0 = ~1,
##     formula.p1 = ~1, data = yi, family = "R-S")
##
##  Results:
##
##  Estimated coefficients for h(mu):
## (Intercept)
##    0.605174
##
##  Estimated coefficients for h(sigma):
## (Intercept)
##   0.2038938
##
##  Estimated coefficients for h(p0):
## (Intercept)
##  0.03000002
##
##  Estimated coefficients for h(p1):
## (Intercept)
##        0.05
```

```
##
##  Convergence
## [1] 0
##
##  message
## [1] "relative convergence (4)"
##
##  iterations
## [1] 22
##
##  Log-likelihood
## [1] 488.2683
```

En el anterior resultado se obtienen varios aspectos importantes de la salida del ajuste de la distribución y leyendo de arriba hacia abajo, primero que todo nos muestra la distribución ajustada, luego para cada el valor ajustado para cada parámetro de la distribución ZOIP, luego un indicador de convergencia del ajuste, donde 0 indica la convergencia, después un mensaje sobre la convergencia (resultados heredados de la función `nlimnb`), después se encuentra el número de iteraciones que fueron necesarias para que convergiera el ajuste de la distribución, por último se encuentra valor de la log-verosimilitud que nos permitirá hacer comparaciones entre ajustes de distribuciones.

Al aplicar a la distribución ajustada (`mod`) la función `summary` se obtiene el siguiente resultado:

```
summary(mod)
## ------------------------------------------------------------
## Fixed effects for identity(mu)
## ------------------------------------------------------------
##              Estimate Std. Error z value  Pr(>|z|)
## (intercept) 0.6066914  0.0031636  191.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
## ------------------------------------------------------------
## Fixed effects for identity(sigma)
## ------------------------------------------------------------
##             Estimate Std. Error z value  Pr(>|z|)
## (intercept) 0.196643   0.004322  45.498 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for identity(p0)
## ------------------------------------------------------------
##              Estimate Std. Error z value Pr(>|z|)
## (intercept) 0.0339992  0.0057308  5.9327 2.98e-09 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for identity(p1)
## ------------------------------------------------------------
##              Estimate Std. Error z value  Pr(>|z|)
## (intercept) 0.0450005  0.0065556  6.8644 6.675e-12 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## ------------------------------------------------------------
```

Con la función `summary` aplicada a la distribución ZOIP ajustada, se obtiene más detalles de los parámetros estimados, primero se obtiene el valor estimado (Estimate), su error estándar (Std.Error), el valor Z del estimador (z value) y el valor p que nos indicara la significancia del parámetro estimado (`Pr(>|z|)`).

En el resultado anterior se obtienen los valores de $\hat{\mu} = 0.6066914$, $\hat{\sigma} = 0.196643$, $\hat{p}_0 = 0.0339992$ y $\hat{p}_1 = 0.0450005$, que son a los parámetros con los que se simulo y_i. Además cabe resaltar que en la función `RM.ZOIP` para ajustar distribuciones de probabilidad no es necesario colocar funciones de enlace ni espacio de busqueda de los parámetros, ya que estos son introducidas automáticamente de acuerdo a el valor tomado en `family`.

Así como la función `summary` puede ser aplicada a un objeto de la clase `ZOIP` o `ZOIPM`, en este trabajo se implementaron otros tipo de funciones asociadas a métodos S3 de R a objetos de clase `ZOIP` o clase `ZOIPM`, tales como la función `print` y `coef`, que permiten mostrar los resultados del modelo ajustado en general y los parámetros regresores estimados, respetivamente.

2.5. Aplicación

En esta sección se muestran varios resultados sobre el ajuste de una distribución ZOIP, primero se realizó un estudio de simulación para observar la convergencia de la estimación de los parámetros de la distribución, y en segunda instancia se ajustó una distribución ZOIP a datos reales sobre la utilización de una tarjeta de crédito de una entidad financiera.

2.5.1. Datos simulados

En este estudio de simulación se analizan diferentes aspectos de la capacidad de estimación que tiene el método de máxima verosimilitud sobre los parámetros de la distribución ZOIP. Se generaron muestras de una distribución ZOIP bajo las diferentes distribuciones y parámetrizaciones con tamaños de muestra n de: 5, 10, 15 y así sucesivamente hasta 500, y se realizaron 1000 réplicas para cada tamaño de muestra, posteriormente se calculó la mediana de cada una de las estimaciones de los parámetros, y así poder analizar la capacidad de convergencia de las metodologías implementadas en la distribución ZOIP y en el paquete **ZOIP**.

En el primer escenario del estudio de simulación se generaron los datos de una distribución ZOIP-beta($\mu = 0.6, \sigma = 0.2, p_0 = 0.03, p_1 = 0.05$) para el caso de la parametrización de Rigby y Stasinopoulos (2005), ZOIP-beta($\mu = 0.6, \sigma = 24, p_0 = 0.03, p_1 = 0.05$) para el caso de la parametrización de Ferrari y Cribari-Neto (2004), ZOIP-beta($\mu = 14.4, \sigma = 9.6, p_0 = 0.03, p_1 = 0.05$) en la parametrización original, cabe aclarar que las tres parametrizaciones anteriores generan exactamente la misma distribución, esto gracias a las ecuaciones definidas en (2-7), (2-8), (2-12) y (2-13), de igual manera se generó la misma cantidad de datos simulados para la distribución ZOIP-simplex($\mu = 0.4, \sigma = 0.2, p_0 = 0.03, p_1 = 0.05$).

En la Figura **2-2** se presentan las medianas de la estimación de los parámetros para cada tamaño de muestra, de esta figura se observa que independientemente de la distribución y parametrización escogida en la distribución ZOIP, todos las estimaciones convergen al valor verdadero del parámetro a medida que aumenta el tamaño de muestra n, de la Figura **2-2** se nota que las estimaciones de σ cuando son parámetros con significado de dispersión como es en la distribución beta con parametrización Rigby y Stasinopoulos (2005) y en la distribución simplex, tienden a dar valores subestimados, por otra parte, en las distribuciones que σ tiene significado de forma y precisión tienden a dar valores sobrestimados. Se observa que las estimaciones de los parámetros de inflación a pesar de que son pequeñas dan resultados muy satisfactorios y casi sin variación en su forma de estimación de distribución a distribución.

Como medida global del proceso de estimación se eligió el MAPE (Error porcentual absoluto medio. $(\sum_{i=1}^{n} |y_i - \hat{y}_i/y_i|)/n$) debido a los cambios de escala entre los diferentes parámetros de las diferentes distribuciones y parametrizaciones. Esta media se realizó como un promedio

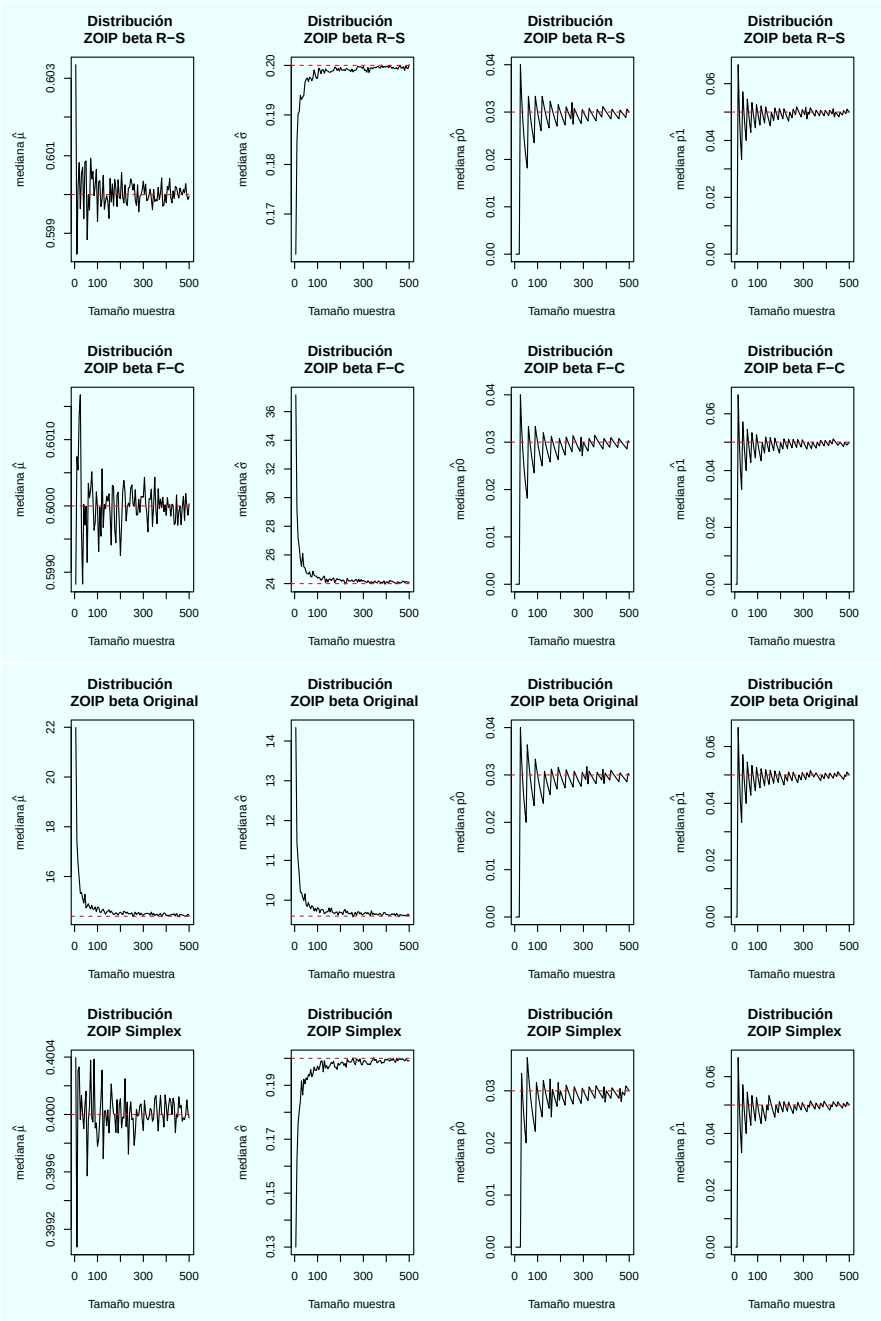

Figura 2-2: Mediana de los parámetros estimados en el escenario 1 para distintas parametrizaciones y valores de n, las líneas rojas representan el verdadero valor del parámetro.

de los MAPES generados por cada uno de los parámetros de la distribución ZOIP en cada tamaño de muestra. En la Figura **2-4**a se presenta el MAPE para las diferentes distribuciones y parametrizaciones estimadas, se observa como a medida que el tamaño de muestra aumenta, el MAPE va decreciendo rápidamente, aunque después de un tamaño de muestra de 200, el MAPE decrece de una manera más lenta, además los errores de estimación son muy parecidos entre los cuatro casos de simulación, la estimación sobre los parámetros de la distribución ZOIP-simplex tiene un error un poco más grande, pero no es significativo sobre los demás casos.

En el segundo escenario de simulación se generó el mismo ejercicio de simulación anterior sobre las mismas distribuciones y parametrizaciones, solo que los valores de p_0 y p_1 cambian por 0.3 y 0.2, respectivamente. Dando así que el 50 % de los datos se vean contaminados por ceros y unos, esto para ver si de alguna forma afecta el aumento de la presencia de ceros y unos sobre las estimaciones de los parámetros de la parte continua de la distribución ZOIP. En la Figura **2-3** se presentan las estimaciones de los parámetros de la simulación con inflaciones al 50 % para diferentes tamaños de muestras, en general se observa que no se ven cambios muy significativos sobre la Figura **2-2** en los parámetros de μ y σ, sin embargo, en la estimación de p_0 se tienden a dar valores subestimados con relación al estudio de simulación anterior y con el parámetro p_1 aunque las estimaciones son muy acertadas sobre el valor real desde tamaños de muestra pequeños, en algunas ocasiones se producen pequeñas perturbaciones no muy alejados del valor real.

Parámetro	Caso	MAPE escenario 1 %	MAPE escenario 2 %
μ	Caso R-S	0.61	0.86
	Caso F-C	0.50	0.85
	Caso original	0.53	0.70
	Caso simplex	0.47	0.63
σ	Caso R-S	2.53	3.40
	Caso F-C	5.10	6.90
	Caso original	5.30	6.98
	Caso simplex	5.30	7.37
p_0	Caso R-S	20.5	5.36
	Caso F-C	19.7	5.42
	Caso original	19.8	5.43
	Caso simplex	20.8	5.51
p_1	Caso R-S	15.2	7.28
	Caso F-C	16.0	7.00
	Caso original	15.7	7.00
	Caso simplex	16.2	7.12
	Promedio	10.57	5.26

Tabla 2-1: MAPE de las estimaciones para cada parámetro en diferentes parametrizaciones en los dos estudios de simulación.

En la Figura **2-4**b se presenta el MAPE para el estudio de simulación del escenario 2, se puede ver como se obtienen MAPES muy parecidos a los del estudio de simulación del escenario 1, pero cabe resaltar como se comete menos error sobre la estimación de los parámetros de

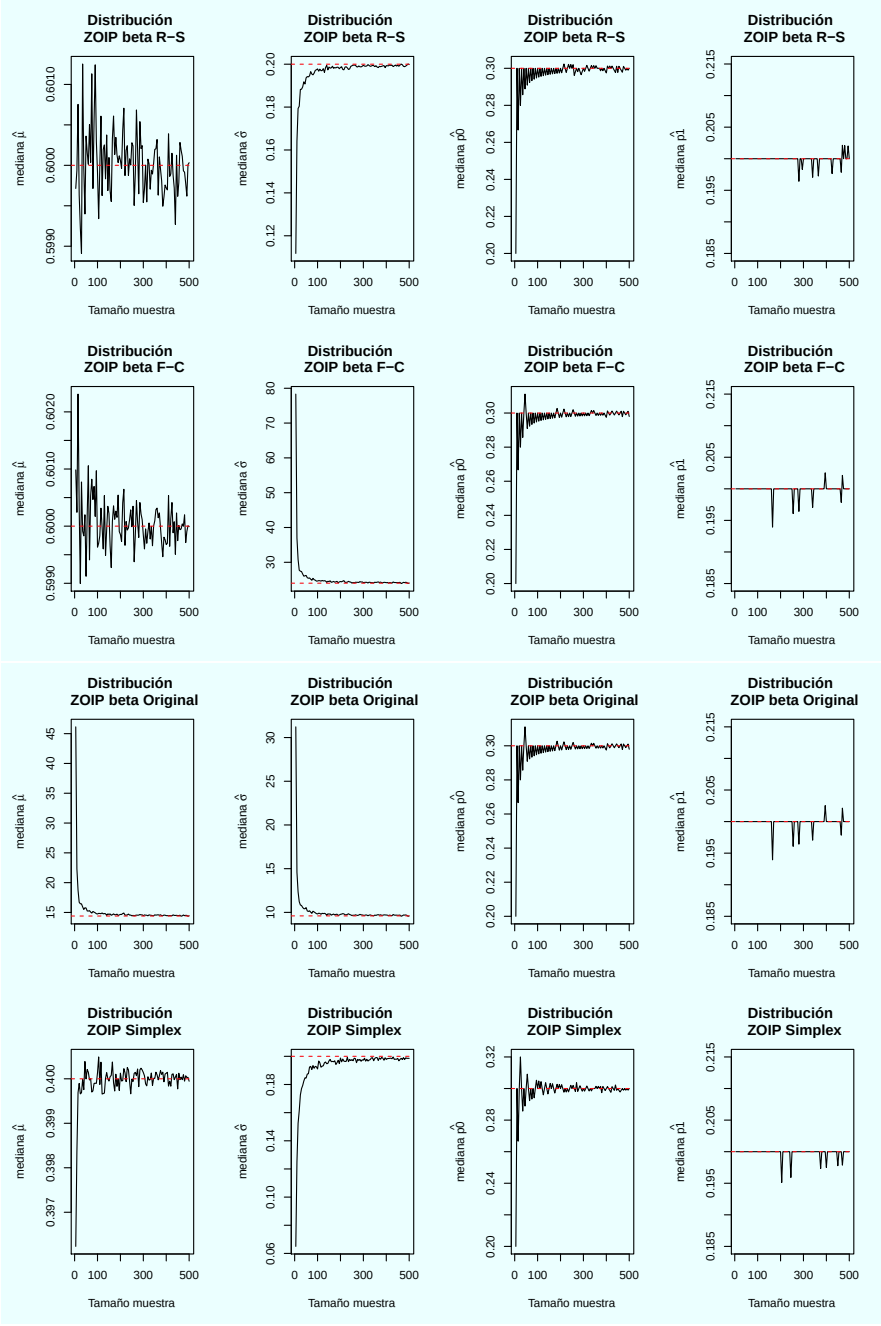

Figura 2-3: Simulación de distribución ZOIP para distintas parámetrizaciones con parámetros de inflación grandes, distribuciones y valores de n.

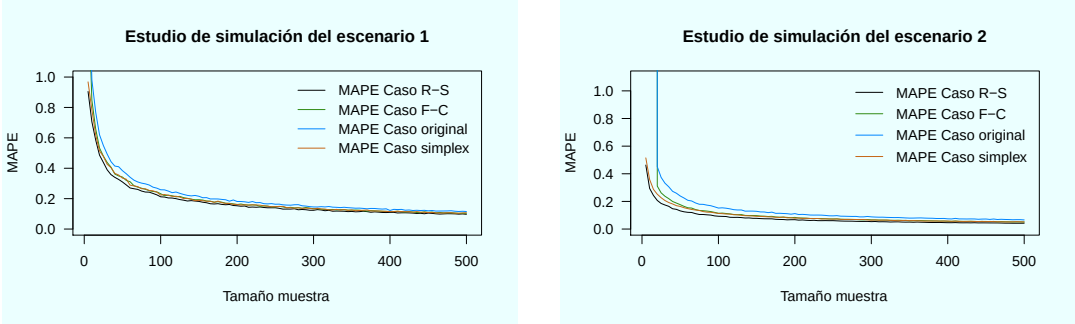

Figura 2-4: MAPE (Error porcentual absoluto medio) para los dos escenarios de simulación y para distintas parametrizaciones y valores de n.

la distribución beta con parametrización Rigby y Stasinopoulos (2005). En la tabla **2-1** se presenta el MAPE para cada parámetro de cada parametrización para ambos estudios de simulación, es claro ver como en general el estudio de simulación del escenario 2 produce un MAPE menor que el del escenario 1, esto es causado por que en el escenario 1 de simulación los errores de pronóstico son más grandes en los parámetros de inflación que en el escenario 2. Por todo lo visto anteriormente se puede concluir que el crecimiento de los parámetros de inflación no afecta de manera significativa la estimación de los parámetros de la parte continua de la distribución ZOIP, pero si en una mejor estimación de los parámetros de inflación.

2.5.2. Datos reales

En esta sección se presenta el ajuste de una distribución ZOIP a datos reales sobre la utilización de una tarjeta de crédito en un banco, para una entidad financiera grande como un banco es de vital importancia conocer el comportamiento del porcentaje de utilización de sus tarjetas de crédito (tdc), se define a y como el porcentaje de uso de una tdc, en la Figura **2-5** se presenta el histograma del porcentaje de utilización de las tdc y es claro notar que y se encuentra entre cero y uno, pero adicional es muy común ver que las tdc no sean utilizadas $(y = 0)$ y también que las tdc sean utilizadas en la totalidad de su cupo asignado $(y = 1)$, por lo que se trata a y como una variable aleatoria perteneciente a datos proporcionales inflados con ceros y unos. Se tiene un total de 9206 tdc, que representan el porcentaje de utilización de las tdc para un trimestre del año 2014 del banco. Se quiere estudiar el ajuste de una distribución ZOIP, para ello se utiliza el paquete en R llamado **ZOIP** mediante su función `RM.ZOIP`.

En la Tabla **2-2** se muestran resultados de los cuatro parámetros estimados vía máxima verosimilitud para la distribución ZOIP, en ellas vemos cómo cambian los valores de los parámetros según la parametrización escogida, los valores de log-verosimilitud no indican

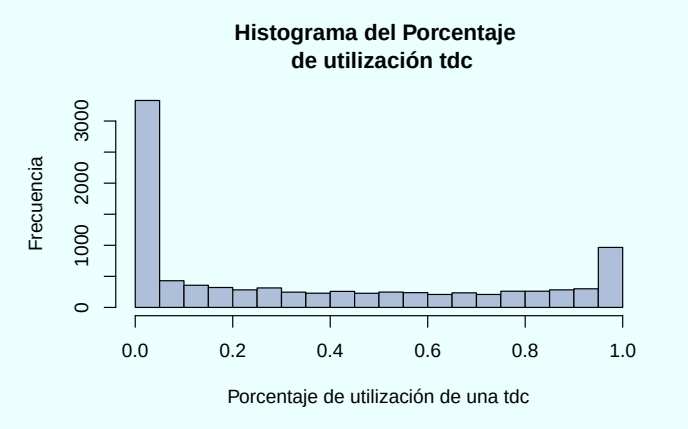

Figura 2-5: Histograma del porcentaje de utilización de las tdc en un banco.

Familia	Parámetro	Estimación	Error estándar	Valor P	Log-Verosimilitud
R-S	μ	0.4040	0.0037	$< 2.2e^{-16}$	5854.067
	σ	0.6601	0.0027	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
F-C	μ	0.4040	0.0037	$< 2.2e^{-16}$	5854.067
	σ	0.4040	0.0037	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
original	μ	0.5233	0.0080	$< 2.2e^{-16}$	5854.067
	σ	0.7719	0.0130	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
simplex	μ	0.5741	0.0010	$< 2.2e^{-16}$	54425.63
	σ	4885.44	18.2430	$< 2.2e^{-16}$	
	p_0	0.1497	0.0032	$< 2.2e^{-16}$	
	p_1	0.0090	0.0004	$< 2.2e^{-16}$	

Tabla 2-2: Ajuste de diferentes distribuciones ZOIP en el porcentaje de utilización de una tdc.

que el mejor modelo ajustado es un ZOIP-beta, ya que es bastante menor el valor de log-verosimilitud de una distribución ZOIP-simplex, además que en las estimaciones de los parámetros de la distribución ZOIP-simplex no se tuvo una convergencia, por lo tanto los valores son muy distintos para el parámetro de dispersión a los vistos en la distribución ZOIP-beta, inclusive muy elevados. Además, el valor de μ es mayor que las de la parametrización en Rigby y Stasinopoulos (2005) y Ferrari y Cribari-Neto (2004), un 17 % más.

En la Figura **2-6** se presenta gráficamente el ajuste de la distribución ZOIP para diferentes parametrizaciones al porcentaje de utilización de las tdc, la línea azul que representa la distribución ZOIP ajustada, es de notar que dicha línea azul es exactamente igual en las tres ocasiones que se ajusta la distribución ZOIP-beta y se ve como sigue el comportamiento

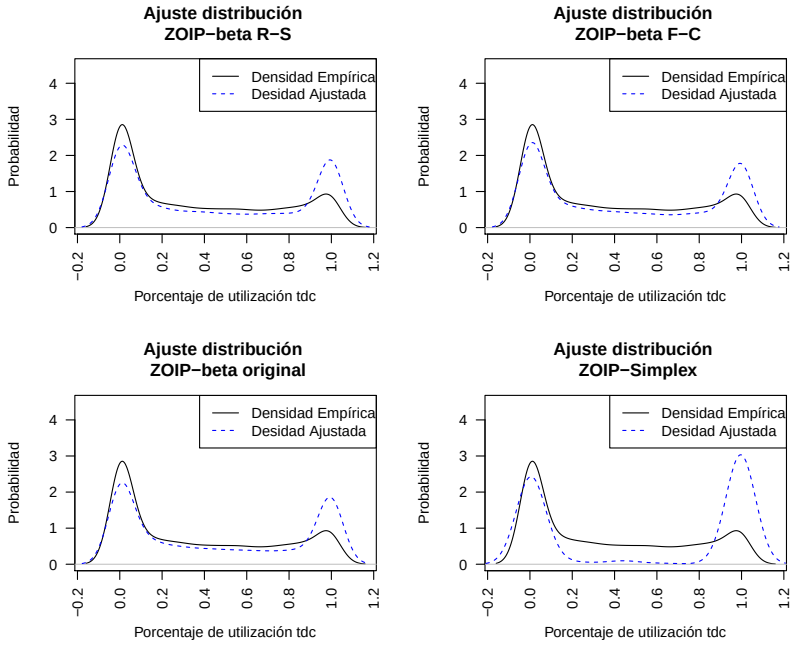

Figura 2-6: Ajuste de diferentes distribuciones y parametrizaciones ZOIP al porcentaje de utilización de una tdc.

original del porcentaje de utilización de las tdc. Es bueno resaltar también como en la Figura 1d de la Figura **2-6** no se nota un buen ajuste para los valores entre cero y uno. Por todo anterior es recomendable decir que el porcentaje de utilización de las tdc de este banco se comportan como una distribución ZOIP-beta con los parámetros descritos en la tabla **2-2**, según la parametrización deseada y no como una distribución ZOIP-simplex.

2.6. Conclusiones

La distribución ZOIP y el paquete **ZOIP** de R permiten ajustar distribuciones para datos provenientes de porcentajes, tasas o proporciones que se encuentren inflados con ceros y/o unos, dicha distribución está compuesta por cuatro parámetros, que son estimados vía máxima verosimilitud y en el cual de acuerdo a los estudios de simulación realizados estos convergen a los valores reales con un tamaño de muestra relativamente pequeño, además se observa como la estimación de los parámetros de la parte continua no se ven afectados por el aumento de la presencia de ceros y unos en los datos, pero si la estimación de los parámetros de la parte discreta. Por otra parte, se observa como el Ajuste de la distribución ZOIP-beta explica el comportamiento de la distribución del porcentaje de utilización de una tarjeta de crédito en un banco.

La distribución ZOIP y el paquete **ZOIP** de R permiten de una manera muy versátil utilizar y ajustar diferentes parametrizaciones y distribuciones para datos proporcionales. Además permite Utilizar y ajustar distribuciones para datos proporcionales que se encuentran inflados solo con ceros o solo con unos, de una manera práctica.

3 Capítulo 3: Modelo de regresión ZOIP con efectos fijos

En muchos casos de estudios es factible preguntarse cómo puede ser explicada una variable aleatoria proveniente de datos proporcionales a partir de diferentes variables, es decir un modelo de regresión para datos proporcionales, el modelo más conocido en la literatura para este tipo de datos es la regresión beta, donde Paolino (2001) estima mediante máxima verosimilitud modelos de variables dependientes de una distribución beta, utilizando la parametrización original, más adelante Ferrari y Cribari-Neto (2004) reparametrizan la distribución e introducen la regresión beta bajo esta nueva parametrización, más adelante en el paquete **betareg** de R (Zeileis et al., 2010) implementan dicha regresión. Por otro parte Rigby y Stasinopoulos (2005) también realizan otra reparametrización de la distribución beta original, basado en parámetros como la media y la dispersión, además introducen un modelo de regresión beta basado en dicha distribución y lo implementan en el paquete **gamlss** de R, sin embargo, existen otros tipos de regresiones basadas en otras distribuciones, como la regresión simplex, que se encuentra bajo la distribución para datos proporcionales simplex (Jørgensen y Barndorff-Nielsen, 1991), dicha regresión fue realizada por Qiu et al. (2008) e implementada en el paquete **simplexreg** de R (Zhang et al., 2016).

Sin embargo, los anteriores modelos de regresión son realizados para datos proporcionales no inflados con ceros o unos, es por esto que Ospina y Ferrari (2012) realizan un modelo de regresión inflado con cero o con uno, no con ambos, bajo la distribución beta inflada de Ospina y Ferrari (2010) con parametrización Ferrari y Cribari-Neto (2004), de igual manera Rigby y Stasinopoulos (2005) implementan los modelos de regresión beta inflados en ceros y/o unos, y se encuentran implementados en el paquete **gamlss** de R (Rigby y Stasinopoulos, 2008), sin embargo, para la utilización del modelo de regresión inflado solo con ceros o unos o con ambos, se deben utilizar funciones distintas dentro del paquete para ajustar los tres diferentes modelos de regresión. Además no existen paquetes en R que logren ajustar un modelo de regresión beta inflado con ceros y/o unos bajo las parametrizaciones originales y de Ferrari y Cribari-Neto (2004), por otra parte a pesar de que existen desarrollos teóricos sobre el modelo de regresión simplex inflado con ceros y/o unos (Galvis y Lachos, 2014), no existe un paquete en R que permita realizar un ajuste sobre dicho modelo de regresión.

Es por esto que en este trabajo se implementa de manera teórica y de forma práctica me-

diante el paquete **ZOIP** en el sistema de computación R (R Core Team, 2017) y disponible en el repositorio web `GitHub`, un modelo de regresión para datos proporcionales inflados con ceros y/o unos (Modelo de regresión ZOIP) que permita mediante una misma función ajustar modelos en diferentes distribuciones para datos proporcionales y en diferentes parametrizaciones.

Este capítulo se encuentra organizado de la siguiente manera: primero se presenta el modelo de regresión ZOIP que es basado en la distribucion ZOIP visto en el capitulo anterior y su debida estimación, mediante maxima verosimilitud, en la siguiente sección se presenta la implementación del modelo de regresión ZOIP en el paquete **ZOIP** de R y por último se presentan unas aplicaciones a datos simulados y a datos reales.

3.1. Modelo de regresión ZOIP

Una clase general de modelos de regresión ZOIP puede definirse como sigue. Sea $y_1, y_2, \ldots, y_n$ variables aleatorias independientes tal que cada y_i, para $i = 1, \ldots, n$, tiene función de densidad de probabilidad (2-17) con parámetros $\mu = \mu_i$, $\sigma = \sigma_i$, $p_0 = p_{0i}$, y $p1 = p_{1i}$. Se asume que μ_i, σ_i, p_{0i} y p_{1i} se definen como

$$\begin{aligned} h_1(\mu_i) &= \mathbf{x}_{i1}^\top \boldsymbol{\beta}_1, \\ h_2(\sigma_i) &= \mathbf{x}_{i2}^\top \boldsymbol{\beta}_2, \\ h_3(p_{0i}) &= \mathbf{x}_{i3}^\top \boldsymbol{\beta}_3, \\ h_4(p_{1i}) &= \mathbf{x}_{i4}^\top \boldsymbol{\beta}_4 \end{aligned} \tag{3-1}$$

donde $\mathbf{x}_{i1} = (x_{i11}, x_{i12}, \ldots, x_{i1k_1})$, $\mathbf{x}_{i2} = (x_{i21}, x_{i22}, \ldots, x_{i2k_2})$, $\mathbf{x}_{i3} = (x_{i31}, x_{i32}, \ldots, x_{i1k_3})$ y $\mathbf{x}_{i4} = (x_{i41}, x_{i42}, \ldots, x_{i1k_4})$, son vectores de covariables conocidos de dimensión k_1, k_2, k_3 y k_4 respectivamente, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1k_1})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2k_2})$, $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \ldots, \beta_{3k_3})$ y $\boldsymbol{\beta}_4 = (\beta_{41}, \beta_{42}, \ldots, \beta_{4k_4})$ son vectores de parámetros de regresión desconocidos. Además se asume que las funciones de enlace $h_1(\cdot)$, $h_2(\cdot)$, $h_3(\cdot)$ y $h_4(\cdot)$ son conocidas y apropiadas para mapear de los reales a los valores admisibles del parámetro, ademas son funciones estrictamente monótonas y doblemente diferenciables. Las posibles funciones para el parámetro μ y σ son logit, probit, clog-log, o log dependiendo de la parametrización, para los parámetros de inflación p_0 y p_1 son posibles funciones de enlace como logit, probit, clog-log. Estudios sobre funciones enlace mal especificadas sobre modelos de regresión beta se encuentran en de Andrade (2007).

3.1.1. Inferencia estadística

Para estimar los parámetros del modelo de regresión ZOIP, se usará el método de máxima verosimilitud. La función de verosimilitud para $\boldsymbol{\theta} = (\boldsymbol{\beta_1}^\top, \boldsymbol{\beta_2}^\top, \boldsymbol{\beta_3}^\top, \boldsymbol{\beta_4}^\top)^\top$, basado en una muestra de observaciones independientes, es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\mathbf{y}_i; \mu_i, \sigma_i, p_{0i}, p_{1i}) \tag{3-2}$$

donde para el caso de ZOIP-beta original $\mu_i = p_i$, $\sigma_i = q_i$; si la distribución ZOIP-beta fuese con parametrización de Ferrari y Cribari-Neto (2004) el único parámetro que cambiaría es $\sigma_i = \phi_i$, el resto de los parámetros no tendrían modificaciones según su parametrización o distribución.

La función de verosimilitud definida en (3-2) al aplicar logaritmo natural se obtiene la función de log verosimilitud definida como:

$$\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\beta_3}) + \ell_2(\boldsymbol{\beta_4}) + \ell_3(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$$

Note que la función de verosimilitud es factorizada en tres términos, dos de ellos del componente discreto y uno compuesto por $\boldsymbol{\beta_1}$ y $\boldsymbol{\beta_2}$ del componente continuo, por tanto los parámetros son separables (Pace y Salvan, 1997), así la máxima verosimilitud puede ser tratada por separado y por lo tanto:

$$\ell_1(\boldsymbol{\beta_3}) = \sum_{i=1}^{n} p_{0i}^{S_0(y_i)} (1 - p_{0i})^{1 - S_0(y_i)}$$

$$\ell_2(\boldsymbol{\beta_4}) = \sum_{i=1}^{n} p_{1i}^{S_1(y_i)} (1 - p_{1i})^{1 - S_1(y_i)}$$

$$\ell_3(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = \sum_{i=1:y_i \in (0,1)}^{n} f(y_i; \mu_i, \sigma_i)$$

Con

$$S_j(y_i) = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{si } y_i \neq j \end{cases} \quad ; \quad j = 1, 2$$

Con $p_{0i} = h_3^{-1}(\mathbf{x}_{i3}^\top \boldsymbol{\beta}_3)$, $p_{1i} = h_4^{-1}(\mathbf{x}_{i4}^\top \boldsymbol{\beta}_4)$, $\mu_i = h_1^{-1}(\mathbf{x}_{i1}^\top \boldsymbol{\beta}_1)$ y $\sigma_i = h_2^{-1}(\mathbf{x}_{i2}^\top \boldsymbol{\beta}_2)$ como se definio en (3-1). La funcion de verosimilitud depende de tres terminos, el primero depende de $\boldsymbol{\beta}_3$

(componente discreto para inflación en cero), el segundo de $\boldsymbol{\beta}_4$ (componente discreto para explicar la inflación en uno) y el tercero depende de $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ (Componentes para explicar la parte continua), por lo tanto los parámetros son separables y la inferencia de máxima verosimilitud para $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ se puede hacer por separado de la de $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$, como si conociera a $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$ y viceversa. (Ospina y Ferrari, 2012).

No existen expresiones que den una solución cerrada analíticamente para encontrar los máximos de las funciones de log verosimilitudes descritas anteriormente, para así hallar los estimadores de máxima verosimilitud de los parámetros de regresión de cada uno de los componentes de la distribución ZOIP. Por lo que es necesario utilizar algoritmos de optimización no lineal como el método de Newton-Raphson o Fisher's scoring, para nuestro caso utilizaremos el algoritmo de optimización dado por la función `nlminb` o `optim` del paquete `stats` de R e implementado en el paquete **ZOIP** de R para el modelo de regresión ZOIP.

3.2. Modelo de regresión ZOIP en el Paquete ZOIP

En esta sección presentaremos como el paquete **ZOIP** realizado en R ajusta un modelo de regresión ZOIP con efectos fijos, vía máxima verosimilitud.

3.2.1. Función RM.ZOIP

La función `RM.ZOIP` estima los parámetros de un modelo de regresión ZOIP con y sin covariables vía máxima verosimilitud utilizando el optimizador `nlminb` o `optim`. La estructura de la función `RM.ZOIP` es la siguiente:

```
RM.ZOIP(formula.mu, formula.sigma = ~1, formula.p0 = ~1,
    formula.p1 = ~1, data, link = c("identity", "identity",
        "identity", "identity"), family = "R-S", optimizer = "nlminb")
```

Los argumentos de la función `RM.ZOIP` son:

- `formula.mu`: Formula que define la función de regresión para el parámetro μ, Un valor posible es `y ~ x1 + x2`, es necesario definir la variable respuesta (y).
- `formula.sigma`: Formula que define la función de regresión para el parámetro σ, un valor posible es `~ x1`, por defecto `~ 1`.
- `formula.p0`: Formula que define la función de regresión para el parámetro p_0, un valor posible es `~x1`, por defecto `~1`.
- `formula.p1`: Formula que define la función de regresión para el parámetro p_1, un valor posible es `~x1`, por defecto `~1`.
- `data`: Es el conjunto de datos en formato `data.frame` donde debe contener las nombres de las columnas tal cual como están en las fórmulas.

- `family`: Elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP, si toma el valor de ‘‘`R-S`’’ se utilizara la distribución beta con parametrización Rigby y Stasinopoulos (2005), si toma el valor de ‘‘`F-C`’’ se utilizara la distribución beta parametrización Ferrari y Cribari-Neto (2004), el valor de ‘‘`Original`’’ se utilizara la distribución beta con parametrización original, ‘‘`Simplex`’’ utilizará la distribución simplex.
- `link`: Es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y formula. Si el modelo de regresión no posee covariables se debe utilizar como función enlace la opción `identity`, independientemente del valor escogido en familia, opciones posibles son `logit`, `log`, por defecto `link=c(‘‘identity’’,‘‘identity’’,‘‘identity’’,‘‘identity’’)`.
- `optimizer`: Elección del optimizador, utilizado para encontrar la convergencia de la máxima verosimilitud. se puede elegir el valor de ‘‘`nlminb`’’ o ‘‘`optim`’’, por defecto ‘‘`nlminb`’’

En el siguiente ejemplo nos concentraremos en el ajuste de un modelo regresión ZOIP, para ello se mostrará el código utilizado y la salida de la función `RM.ZOIP`, para una variable aleatoria simulada de una distribución ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) y dos covariables simuladas a partir de una distribución uniforme entre cero y uno, el tamaño de la muestra simulada es 1000. Esto replicando exactamente uno de los casos de simulación vistos en la próxima sección.

Primero se simula la variable respuesta a partir de la función `rZOIP` con los debidos valores de los parámetros para cada observación, y las covariables.

```
library(ZOIP)
n <- 1000
x1 <- runif(n)
x2 <- runif(n)

c1 <- 0.2
c2 <- -1
mu_i <- inv.logit(c1 + c2 * x1)

b1 <- 0.3
b2 <- 3
b3 <- 0.9
sigma_i <- inv.logit(b1 + b2 * x1 + b3 * x2)
```

```
d1 <- 0.07
p0_i <- rep(d1, n)

e1 <- 0.02
e2 <- -4
p1_i <- inv.logit(e1 + e2 * x2)

param <- cbind(mu_i, sigma_i, p0_i, p1_i)
y_i <- apply(param, 1, function(x)
{
    rZOIP(1, mu = x[1], sigma = x[2], p0 = x[3], p1 = x[4],
        family = "R-S")
})

data <- as.data.frame(cbind(y_i, x1, x2))

link <- c("logit", "logit", "identity", "logit")
mod <- RM.ZOIP(formula.mu = y_i ~ x1, formula.sigma = ~x1 +
    x2, formula.p0 = ~1, formula.p1 = ~x2, data = data,
    link = link, family = "R-S")

mod
```

Los resultados obtenidos se muestran a continuación.

```
## Call:
## RM.ZOIP(formula.mu = y_i ~ x1, formula.sigma = ~x1 + x2, formula.p0 = ~1,
##     formula.p1 = ~x2, data = data, link = link, family = "R-S")
##
##  Results:
##
##  Estimated coefficients for h(mu):
## (Intercept)          x1
##   0.3395118  -1.4808985
##
```

```
##  Estimated coefficients for h(sigma):
## (Intercept)          x1          x2
##   0.7006753   2.4493508   0.5328350
##
##  Estimated coefficients for h(p0):
## (Intercept)
##  0.08199999
##
##  Estimated coefficients for h(p1):
## (Intercept)          x2
##  -0.1252807  -3.0092159
##
##  Convergence
## [1] 0
##
##  message
## [1] "relative convergence (4)"
##
##  iterations
## [1] 39
##
##  Log-likelihood
## [1] 7688.628
```

En el anterior resultado se obtienen varios aspectos importantes de la salida del modelo y leyendo de arriba hacia abajo, primero que todo nos muestra el modelo ajustado, luego para cada parámetro de la distribución ZOIP los valores estimados para cada uno de los parámetros de regresión asociados a cada covariable, luego un indicador de convergencia, donde 0 indica la convergencia, después un mensaje sobre la convergencia (resultados heredados de la función `nlimnb`), después se encuentra el número de iteraciones que fueron necesarias para que convergiera el modelo, por último se encuentra valor de la log-verosimilitud que nos permitirá hacer comparaciones entre modelos.

Al aplicar al modelo ajustado (`mod`) la función `summary` se obtiene el siguiente resultado:

```
summary(mod)
## ------------------------------------------------------------
## Fixed effects for logit(mu)
## ------------------------------------------------------------
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.339512   0.098754  3.4379 0.0005862 ***
## x1          -1.480898   0.179015 -8.2725 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for logit(sigma)
## ------------------------------------------------------------
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.70068    0.10726  6.5323 6.476e-11 ***
## x1           2.44935    0.13775 17.7812 < 2.2e-16 ***
## x2           0.53283    0.13273  4.0143 5.961e-05 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for identity(p0)
## ------------------------------------------------------------
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 0.0820000  0.0086762  9.4512 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for logit(p1)
## ------------------------------------------------------------
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.12528    0.15084 -0.8305   0.4062
## x2          -3.00922    0.34242 -8.7881   <2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## ------------------------------------------------------------
```

Con la función `summary` aplicada al modelo de regresión ZOIP, se obtiene más detalles de los parámetros regresores estimados para cada parámetro del modelo de regresión ZOIP, se obtiene el valor estimado (Estimate), su error estándar (Std.Error), el valor Z del estimador

(z value) y el valor p que nos indicara la significancia del parámetro estimado (`Pr(>|z|)`), es de notar que para cada parámetro del modelo de regresión ZOIP le es mostrado la respectiva función enlace utilizada y definida al inicio del modelo.

3.3. Aplicación

En esta sección se muestran diferentes resultados sobre el ajuste de un modelo de regresión ZOIP, por medio del paquete **ZOIP**, primero se realizó un estudio de simulación para analizar la convergencia de la estimación de los parámetros regresores de una regresión ZOIP, y en segunda instancia se ajusta un modelo de regresión ZOIP a datos reales, sobre cómo puede ser explicado el porcentaje de utilización de una tarjeta de crédito (tdc) de una entidad financiera con diferentes variables del negocio.

3.3.1. Datos simulados

En el estudio de simulación se analizan diferentes aspectos de la capacidad de estimación que tiene el método de máxima verosimilitud sobre los parámetros regresores de un modelo de regresión ZOIP. Para comprobar esto se generaron muestras pertenecientes a una distribución ZOIP a partir de dos variables aleatorias uniformes cero uno, con tamaños de muestra de 25, 50, 75, y 100 a partir de este punto no se realizarán incrementos de 25 si no de 100, es decir 100, 200, 300 hasta 3500, y se realizaron 1000 réplicas para cada tamaño de muestra, posteriormente se calculó la mediana de la estimación de cada parámetro regresor para cada distribución y parametrización utilizada. A continuación, se muestra la estructura simulada para cada parámetro de la distribución ZOIP.

Si $y_i \sim ZOIP(\mu_i, \sigma_i, p_{0i}, p_{1i})$,

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i}, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \\ h_3(p_{0i}) &= \beta_0, \\ h_4(p_{1i}) &= \beta_0 + \beta_1 x_{2i}, \end{aligned} \tag{3-3}$$

donde para la regresión de μ: $\beta_0 = 0.2$ y $\beta_1 = -1$, para la de σ se escogieron dos escenarios distintos si la regresión a modelar es ZOIP-beta parametrización de Rigby y Stasinopoulos (2005), entonces: $\beta_0 = 0.3$, $\beta_1 = -2$, $\beta_2 = -4$, para las demás parametrización y distribuciones $\beta_0 = 0.3$, $\beta_1 = 3$, $\beta_2 = 0.9$, esto para tener una variabilidad de los datos moderada. Para el parámetro p_0: $\beta_0 = 0.07$ y para p_1: $\beta_0 = 0.02$ y $\beta_1 = -4$ para todos los casos posibles de selección de la regresión ZOIP, y $x_{1i} \sim U(0,1)$, $x_{2i} \sim U(0,1)$. Las funciones de enlace

adecuadas para cada distribución y parametrización se muestran en la tabla **3-1**.

Familia	Parámetro	$h(\cdot)$
R-S	μ	Logit
	σ	Logit
	p_0	NA
	p_1	Logit
F-C	μ	Logit
	σ	Log.
	p_0	NA
	p_1	Logit
original	μ	Log.
	σ	Log.
	p_0	NA
	p_1	Logit
simplex	μ	Logit
	σ	Log.
	p_0	NA
	p_1	Logit

Tabla 3-1: Funciones de enlace adecuadas para cada parámetro, según su distribución y/o parametrización.

En la figura **3-1** se describen los valores estimados para diferentes valores de tamaño de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización de Rigby y Stasinopoulos (2005), en ella se ve como todos los parámetros estimados oscilan alrededor del valor real del parámetro que es representado por la línea roja, sin embargo, se nota como unos parámetros tienen una oscilación mayor que otros, como es el caso de los parámetros de intercepto de la media y el del parámetro de inflación de unos, asociada a p_1. Los de más parámetros convergen rápidamente a sus valores reales, como los parámetros que representan la variabilidad (σ) y el parámetro de p_0.

En la figura **3-2** se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización de Ferrari y Cribari-Neto (2004), en dicha figura se nota como la estimación de los parámetros asociados con la media tienen una oscilación mayor que los demás parámetros, sin embargo, en todos los parámetros se obsserva como a medida que el tamaño de muestra es más grande la oscilación de los parámetros es menor y van convergiendo satisfactoriamente a sus valores reales.

En la figura **3-3** se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización original, se puede ver como con los valores del escenario de simulación elegidos, se obtiene una distribución ZOIP con mayor variabilidad, por lo que los valores de los parámetros asociados a σ tienen una mayor oscilación, sin embargo, este oscila solo en un 0.01 de sus unidades, lo que no es preocupante. Por otra parte, se observa como el parámetro de intercepto del parámetro de inflación de unos (p_1) si oscila mucho más ya que este tiene una desviación estándar de 0.04 en prome-

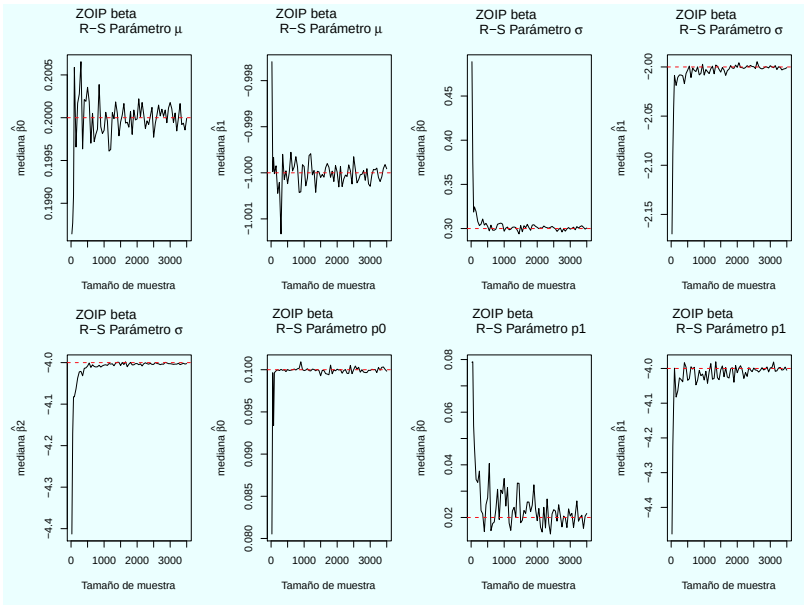

Figura 3-1: Simulación de un modelo de regresión ZOIP-beta para la parametricazion R-S con diferentes valores de n.

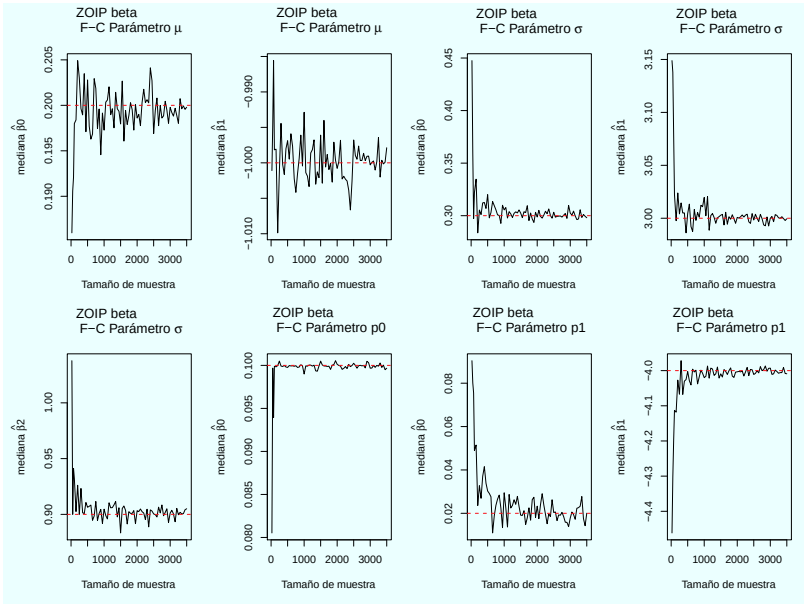

Figura 3-2: Simulación de un modelo de regresión ZOIP-beta para la parametricazion F-C con diferentes valores de n.

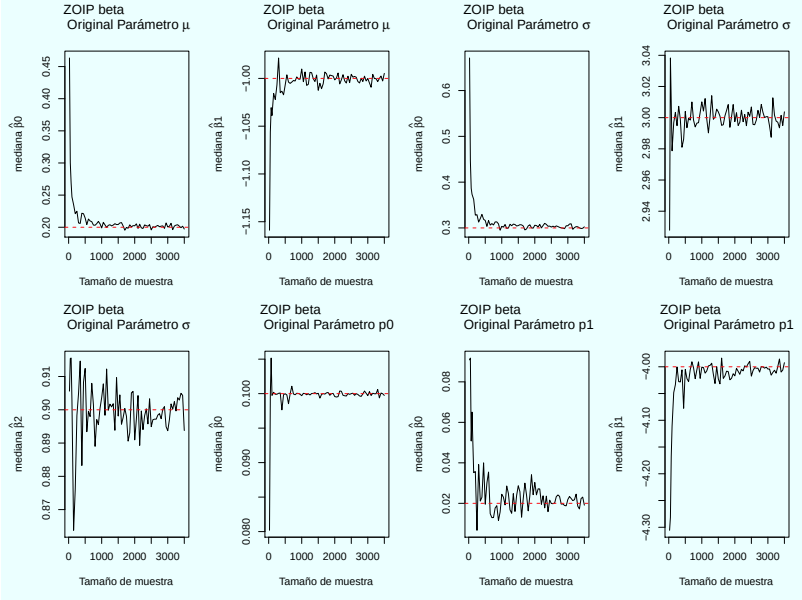

Figura 3-3: Simulación de un modelo de regresión ZOIP-beta para la parametricazion original con diferentes valores de n.

dio, pero se observa como a través de que el tamaño de muestra es mayor la oscilación va disminuyendo, por lo que se sospecha que se necesita un mayor tamaño de muestra para que esta converja con mayor satisfacción.

En la figura **3-4** se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-simplex, Se nota como todos los parámetros oscilan alrededor de los valores verdaderos y como estas oscilaciones se van reduciendo a través de que el tamaño de muestra crece, sin embargo, unos parámetros toman mayor tiempo de convergencia como es el parámetro β_1 asociado al parámetro de dispersión (σ).

Familia	μ		σ			p_0	p_1	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_0$	$\hat{\beta}_1$
R-S	1.25	0.32	1.45	2.55	1.38	4.86	383.09	4.88
F-C	14.22	3.96	22.21	2.9	10.14	4.86	91.21	4.88
original	22.34	8.03	22.55	3.62	8.69	4.84	90.58	4.96
simplex	13.93	5.89	24.49	3.11	11.01	4.85	91.15	4.81

Tabla 3-2: Mediana del MAPE (Error porcentual absoluto medio) en porcentaje para los diferentes parametros en las diferentes parametrizaciones.

En la tabla **3-2** se muestra la mediana del MAPE de los diferentes parámetros regresores para cada posible caso de la distribución o parametrización de la distribución ZOIP, en dicha tabla

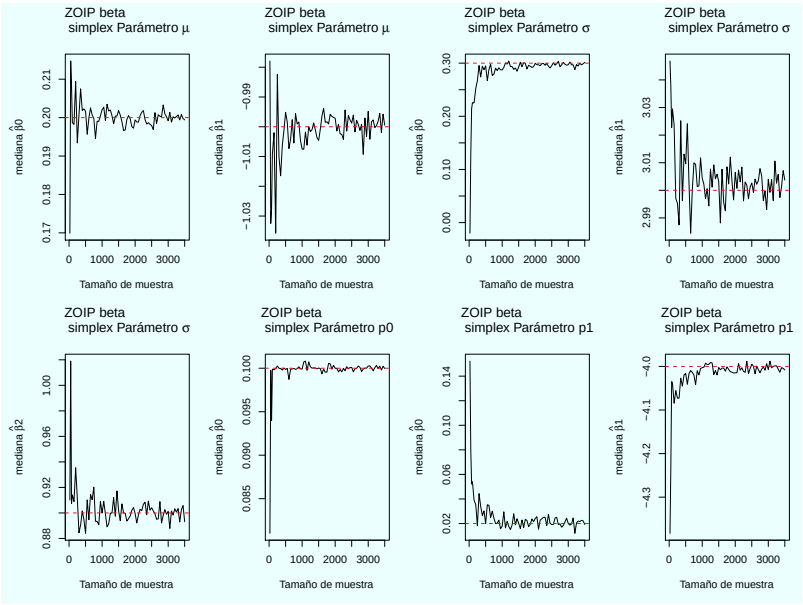

Figura 3-4: Simulación de un modelo de regresión ZOIP-simplex con diferentes valores de n.

se nota como el MAPE en los interceptos de cualquier regresión asociada a los parámetros de la distribución ZOIP son un poco más grandes que los demás parámetros regresores de cada regresión, además se comete un MAPE más grande en las regresiones asociadas a todos los parámetros de inflación, esto nos permite concluir que hallar los parámetros verdaderos en los parámetros de inflación es un poco más difícil que en los parámetros de localización y escala como lo son μ y σ, esto se debe a que se posee una menor cantidad de datos en cero y uno, en este escenario de simulación elegido. Por otro lado, el intercepto asociado a la regresión del parámetro de inflación de los unos posee un MAPE muy grande, por lo que nos permite concluir que a pesar de que los diferentes parámetros estimados en la simulación oscilan alrededor del valor real este todavía tiene una variabilidad muy grande por lo que hace que este MAPE sea grande y el parámetro no haya convergido con un tamaño de muestra de 3500.

En la figura **3-5** se muestra la mediana del MAPE de la mediana del MAPE de todos los parámetros asociados a cada parametrización y distribución para diferentes tamaños de muestra, en ella se evidencia como el caso de la regresión ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) tiene un MAPE menor, donde este tiene asociados unos parámetros distintos con una distribución ZOIP con menor variabilidad, por lo que no es del todo comparable con las demás parametrizaciones y distribuciones, se nota un MAPE menor al 20 % a partir de un tamaño de muestra mayor a 500, por lo que se puede concluir que con un tamaño de muestra mayor a 500 el modelo tendrá un MAPE aceptable para la estimación de todos los parámetros de la regresión ZOIP, sin embargo, esto siempre dependerá de la

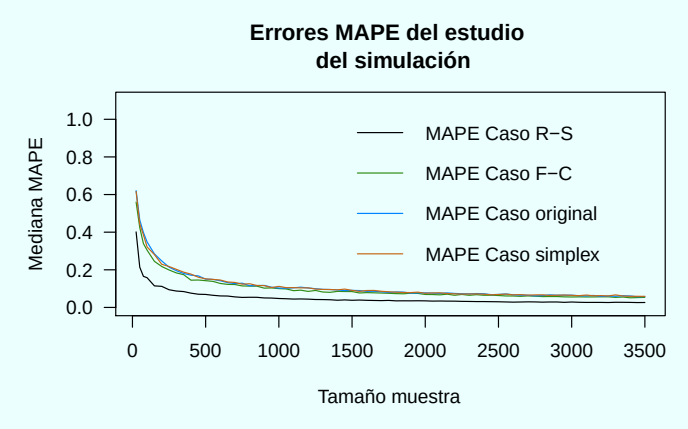

Figura 3-5: Mape (Error porcentual absoluto medio) para modelo de regresion ZOIP simulado para distintas parametrizaciones y valores de n.

variabilidad que posean los datos.

3.3.2. Datos reales

En una entidad financiera tiene gran importancia conocer el comportamiento del porcentaje de utilización de las tarjetas de crédito (tdc), con el fin de conocer el comportamiento de la cartera de tarjeta de crédito, además de detectar los diferentes factores que pueden afectar este tipo de cartera. Se define a y como el porcentaje de uso de una tdc, es de notar que y se encuentra entre cero y uno, pero adicional es normal que se tengan tdc que no sean utilizadas ($y = 0$) y tdc que están utilizadas en la totalidad de su cupo asignado ($y = 1$), por lo que se trata a y como una variable aleatoria perteneciente a datos proporcionales inflados con ceros y unos, es decir y puede ser explicada a partir de una distribución ZOIP. Se tiene un total de 9206 tarjetas de crédito. Se quiere estudiar el impacto de algunas variables sobre el porcentaje de utilización de una tdc, para ello se busca ajustar un modelo de regresión ZOIP mediante la función `RM.ZOIP` del paquete **ZOIP** de R, que permita explicar el comportamiento del porcentaje de utilización de una tdc mediante las siguientes tres variables, *Score:* variable entre cero y 1000 que para nuestro caso se cambiara de escala entre cero y uno, está explica la calificación del comportamiento de pago del cliente asociada a la tdc, que pertenece a la entidad financiera, donde cero es la peor calificacion y uno un comportamiento de pago ideal; *Prom Cuotas:* se define como el promedio de cantidad de cuotas al que ha diferido sus compras en los últimos seis meses; *Cupo tdc Entidad:* es el cupo total asignado a la tdc, esta será tratada como el logaritmo de su cupo más uno, para una mayor estabilidad de su varianza.

En el modelo de regresión ZOIP se deben definir cuatro diferentes modelos de regresión

para ser ajustados, de tal forma que nos permita ver el efecto de las variables descritas anteriormente sobre cada uno de los parámetros de la distribución ZOIP, adicionalmente dependiendo de la parametrización o distribución que se esté utilizando, se debe utilizar una función enlace adecuada para cada regresión, en las ecuaciones dadas en (3-4) se pueden ver los modelos de regresión a aplicar en cada parámetro, en estas ecuaciones se ve que dependen de una función enlace $h(\cdot)$, en la tabla **3-3** se muestran las diferentes funciones enlaces adecuadas para cada parámetro dependiendo de la distribución escogida y/o la parametrización.

Si $y_i \sim ZOIP(\mu_i, \sigma_i, p_{0i}, p_{1i})$,

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \\ h_3(p_{0i}) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \\ h_4(p_{1i}) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \end{aligned} \tag{3-4}$$

donde y_i es porcentaje de utilizacion de la i-esima tdc, x_{1i}: es el valor del score del i-esimo individuo asociada a la tarjeta de crédito, x_{2i}: es valor del promedio de cuotas al que difiere sus compras de la i-esima tarjeta de crédito, x_{3i}: es el valor del cupo otorgado de la i-esima tarjeta de crédito.

Familia	Parámetro	$h(\cdot)$
R-S	μ	Logit
	σ	Logit
	p_0	Logit
	p_1	Logit
F-C	μ	Logit
	σ	Log.
	p_0	Logit
	p_1	Logit
original	μ	Log.
	σ	Log.
	p_0	Logit
	p_1	Logit
simplex	μ	Logit
	σ	Log.
	p_0	Logit
	p_1	Logit

Tabla 3-3: Funciones de enlace adecuadas para cada parámetro, según su distribución y/o parametrización para el modelo de regresión ZOIP en el porcentaje de utilización de una tdc.

En las ecuaciones dadas en (3-4) se puede ver como los parámetros μ, σ, p_0, p_1 serán explicados por las variables score, cupo asignado a su tdc y el promedio al que difiere sus compras, bajo estas ecuaciones podemos observar como se explica el porcentaje utilización de una tdc, la variabilidad de este porcentaje de utilización, el por que un cliente no llega a utilizar

nunca su tdc y adicional y contrariamente a lo anterior, por que un cliente utiliza la total capacidad de su tdc.

Familia	Parámetro	β's	Estimación	Error estándar	Valor P	Log-Verosimilitud	Iteraciones
R-S	μ	$\hat{\beta}_0$	-0.046	0.050	0.3618	-5414.738	125
		$\hat{\beta}_1$	-0.354	0.107	0.0009		
		$\hat{\beta}_2$	0.022	0.002	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	-0.025	0.009	0.0074		
	σ	$\hat{\beta}_0$	0.822	0.038	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	-0.197	0.078	0.0114		
		$\hat{\beta}_2$	-0.006	0.002	0.0013		
		$\hat{\beta}_3$	-0.003	0.007	0.6741		
	p_0	$\hat{\beta}_0$	-1.496	0.101	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.724	0.185	$8.87e^{-5}$		
		$\hat{\beta}_2$	-0.153	0.009	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	0.002	0.015	0.1243		
	p_1	$\hat{\beta}_0$	-1.480	0.095	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	-0.630	0.254	0.0132		
		$\hat{\beta}_2$	0.011	0.006	0.0666		
		$\hat{\beta}_3$	-0.069	0.022	0.0022		

Tabla 3-4: Parametros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) en el porcentaje de utilización de una tdc.

Familia	Parámetro	β's	Estimación	Error estándar	Valor P	Log-Verosimilitud	Iteraciones
F-C	μ	$\hat{\beta}_0$	-0.045	0.050	0.3667	-5414.605	105
		$\hat{\beta}_1$	-0.354	0.107	0.0009		
		$\hat{\beta}_2$	0.022	0.002	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	-0.025	0.009	0.0074		
	σ	$\hat{\beta}_0$	0.068	0.045	0.1286		
		$\hat{\beta}_1$	0.238	0.094	0.0117		
		$\hat{\beta}_2$	0.007	0.002	0.0012		
		$\hat{\beta}_3$	0.003	0.008	0.6854		
	p_0	$\hat{\beta}_0$	-1.496	0.101	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.724	0.185	$8.87e^{-5}$		
		$\hat{\beta}_2$	-0.153	0.009	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	0.002	0.015	0.1243		
	p_1	$\hat{\beta}_0$	-1.480	0.095	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	-0.630	0.254	0.0132		
		$\hat{\beta}_2$	0.011	0.006	0.0666		
		$\hat{\beta}_3$	-0.069	0.022	0.0022		

Tabla 3-5: Parametros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Ferrari y Cribari-Neto (2004) en el porcentaje de utilización de una tdc.

En las tablas **3-4**, **3-5**, **3-6**, **3-7** se tiene información sobre como las covariables influyen en los parámetros de los cuatro diferentes modelos ajustados de la regresión ZOIP, primero se

Familia	Parámetro	β's	Estimación	Error estándar	Valor P	Log-Verosimilitud	Iteraciones
original	μ	$\hat{\beta}_0$	-0.649	0.048	$< 2.2e^{-16}$	-5415.386	121
		$\hat{\beta}_1$	0.035	0.103	0.7311		
		$\hat{\beta}_2$	0.019	0.002	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	-0.011	0.009	0.2267		
	σ	$\hat{\beta}_0$	-0.611	0.054	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.397	0.111	0.0003		
		$\hat{\beta}_2$	-0.002	0.003	0.4724		
		$\hat{\beta}_3$	0.015	0.010	0.1456		
	p_0	$\hat{\beta}_0$	-1.496	0.101	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.724	0.185	$8.87e^{-5}$		
		$\hat{\beta}_2$	-0.153	0.009	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	0.002	0.015	0.1243		
	p_1	$\hat{\beta}_0$	-1.480	0.095	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	-0.630	0.254	0.0132		
		$\hat{\beta}_2$	0.011	0.006	0.0666		
		$\hat{\beta}_3$	-0.069	0.022	0.0022		

Tabla 3-6: Parametros regresores estimados de un modelo de regresión ZOIP-beta con parametrización original en el porcentaje de utilización de una tdc.

Familia	Parámetro	β's	Estimación	Error estándar	Valor P	Log-Verosimilitud	Iteraciones
simplex	μ	$\hat{\beta}_0$	0.180	0.050	0.0003	-22385.78	117
		$\hat{\beta}_1$	-3.890	0.206	$< 2.2e^{-16}$		
		$\hat{\beta}_2$	0.166	0.004	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	-0.128	0.019	$5.85e^{-12}$		
	σ	$\hat{\beta}_0$	11.59	0.062	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.063	0.240	0.7918		
		$\hat{\beta}_2$	0.134	0.004	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	0.344	0.023	$< 2.2e^{-16}$		
	p_0	$\hat{\beta}_0$	-1.496	0.101	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	0.724	0.185	$8.87e^{-5}$		
		$\hat{\beta}_2$	-0.153	0.009	$< 2.2e^{-16}$		
		$\hat{\beta}_3$	0.002	0.015	0.1243		
	p_1	$\hat{\beta}_0$	-1.480	0.095	$< 2.2e^{-16}$		
		$\hat{\beta}_1$	-0.630	0.254	0.0132		
		$\hat{\beta}_2$	0.011	0.006	0.0666		
		$\hat{\beta}_3$	-0.069	0.022	0.0022		

Tabla 3-7: Parametros regresores estimados de un modelo de regresión ZOIP-simplex en el porcentaje de utilización de una tdc.

puede ver que al modelar el parámetro de la media, es decir, el porcentaje medio de utilización de la tdc, la variable regresora score afecta de manera negativa y significativa en cada uno de los modelos ajustados, excepto en el modelo con parametrización original, en el cual el parámetro no dio significativo, esto nos indica que a un mejor comportamiento de pago, menos utilización de la tdc, sobre el parámetro del promedio de cuotas diferidas vemos como el parámetro β_2 sin excepción alguna en todos los modelos es la variable más significativa que permite explicar la proporción media de la utilización de una tdc, esto nos indica que a medida que los clientes difieren a mayores cuotas sus compras con la tdc es-

tarán utilizando más su tdc, algo muy lógico, desde el punto de vista del problema; ahora el parámetro β_3 asociado al cupo de su tdc tiene un efecto negativo aunque no muy significativo sobre la variable respuesta, lo cual indicaría que a mayor cupo en su tdc un poco menos de utilización de la tdc estará acompañado.

Al analizar el efecto de la variabilidad de la utilización de las tdc, se observa como el parámetro de score tiene un efecto positivo sobre la precisión de la utilización de la tdc, en el modelo ZOIP-beta parametrización Rigby y Stasinopoulos (2005) este efecto parece ser negativo, pero σ al estar representando la dispersión y no la precisión, estaría dando un efecto positivo sobre la precisión, por lo tanto a mejor comportamiento de pago la utilización de la tdc serán más parecidas entre los individuos, cabe resaltar que β_1 no es significativo en el modelo simplex. Además el parámetro β_3 asociado al cupo de la tdc no influye sobre la variabilidad en ninguno de los modelos propuestos, excepto en modelo de regresión ZOIP-simplex que tiene un efecto positivo sobre la variabilidad de los porcentajes de utilización de las tdc.

Por otra parte el efecto de que alguien no utilice su tdc es exactamente igual en los cuatro modelos propuestos, esto por la metodología de estimación de máxima verosimilitud, y se puede observar como el comportamiento de pago y la cantidad de cuotas a las que se difiere las compras el cliente afectan de manera significativa la no utilización de las tdc. Además es de resaltar que si los clientes no difieren a grandes cuotas sus compras y su comportamiento de pago es muy bueno estos clientes tendrán mayor probabilidad de no utilizar para nada las tdc.

En el porcentaje de utilización global de la tdc, vemos como las tres covariables incluidas en el modelo afectan de manera significativa la utilización por completo de la tdc y vemos como la variable que es más significativa es el cupo de las tdc y que este tiene un efecto negativo sobre la probabilidad de utilizar por completo mi tdc, lo que me indica que a mayor cupo menor probabilidad de utilizar por completo de mi tdc (efecto que habíamos evidenciado en la explicación de la utilización media de las tdc), el mismo efecto pasa sobre la variable score que me indica que a peor comportamiento de pago más utilización la tdc, sin embargo, si el cliente difiere a grandes cuotas su compras este tendera a tener una mayor probabilidad de utilizar por completo su tdc. cabe resaltar que el efecto de las covariables sobre la probabilidad de utilizar por completo la tdc es totalmente contrario al efecto causado sobre la probabilidad de no utilizar para nada la tdc, algo lógico y esperado.

Al analizar el valor de la log-verosimilitud se observa que el mejor modelo de regresión que explica el porcentaje de utilización de una tdc en esta entidad financiera es la regresión ZOIP-beta, ya que posee un valor de log-verosimilitud menor que el modelo de regresión ZOIP-simplex, sin embargo, no importa la parametrización que se tenga en la regresión ZOIP-

beta, porque el valor de la log-verosimilitud son significativamente iguales, sin embargo, el modelo que requiere menor número de iteraciones para ser ajustado es la regresión ZOIP-beta parametrización Ferrari y Cribari-Neto (2004) seguido por la parametrización original y por último la parametrizacion de Rigby y Stasinopoulos (2005).

3.4. Conclusión

El modelo de regresión ZOIP, es un modelo de regresión de efectos fijos que es desarrollado bajo la distribución ZOIP y el cual se encarga de encontrar las covariables o factores que más influyen en una variable respuesta cuya distribución asociada es una distribución ZOIP. La estimación del efecto de las covariables sobre la variable respuesta se realiza mediante máxima verosimilitud, dicha maximización de la verosimilitud no tiene una solución cerrada analíticamente, por lo que se realiza computacionalmente y el paquete **ZOIP** de R da una solución a esto. Según los estudios de simulación realizados en este trabajo, las estimaciones convergen con un tamaño de muestra moderado a sus valores reales; aunque en ocasiones ocurre que las convergencias de los efectos de las covariables asociadas a los parámetros de inflación requieren un mayor número de muestras para demostrar su convergencia.

En el modelo de regresión ZOIP y el paquete **ZOIP** de R es posible realizar de una manera muy sencilla los principales tipos de regresiones para datos proporciones inflados con ceros y unos que existen, como la regresión simplex y la regresión beta bajo diferentes parametrizaciones, además el modelo de regresión ZOIP permite realizar ajustes a modelos inflados unilateralmente, es decir, donde solo haya datos proporcionales con valores cero o uno, e incluso realizar un ajuste sobre un modelo de regresión para datos proporcionales sin inflaciones.

4 Capítulo 4: Modelo de regresión ZOIP con efectos mixtos

Los modelos de regresión mixtos han sido de mucho interés en la última época, por su capacidad de estimar el efecto de una variable sobre el modelo, a través de la estimación de la varianza de la distribución de la variable, estos fueron introducidos de una manera general por Laird y Ware (1982). Es por esto que los modelos de regresión mixtos también han sido implementados cuando la variable respuesta es tomada por una variable aleatoria perteneciente a datos proporcionales, tal es el caso de Rigby y Stasinopoulos (2005) en los modelos aditivos generalizados para localización, escala y forma (Gamlss) que implementan el modelo de regresión beta con intercepto aleatorio normal, así otros autores como Verkuilen y Smithson (2012) y Bonat et al. (2012) proponen modelos de regresión beta con efectos aleatorios normales, estimados a partir de máxima verosimilitud marginal y metodologías bayesianas. Figueroa-Zuñiga et al. (2013) extienden el modelo propuesto por Ferrari y Cribari-Neto (2004) a un modelo con efectos fijos y aleatorios bajo la distribución normal y bajo la distribución t en estructuras de regresión tanto para el parámetro de la media, como del parámetro de precisión, la estimación de los parámetros del modelo de regresión fue realizado bajo una perspectiva bayesiana, mediante implementaciones computacionales del muestreador de Gibbs. Usuga y Giampaoli (2013) desarrollan el modelo de regresión beta mixto para datos proporcionales longitudinales, bajo intercepto y pendiente aleatoria normal y no normal, la estimación de los parámetros es realizada vía máxima verosimilitud y la cuadratura de Gauss-Hermite adaptativa. Otros autores como Song et al. (2004) implementan un modelo de regresión mixto para una variable respuesta bajo la distribución simplex, Bonat et al. (2015) también realiza un análisis de verosimilitud del modelo beta mixto, donde la estimación de los parámetros de regresión es realizada bajo algoritmos MCMC.

Como se pudo notar anteriormente, se ve que las metodologías de estimación de los parámetros del modelo de regresión mixto no tienen una solución cerrada analíticamente y es un poco más complicada cuando se trata de una variable respuesta perteneciente a datos proporcionales, por lo que se utilizan ciertas aproximaciones o en la mayoría de los casos metodologías y algoritmos bayesianos para la estimación de los parámetros regresores. Una de las metodologías utilizadas es la aproximación de la función de verosimilitud vía la cuadratura de Gauss-Hermite adaptativa, utilizada e implementada en los modelos de regresión beta mixtos por Usuga y Giampaoli (2013) y es utilizada para la estimación del modelo

de regresión ZOIP mixto, dicha cuadratura fue implementada anteriormente por Fahrmeir y Tutz (2001) sobre los modelos lineales generalizados. Diversos estudios para la estimación de parámetros sobre modelos estadísticos han sido implementados mediante esta técnica, por ejemplo, el trabajo realizado por Liu y Pierce (1994) y Smithson y Verkuilen (2006) que estima los parámetros del modelo de regresión beta bajo la cuadratura de Gauss-Hermite. Por otra parte, se han realizado diversas modificaciones sobre la cuadratura de Gauss-Hermite original, tales como la cuadratura de Gauss-Hermite adaptativa y algunas mejoras sobre esta como la cuadratura de Gauss-Hermite adaptativa con *pruning* (Hernández et al., 2013).

Los modelos de regresión mixtos que se han mencionado anteriormente sobre datos proporcionales, no se encuentran con presencia de datos en cero y/o uno, es decir inflados con ceros y/o unos, por lo que otros autores como Ospina y Ferrari (2010) presentaron una distribución beta inflada con ceros o con unos, mediante una combinación de una distribución discreta y una distribución continua dada por la distribución beta con parametrización de Ferrari y Cribari-Neto (2004) y la cual dio pie para que más adelante Ospina y Ferrari (2012) propusieran una clase general de modelos de regresión beta inflados en cero y uno. Recientemente Kosmidis et al. (2015) han estudiado los modelos de regresión inflados para datos proporcionales, pero basados en una distribución distinta a la propuesta por Ospina y Ferrari (2012), sin embargo, cabe aclarar que los anteriores modelos son modelos de regresión para efectos fijos, es decir, no incluyen algún efecto aleatorio, por lo que otros autores como Galvis y Lachos (2014) incluyen efectos aleatorios dentro de los modelos de regresión inflados con ceros y/o unos, basados en la distribución propuesta por Ospina y Ferrari (2010)y otras distribuciones para datos proporcionales, como la distribución simplex y beta-rectangular, la estimación de los parámetros de regresión se realizó mediante metodologías bayesianas, MCMC.

Este capítulo se encuentra organizado de la siguiente manera: primero se presenta el modelo de regresión ZOIP mixto basado en la distribución ZOIP visto en el capítulo 2 y su debida estimación, mediante máxima verosimilitud, se muestra los diferentes tipos de cuadratura de Gauss-Hermite, para que posteriormente se muestre la aproximación de la función de verosimilitud vía la cuadratura de Gauss-Hermite adaptativa multidimensional, en la siguiente sección se presenta la implementación del modelo de regresión ZOIP mixto en el paquete **ZOIP** de R y por último se presenta unas aplicaciones a datos simulados y a datos reales.

4.1. Modelo de regresión ZOIP mixto

Una escritura jerárquica de dos niveles considerada para un modelo con variable respuesta dada por la distribución para datos proporcionales inflados con ceros y/o unos (ZOIP), vista en la sección 2.2. Denotando a y_{ij} como la $j-$ésima medida del $i-$ésimo grupo, además si asumisos interceptos aleatorios γ_{i1} y γ_{i2}, los cuales son independientes y cada uno sigue

una distribución normal con media cero y desviación estándar λ_1 y λ_2, respectivamente. Asumimos también que los interceptos aleatorios γ_{i1} y γ_{i2} son independientes entre sí. Una escritura matemática para el modelo es el siguiente:

$$y_{ij}|\gamma_{i1},\gamma_{i2} \overset{\text{ind}}{\sim} ZOIP(\mu_{ij},\sigma_{ij},p_{0ij},p_{1ij}),$$

$$\gamma_{i1} \overset{\text{i.i.d}}{\sim} N(0,\lambda_1^2),$$
$$\gamma_{i2} \overset{\text{i.i.d}}{\sim} N(0,\lambda_2^2),$$

Para $i = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, n_i$ Los parámetros μ, σ, p_0 y p_1 son modelados linealmente en función de un conjunto de covariables respectivamente, por:

$$h_1(\mu_{ij}) = \mathbf{x}_{ij1}^\top \boldsymbol{\beta}_1 + \gamma_{i1},$$

$$h_2(\sigma_{ij}) = \mathbf{x}_{ij2}^\top \boldsymbol{\beta}_2 + \gamma_{i2},$$
$$h_3(p_{0ij}) = \mathbf{x}_{ij3}^\top \boldsymbol{\beta}_3,$$
$$h_4(p_{1ij}) = \mathbf{x}_{ij4}^\top \boldsymbol{\beta}_4$$

donde $\mathbf{x}_{ij1}$, $\mathbf{x}_{ij2}$, $\mathbf{x}_{ij3}$ y $\mathbf{x}_{ij4}$, son vectores de covariables conocidos de dimensión k_1, k_2, k_3 y k_4 respectivamente. $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$ son vectores de parámetros desconocidos fijos asociados a las covariables y γ_{i1}, γ_{i2} son los interceptos aleatorios asociados al $i-$ésimo grupo. Además las funciones $h_1(\cdot)$, $h_2(\cdot)$, $h_3(\cdot)$ y $h_4(\cdot)$ son funciones de enlace conocidas y apropiadas para mapear de los reales a los valores admisibles del parámetro, ademas son funciones estrictamente monótonas y doblemente diferenciables. Las posibles funciones para el parámetro μ y σ son logit, probit, clog-log, o log dependiendo de la parametrización, para los parámetros de inflación p_0 y p_1 son posibles funciones de enlace como logit, probit, clog-log.

4.1.1. Inferencia estadística

Para la estimación de los parámetros del modelo de regresión con intercepto aleatorio en la media y la varianza, para datos proporcionales inflados con ceros y/o unos, por medio de máxima verosimilitud, es necesario hallar la función de verosimilitud.
Considere $\boldsymbol{\theta} = (\boldsymbol{\beta_1}^\top, \boldsymbol{\beta_2}^\top, \boldsymbol{\beta_3}^\top, \boldsymbol{\beta_4}^\top, \lambda_1, \lambda_2)^\top$ un vector de parámetros a ser estimado en el espacio:

$$\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^k | \boldsymbol{\beta_1} \in \mathbb{R}^{k_1}, \boldsymbol{\beta_2} \in \mathbb{R}^{k_2}, \boldsymbol{\beta_3} \in \mathbb{R}^{k_3}, \boldsymbol{\beta_4} \in \mathbb{R}^{k_4}, \lambda_1 \in \mathbb{R}^+, \lambda_2 \in \mathbb{R}^+\right\}$$

en el que $k = k_1 + k_2 + k_3 + k_4 + 2$, tenemos que una distribución marginal de $\mathbf{y}_i =$

$(y_{1i}, \ldots, y_{n_i} i)^\top$ es dada por:

$$f(\mathbf{y}_i; \boldsymbol{\theta}) = \int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f(y_{ij}|\gamma_{i1}, \gamma_{i2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}) \cdot f(\gamma_{i1}; \lambda_1) f(\gamma_{i2}; \lambda_2) d\gamma_{i1} d\gamma_{i2},$$

Entonces una función de verosimilitud para las observaciones $\mathbf{y} = (y_1, \cdots, y_N)^\top$ es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(\mathbf{y}_i; \boldsymbol{\theta})$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f(y_{ij}|\gamma_{i1}, \gamma_{i2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}) \cdot f(\gamma_{i1}; \lambda_1) f(\gamma_{i2}; \lambda_2) d\gamma_{i1} d\gamma_{i2},$$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} log \left[\int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f(y_{ij}|\gamma_{i1}, \gamma_{i2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}) \cdot f(\gamma_{i1}; \lambda_1) f(\gamma_{i2}; \lambda_2) d\gamma_{i1} d\gamma_{i2} \right], \tag{4-1}$$

donde $f(y_{ij}|\gamma_{i1}, \gamma_{i2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4})$ es la función de densidad de probabilidad condicional de y_{ij} que distribuye $ZOIP(\mu, \sigma, p_0, p_1)$. γ_{i1}, γ_{i2} y $f(\gamma_{i1}; \lambda_1)$ y $f(\gamma_{i2}; \lambda_2)$ son funciones de densidades de probabilidad normales de γ_{i1} y γ_{i2}, respectivamente. Para encontrar el máximo de la función $\ell(\boldsymbol{\theta})$ no es posible mediante una manera cerrada analíticamente, una dificultad adicional es la solución de la integral para encontrar la maximización de la función $\ell(\boldsymbol{\theta})$, por lo que es necesario utilizar técnicas computacionales para la solución de esta, técnicas tales como aproximaciones de Laplace, algoritmos EM, integración Monte Carlo o técnicas bayesianas. Para solucionar dicha función de log-verosimilitud se utilizo el método de integración numérica Gauss-Hermite adaptativa multidimensional con y sin *pruning*, tal como se describe en la siguiente sección.

4.1.2. Cuadratura de Gauss-Hermite

Cuadratura de Gauss-Hermite unidimensional

La cuadratura de Gauss-Hermite (GQ) es una herramienta útil para aproximar una integral de una función $g(x)$ sobre $\Re$ con una suma ponderada, donde la variable x es reemplazada por una cuadratura de n puntos o nodos. Cada punto de la cuadratura es denotado por p_i, es evaluado en la función y los resultados son ponderados por los pesos de la cuadratura

asociados w_i.

$$\int_{\Re} g(x)dx \approx \sum_{i=1}^{n} g(p_i)exp(p_i^2)w_i.$$

El conjunto de los n puntos de la cuadratura $\mathbf{P} = \{p_1, p_2, \ldots, p_n\}$ corresponde a las raíces del polinomio de Hermite dado por:

$$H_n(x) = (-1)^n e^{-x^2} \frac{d^n}{dx^n} e^{-x^2},$$

con pesos asociados $\mathbf{W} = \{w_1, w_2, \ldots, w_n\}$ dados por

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2}.$$

Cuadratura de Gauss-Hermite adaptativa

Unidimensional

La cuadratura de Gauss-Hermite adaptativa (AGQ) es propuesta por Liu y Pierce (1994); (Pinheiro y Bates, 1995), es básicamente una transformación de los puntos asociados a la cuadratura, centrando y extendiendo alrededor del máximo valor de $\hat{x}$ de la función $log(g(x))$. La transformación de los puntos de la cuadratura p_i definido como p_i^*, está dado por $p_i^* = \sqrt{2}\hat{\sigma} p_i + \hat{x}$ donde:

$$\hat{\sigma}^2 = \left[-\frac{d^2}{dx^2} log(g(x)) \Big|_{x=\hat{x}} \right]^{-1}.$$

Así, la aproximación de la integral de $g(x)$ sobre $\Re$ está dado por:

$$\int_{\Re} g(x)dx \approx \sqrt{2}\hat{\sigma} \sum_{i=1}^{n} g(p_i^*)exp(p_i^2)w_i.$$

Multidimensional

Si extendemos la AGQ a una integral de dimensión q de la función $g(x)$ sobre $\Re^q$, en este caso, con una cuadratura de n puntos, $\mathbf{Z}$ está basado en el producto cartesiano de $\mathbf{P}$, y los pesos de la cuadratura de $\mathbf{A}$ están basados similarmente en el producto Kronecker, denotado por $\otimes$, los pesos originales $\mathbf{W}$, son dados:

$$\mathbf{Z} = \underbrace{P \times \ldots \times P}_{q \text{ veces}} = P^q,$$

$$\mathbf{A} = \underbrace{W \otimes \ldots \otimes W}_{q \text{ veces}}.$$

Así, la expresión para la integral aproximada de $g(x)$ sobre $\Re^q$ está dado por:

$$\int_{\Re^q} g(x)dx \approx |\hat{Q}|^{1/2} 2^{q/2} \sum_{i=1}^{n^q} g(z_i^*) exp(z_i^\top z_i) a_i,$$

donde z_i y a_i corresponden a los elementos de $\mathbf{Z}$ y $\mathbf{A}$, respectivamente. Los nuevos puntos de la cuadratura z_i^* estan centrados en el máximo de $\hat{x}$ del $log(g(x))$ y está dado por $z_i^* = \hat{x} + \sqrt{2}\hat{Q}^{1/2} z_i$, donde $\hat{Q}^{1/2}$ corresponde a la descomposición de Cholesky de la curvatura de la matriz $\hat{Q}$, que se encuentra dada por:

$$\hat{Q} = \left[-\frac{d^2}{dx^2} log(g(x)) \bigg|_{x=\hat{x}} \right]^{-1}.$$

Cuadratura de Gauss-Hermite adaptativa con pruning

Es claro que los resultados obtenidos por la AGQ son mejores que los de GQ, debido a que se encuentran centrados, sin embargo, la AGQ requiere un tiempo de optimización más elevado, debido a la transformación de los puntos de cuadratura, pero no todos los puntos de la AGQ aportan de manera significativa un valor sobre la solución de la integral, es por esto que Hernández et al. (2013) desarrolla un mejoramiento de la AGQ, donde elimina los puntos de la cuadratura que no son significativos sobre la solución de la integral, de modo que no afectan de manera significativa los resultados finales de la integral, pero si afectan de manera positiva el tiempo de ejecución, dicho mejoramiento es llamado cuadratura de Gauss-Hermite con *pruning*.

La cuadratura de Gauss-Hermite adaptativa con *pruning* consiste en eliminar puntos de la cuadratura, tales que el peso a_i asociado al punto es menor que un valor de referencia θ, que está dado por:

$$\theta = \frac{w_{[1]} w_{[\frac{n+1}{2}]}}{n^{q-1}}.$$

donde $w_{[1]}$ y $w_{[\frac{n+1}{2}]}$ corresponden respectivamente, a el valor mínimo y la mediana de los pesos originales $\mathbf{W}$. (Ver más en Hernández et al. (2013).)

4.1.3. Aproximación de la función de verosimilitud vía cuadratura de Gauss-Hermite

En la función de verosimilitud definida en (4-1) se tiene que para cada i-esimo grupo se debe resolver la siguiente integral:

$$
\begin{aligned}
I_i &= \int_{\Re^2} \prod_{j=1}^{n_i} f(y_{ij}|\gamma_{i1},\gamma_{i2};\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\boldsymbol{\beta_4}) \cdot f(\gamma_{i1};\lambda_1) f(\gamma_{i2};\lambda_2) d\gamma_{i1} d\gamma_{i2} \\
&= \int_{\Re^2} \prod_{j=1}^{n_i} f(y_{ij}|\gamma_{i1},\gamma_{i2};\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\boldsymbol{\beta_4}) \cdot \frac{exp(\gamma_{i1}^2/2\lambda_1^2)}{\lambda_1\sqrt{2\pi}} \cdot \frac{exp(\gamma_{i2}^2/2\lambda_2^2)}{\lambda_2\sqrt{2\pi}} d\gamma_{i1} d\gamma_{i2}
\end{aligned}
$$

Si se realiza el siguiente cambio de variables

$$
\begin{aligned}
b_{i1} &= \frac{\gamma_{i1}}{\sqrt{2}\lambda_1} \qquad b_{i2} = \frac{\gamma_{i2}}{\sqrt{2}\lambda_2} \\
\therefore b_{i1}^2 &= \frac{\gamma_{i1}^2}{2\lambda_1^2} \qquad b_{i2}^2 = \frac{\gamma_{i2}^2}{2\lambda_2^2} \\
\therefore \gamma_{i1} &= \sqrt{2}\lambda_1 b_{i1} \quad \gamma_{i2} = \sqrt{2}\lambda_2 b_{i2}
\end{aligned}
$$

Por lo anterior se tiene que la integral I_i se convierte en:

$$
I_i = \int_{\Re^2} \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{2}\lambda_1 b_{i1}, \sqrt{2}\lambda_2 b_{i2};\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\boldsymbol{\beta_4}) \cdot \frac{exp(-b_{i1}^2)exp(-b_{i2}^2)}{\pi} db_{i1} db_{i2} \tag{4-2}
$$

La integral definida en (4-2) tiene una forma factible para ser aproximada usando la cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*, vista en la sección 4.1.2, de este modo la integral I_i es aproximada por:

$$
I_i = \sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2};\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\boldsymbol{\beta_4}) \cdot \frac{w_{k_1} w_{k_2}}{\pi}
$$

donde z_{k_1} y z_{k_2} son los puntos de la cuadratura, w_{k_1} y w_{k_2} son los pesos asociados a los puntos de la cuadratura, por lo tanto la función de verosimilitud aproximada es dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}) \cdot \frac{w_{k_1} w_{k_2}}{\pi} \right]$$

y la función de log verosimilitud esta dada por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} log \left[\sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}) \cdot \frac{w_{k_1} w_{k_2}}{\pi} \right] \tag{4-3}$$

Al tener la función de log verosimilitud definida en (4-3), para hallar los estimadores máximos verosímiles se debe utilizar herramientas computacionales, como algoritmos de optimización, tales como las funciones de R, `nlminb` o `optim`. La metodología de estimación de los parámetros del modelo de regresión ZOIP con intercepto aleatorio en la media y la varianza, utilizando máxima verosimilitud vía cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*, se encuentra implementada en el paquete **ZOIP** de R, por medio de la función `RMM.ZOIP`.

4.2. Modelo de regresión ZOIP mixto en el paquete ZOIP

En esta sección se mostrara como ajustar un modelo de regresión ZOIP con interceptos aleatorios en los parámetros de media y varianza, mediante el paquete **ZOIP** de R, utilizando el método de máxima verisimilitud y su aproximación mediante la cuadratura de Gauss-Hermite adaptativa multidimensional con *pruning* o sin *pruning*.

4.2.1. Función RMM.ZOIP

La función `RMM.ZOIP` estima los parámetros de un modelo regresión ZOIP con y sin covariables y con interceptos aleatorios en los parámetros de la media y la dispersión, dicha estimación se realiza vía máxima verosimilitud utilizando la cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*, además se utilizara el optimizador `nlminb` o `optim` para la estimación de los efectos fijos. La estructura de la función `RMM.ZOIP` es la siguiente:

```
RMM.ZOIP(formula.mu, formula.sigma = ~1, formula.p0 = ~1,
    formula.p1 = ~1, data, formula.random, link = c("identity",
        "identity", "identity", "identity"), family = "R-S",
    optimizer = "nlminb", n.points = 11, pruning = TRUE)
```

Los argumentos de la función `RMM.ZOIP` son:

- `formula.mu`: Formula que define la función de regresión de efectos fijos, para el parámetro μ, un valor posible es `y ~ x1 + x2`, es necesario definir la variable respuesta (y).
- `formula.sigma`: Formula que define la función de regresión de efectos fijos, para el parámetro σ, un valor posible es `~ x1`, por defecto `~ 1`.
- `formula.p0`: Formula que define la función de regresión de efectos fijos, para el parámetro p_0, un valor posible es `~x1`, por defecto `~1`.
- `formula.p1`: Formula que define la función de regresión de efectos fijos, para el parámetro p_1, un valor posible es `~x1`, por defecto `~1`.
- `data`: Es el conjunto de datos en formato `data.frame` donde debe contener las nombres de las columnas tal cual como están en las fórmulas.
- `formula.random`: Formula que define el efecto mixto dentro del modelo, debe ser solo el intercepto aleatorio que se tendrá en cuenta en el parámetro de la media y la dispersión, la estructura admisible es la siguiente `formula.random = 1 | G1`, donde `G1` es la variable que indica los grupos o sujetos en el modelo, siempre debe ser definido.
- `family`: Elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP, si toma el valor de `‘‘R-S’’` se utilizara la distribución beta con parametrización Rigby y Stasinopoulos (2005), si toma el valor de `‘‘F-C’’` se utilizara la distribución beta parametrización Ferrari y Cribari-Neto (2004), el valor de `‘‘Original’’` se utilizara la distribución beta con parametrización original, `‘‘Simplex’’` utilizará la distribución simplex.
- `link`: Es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y formula. Si el modelo de regresión no posee covariables se debe utilizar como función enlace la opción `identity`, independientemente del valor escogido en familia, opciones posibles son `logit`, `log`. Por defecto `link=c(‘‘identity’’,‘‘identity’’,‘‘identity’’,‘‘identity’’)`.
- `optimizer`: Elección del optimizador, utilizado para encontrar la convergencia de la máxima verosimilitud en los parámetros de efectos fijos, se puede elegir el valor de `‘‘nlminb’’` o `‘‘optim’’`, por defecto `‘‘nlminb’’`
- `n.points`: Número de puntos a utilizar en la aproximación de la función de verosimilitud por medio de la cuadratura de Gauss-Hermite adaptativa multidimensional, por defecto es 11, se recomienda no dar un valor muy grande a este parámetro, por que afectara de manera significativamente los tiempos de convergencia del modelo.
- `pruning`: Es un valor booleano que indica si se utiliza *pruning* o no, para la cuadratura de Gauss-Hermite adaptativa multidimensional, por defecto es TRUE.

En el siguiente ejemplo se mostrará el ajuste de un modelo de regresión ZOIP mixto, el cual proviene de datos simulados, se muestra el código utilizado para dicha simulación, como se utiliza la función `RMM.ZOIP` para ajustar el modelo y mostrar la salida del modelo, explicando cada uno de sus componentes de la salida y la utilización de la función `summary` sobre el objeto de salida de la función `RMM.ZOIP`. Los datos simulados son provenientes de una distribución ZOIP con parametrización Rigby y Stasinopoulos (2005) con parámetros, como μ y σ que

dependerán de un intercepto aleatorio y una covariable discreta definida como cantidad de días (2, 10, 20, 40), los parámetros de inflación son fijados como $p_0 = p_1 = 0.1$, el número de grupos o sujetos será fijado a partir de N, como 21 grupos.

Se carga el paquete **ZOIP** y se definen los diferentes valores de los parámetros de la distribución ZOIP, para ser simulada.

```
library(ZOIP)
N <- 21  # Numeros de grupos o sujetos

Times <- c(2, 10, 20, 40)  # cantidad de dias

subject <- rep(1:N, each = length(Times))
# numero de sujetos en la muestra repetidos tantas veces
# haya dias

Days <- rep(Times, times = N)
b0i <- rep(rnorm(n = N, sd = 1), each = length(Times))
b1i <- rep(rnorm(n = N, sd = 0.5), each = length(Times))

neta <- (1.6 + b0i) - 1.3 * log(Days)
neta2 <- (0.1 + b1i) - 0.8 * log(Days)

mu <- 1/(1 + exp(-neta))
sigma <- 1/(1 + exp(-neta2))

p0 <- 0.1
p1 <- 0.1
```

Se verifica que no hayan valores de μ y σ iguales a unos y se simula los valores de la variable respuesta.

```
mu[mu == 1] <- 0.999
mu[mu == 0] <- 0.001

sigma[sigma == 1] <- 0.999
sigma[sigma == 0] <- 0.001

family <- "R-S"

Y <- rZOIP(n = length(mu), mu = mu, sigma = sigma, p0 = p0,
    p1 = p1, family = family)
base <- data.frame(Y, Days, subject)
```

Se definen los argumentos de la función `RMM.ZOIP`, tales como las regresiones a ser ajustadas a cada uno de los parámetros de la distribución ZOIP.

```
formula.mu <- Y ~ log(Days)
formula.sigma <- ~log(Days)
formula.p0 <- ~1
formula.p1 <- ~1

formula.random <- ~1 | subject

link <- c("logit", "logit", "identity", "identity")

optimizer <- "nlminb"
n.points <- 11
pruning <- TRUE

mod <- RMM.ZOIP(formula.mu = formula.mu, formula.sigma = formula.sigma,
    formula.p0 = formula.p0, formula.p1 = formula.p1, data = base,
    formula.random = formula.random, link = link, family = family,
    optimizer = optimizer, n.points = n.points, pruning = pruning)
mod
```

Los resultados obtenidos se muestran a continuación.

```
## Call:
## RMM.ZOIP(formula.mu = formula.mu, formula.sigma = formula.sigma,
##     formula.p0 = formula.p0, formula.p1 = formula.p1, data = base,
##     formula.random = formula.random, link = link, family = family,
##     optimizer = optimizer, n.points = n.points, pruning = pruning)
##
##  Results:
##
##  Estimated fixed coefficients for h(mu):
## (Intercept)   log(Days)
##    1.845423   -1.331609
##
##  Estimated fixed coefficients for h(sigma):
## (Intercept)   log(Days)
##   0.2795793  -0.5611416
##
##  Estimated fixed coefficients for h(p0):
## (Intercept)
##  0.08319893
##
##  Estimated fixed coefficients for h(p1):
## (Intercept)
##   0.1191148
##  Estimated random coefficients for h(mu) and h(sigma)
##                                      log(.)
## Random Intercept mu    0.7610678 -0.2730329
## Random Intercept sigma 0.5038028 -0.6855703
##
##  message
## [1] "relative convergence (4)"
##
##  time
## [1] 227.48
##
##  iterations
## [1] 39
##
##  Log-likelihood
## [1] 15.75655
```

En el anterior resultado se obtienen varios aspectos importantes de la salida del modelo y leyendo de arriba hacia abajo, se tienen primero que todo los efectos fijos ajustados al modelo de regresión para el parámetro de la media μ, luego los efectos fijos estimados para el modelo de regresión del parámetro de dispersión σ, luego los parámetros ajustados para la regresión sobre el parámetro de inflación de ceros p_0, después los efectos estimados de la regresión del parámetro de inflación de unos p_1, seguido de la estimación de los interceptos aleatorios para μ y para σ, en ella se muestra una matriz donde la primera columna son los valores de la desviación estándar asociada a la distribución normal, es decir, el valor asociado a λ_1 y λ_2, dichos valores generan el efecto aleatorio sobre los parámetros de la media y la dispersión en su intercepto, respectivamente, la segunda columna nos muestra el valor del logaritmo natural de λ_1 y λ_2, respectivamente; luego el siguiente resultado es el valor de la log verosimilitud del modelo ajustado, para ser utilizados como posibles comparaciones entre modelos, después se muestra un mensaje de convergencia heredado del algoritmo de optimización `nlminb` o `optim`, luego el tiempo que demoro el ajuste del modelo en segundos y por último el número de iteraciones necesarias para la convergencia de encontrar la máxima verosimilitud.

```
summary(mod)
## ------------------------------------------------------------
## Fixed effects for logit(mu)
## ------------------------------------------------------------
##             Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)  1.84542    0.32408   5.6944 1.238e-08 ***
## log(Days)   -1.33161    0.10306 -12.9202 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for logit(sigma)
## ------------------------------------------------------------
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.27958    0.40333  0.6932    0.4882
## log(Days)   -0.56114    0.12949 -4.3335 1.468e-05 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
## ------------------------------------------------------------
## Fixed effects for identity(p0)
## ------------------------------------------------------------
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.083199   0.030112   2.763 0.005727 **
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## Fixed effects for identity(p1)
## ------------------------------------------------------------
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 0.119115   0.035352  3.3694 0.0007532 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## ------------------------------------------------------------
## Random effects for mu and sigma
## ------------------------------------------------------------
##                        Estimate Std. Error z value  Pr(>|z|)
## Random Intercept mu     0.76107    0.23334  3.2616 0.0011077 **
## Random Intercept sigma  0.50380    0.14046  3.5867 0.0003349 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
## ------------------------------------------------------------
## ------------------------------------------------------------
```

En el anterior resultado del paquete **ZOIP** se aplicó la función de método S3 `summary` al objeto que guardo la salida del modelo de regresión ZOIP mixto, esta función dará resultados más detallados de la estimación de los parámetros, tendrá como resultado para cada parámetro de regresión del modelo ZOIP mixto, el valor estimado, el error estándar asociado, el valor Z de la distribución normal y su respectivo valor p; esto ayudará al usuario del paquete **ZOIP** a concluir con más argumentos el ajuste de sus parámetros y covariables dentro del modelo de regresión ZOIP mixto ajustado.

4.3. Aplicación

En esta sección se muestran diferentes resultados sobre el ajuste de un modelo de regresión ZOIP con intercepto aleatorio en el parámetro de la media y la varianza, por medio del paquete **ZOIP**, primero se realizó el ajuste de un modelo de regresión ZOIP mixto sobre

datos reales, aplicado al porcentaje de utilización de la tarjeta de crédito (tdc), valorado por el efecto de la ciudad donde vive la persona que le pertenece la tdc, segundo se realizó un estudio de simulación del modelo de regresión ZOIP mixto, basado en la aplicación a datos reales, este estudio nos permite analizar la convergencia de la estimación de los parámetros asociados a los efectos fijos y aleatorios del modelo de regresión mixto, además poder analizar y determinar cuál es la mejor alternativa para estimar los modelos de regresión ZOIP mixto, sobre las diferentes opciones de utilizar la cuadratura de Gauss-Hermite adaptativa multidimensional.

4.3.1. Datos reales

En una entidad financiera es de importancia estudiar los efectos de ciertas variables sobre el porcentaje de utilización de las tarjetas de crédito de la entidad, porque esto dará respuestas a activación de campañas publicitarias y estudios de mercadeo para incentivar la utilización de las tdc, de acuerdo a la situación o características de la persona dueña de la tdc, es por esto que en esta sección se analizó el efecto que tiene la variable *Total mora* y la *ciudad*, sobre el porcentaje de utilización de la tdc. La variable *Total mora* nos indica el número de meses que la tarjeta ha entrado en mora en toda la vida de la tdc; la *ciudad* nos indicara donde vive el cliente a quien pertenece de la tdc. Estas dos variables son importantes para la entidad financiera, porque indicará o estimará el porcentaje de utilización de una tdc para cuando un cliente ha llegado a estar en mora varios meses y vive en cierta ciudad, por lo que indicará a la entidad financiera, como actuar sobre ciertas personas de interés. Para poder cuantificar dicho efecto se planteó un modelo de regresión ZOIP-beta con intercepto aleatorio en el parámetro de la media y la varianza, dado por la variable *ciudad* y un efecto fijo en la media y la varianza dado por la variable *Total mora*, el modelo es planteado bajo la distribución ZOIP-beta con parametrización de Rigby y Stasinopoulos (2005). El número de datos con que se cuenta es un total de 15 tarjetas de crédito por cada ciudad, el número de ciudades tomadas en cuenta son 10 (Bogotá, Medellín, Cali, Barranquilla, Bucaramanga, Cartagena, Cúcuta, Ibagué, Envigado y Neiva), para así obtener una base total de 150 tdc. el modelo planteado se muestra a continuación:

Si $y_{ij}|\gamma_{i1}, \gamma_{i2} \overset{\text{ind}}{\sim} ZOIP(\mu_{ij}, \sigma_{ij}, p_0, p_1)$,

$$\begin{aligned} h_1(\mu_{ij}) &= \beta_0 + \gamma_{1i} + \beta_1 x_{1ij}, \\ h_2(\sigma_{ij}) &= \beta_0 + \gamma_{2i} + \beta_1 x_{1ij}, \\ h_3(p_{0ij}) &= \beta_0, \\ h_4(p_{1ij}) &= \beta_0, \end{aligned} \tag{4-4}$$

donde y_{ij} es porcentaje de utilización de la j-esima tdc perteneciente a la i-esima ciudad,

$i = 1, 2, \ldots, 10$, es decir, $N = 10$ asociada a las 10 ciudades; $j = 1, 2, \ldots, 15$, es decir, $n_i = 15$ asociada al número de tdc por cada ciudad, en este caso todas las ciudades tienen el mismo número de observaciones, es decir, se cuenta con una muestra balanceada; x_{1ij}: es el valor del número de meses en mora de la j-esima tdc asociada a la i-esima ciudad; γ_{1i} y γ_{2i} son los interceptos aleatorios para los parámetros de la media y la varianza, respectivamente, asociados a la i-esima ciudad y provenientes de la distribución normal con media cero y desviación estándar λ_1 para el parámetro regresor γ_{1i} y λ_2 para el parámetro regresor γ_{2i}. Las funciones enlaces asociadas a cada regresión se tomaran como una función *logit*, esto debido a que el modelo de regresión ZOIP-beta mixto se planteó bajo la parametrización de Rigby y Stasinopoulos (2005) y este deberá utilizar dicha función enlace, tal cual como se explicó en el capítulo anterior.

	Parámetros regresores			
Parámetro	$\hat{\beta}_0$ (Error estándar)	$\hat{\beta}_1$ (Error estándar)	$\hat{\lambda}_1$ (Error estándar)	$\hat{\lambda}_2$ (Error estándar)
μ	-1.13 (0.24)	0.33 (0.13)	0.51 (0.30)	
σ	0.33 (0.20)	0.14 (0.09)		0.40 (0.31)
p_0	0.23 (0.03)			
p_1	0.07 (0.02)			

Tabla 4-1: Estimación de los efectos fijos y aleatorios del modelo de regresión ZOIP mixto para el porcentaje utilización de una tdc.

En la tabla **4-1** se muestra la estimación de los efectos fijos y aleatorios del modelo de regresión ZOIP-beta mixto, estimado vía máxima verosimilitud y mediante la aproximación de la cuadratura de Gauss-Hermite adaptativa multidimensional, utilizando 11 puntos de cuadratura con *pruning*. En esta tabla se evidencia como al tener un numero de meses en mora más alto se aumenta el porcentaje medio y la varianza de utilización de las tdc, además se observa que $\hat{\lambda}_1 = 0.51$ dando así que $\gamma_{1i} \sim N(0, 0.51)$, lo que nos permite ver el efecto del cambio de ciudad sobre el porcentaje medio de utilización de las tdc, el valor $\hat{\lambda}_2$ es de 0.4, evidenciando que $\gamma_{2i} \sim N(0, 0.4)$, lo que nos indicará cuantificar el efecto del cambio de ciudad sobre la variabilidad del porcentaje de utilización de las tdc, por otra parte se evidencia que alrededor de un 20 % de las tdc no se utilizan para nada y que un 7 % tienen utilizado la totalidad de su cupo de la tarjeta de crédito. Por último el modelo fue ajustado mediante la función `RMM.ZOIP` del paquete **ZOIP**, tardando un total de 83.5 segundos, un tiempo relativamente prudente para el ajuste un modelo complejo, como lo es el modelo de regresión ZOIP-beta mixto.

El modelo propuesto en (4-4) se puede reescribir con los parámetros estimados como:

$$y_{ij}|\gamma_{i1}, \gamma_{i2} \overset{\text{ind}}{\sim} ZOIP(\mu_{ij}, \sigma_{ij}, p_0, p_1),$$

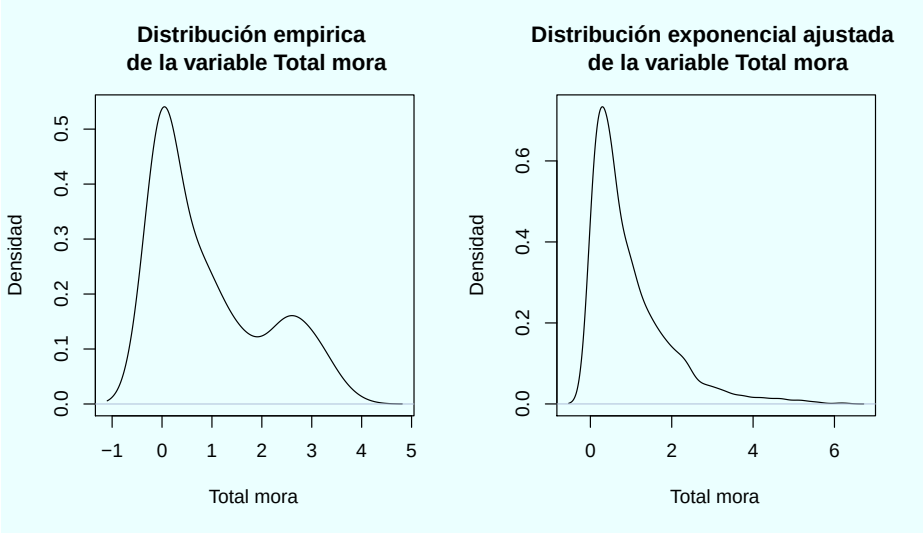

Figura 4-1: Ajuste de la distribución exponencial a la variable *Total mora*

$$
\begin{aligned}
h_1(\mu_{ij}) &= -1.13 + \gamma_{1i} + 0.33x_{1ij}, \\
h_2(\sigma_{ij}) &= 0.33 + \gamma_{2i} + 0.14x_{1ij}, \\
h_3(p_{0ij}) &= 0.23, \\
h_4(p_{1ij}) &= 0.07,
\end{aligned}
\tag{4-5}
$$

donde $\gamma_{1i} \sim N(0, 0.51)$ y $\gamma_{2i} \sim N(0, 0.4)$.

4.3.2. Datos simulados

En esta sección se muestra la realización de un estudio de simulación para verificar la convergencia de los parámetros regresores sobre los valores reales, además se puede concluir a partir de este estudio, cual es la mejor metodología posible para la estimación de los parámetros del modelo de regresión ZOIP mixto. Para dicho estudio se tomó en cuenta el modelo final estimado en la aplicación a datos reales de la sección anterior, así como se describió en el modelo estimado en (4-5). Para poder encontrar la mejor metodología de estimación del modelo de regresión ZOIP mixto se plantean 18 escenarios de simulación y se realizaron 1000 réplicas de cada escenario; los escenarios son resultantes de todas las combinaciones entre, variar el número de puntos de la cuadratura de Gauss-Hermite ($Q = 3, Q = 10, Q = 20$), variar el tamaño de muestra de cada grupo o ciudad ($n_i = 5, n_i = 20, n_i = 50$) y tener en cuenta si la cuadratura de Gauss-Hermite adaptativa multidimensional se realizará con o sin *pruning*.

Para poder variar el tamaño de muestra de cada ciudad es necesario encontrar la distribu-

ción que genera la covariable *Total mora*. En la figura **4-1** se muestra como la distribución exponencial con parámetro $\lambda = 1.075$ describe de forma adecuada el comportamiento de la variable *Total mora* descrita a partir de los datos reales obtenidos en la sección anterior.

Al tener la distribución que genera los datos de la variable *Total mora*, permitirá obtener cualquier número de observaciones de dicha variable, en particular el número de muestras necesarias para los escenarios de simulación propuestos, anteriormente. A continuación se muestra los resultados de la simulación del modelo de regresión ZOIP-beta mixto bajo la parametrización de Rigby y Stasinopoulos (2005), realizada bajo la función `RMM.ZOIP` del paquete **ZOIP** de R.

		Con *pruning*			Sin *pruning*		
Parámetro	β's	$n_i = 5$	$n_i = 20$	$n_i = 50$	$n_i = 5$	$n_i = 20$	$n_i = 50$
μ	$\hat{\beta}_0$	-1.137	-1.110	-1.076	-1.128	-1.120	-1.080
	$\hat{\beta}_1$	0.321	0.327	0.326	0.331	0.330	0.327
	$\hat{\lambda}_1$	0.879	0.576	0.507	0.882	0.568	0.498
σ	$\hat{\beta}_0$	0.445	0.380	0.336	0.452	0.377	0.345
	$\hat{\beta}_1$	0.072	0.118	0.132	0.066	0.121	0.133
	$\hat{\lambda}_2$	0.728	0.450	0.396	0.727	0.456	0.398
p_0	$\hat{\beta}_0$	0.220	0.230	0.230	0.220	0.230	0.230
p_1	$\hat{\beta}_0$	0.060	0.070	0.070	0.060	0.070	0.072
Tiempo(Seg)		115.72	130.85	140.58	61.69	163.36	218.68
Nro Iteraciones		22	30	34	22	30	34

Tabla 4-2: Estimación de los parámetros de regresión y los interceptos aleatorios de un modelo de regresión simulado ZOIP mixto, visto a partir del tamaño de muestra y si se utiliza *pruning* o no.

En la tabla **4-2** se muestra los valores estimados de cada uno de los parámetros de las regresiones ajustadas con respecto al tamaño de muestra n_i y si se utilizó *pruning* o no, los valores estimados de cada parámetro regresor y desviación estándar de las distribuciones normales asociadas a los interceptos aleatorios, fueron hallados, a partir de la mediana de los valores estimados de las réplicas de los escenarios de simulación que cumplen con cada combinación entre los diferentes tamaños de muestra y si fue con *pruning* o no. En dicha tabla se nota como los valores de λ_1 y λ_2 asociados a los interceptos aleatorios van convergiendo a su valor real a través que el tamaño de muestra va aumentando sin importar si se hizo con o sin *pruning*, además se nota como la estimación de la pendiente va convergiendo a través que le tamaño de muestra va creciendo, este parámetro da un valor muy plausible desde tamaños de muestra pequeños, sin importar si se hizo con o sin *pruning*, este último fenómeno se evidencia también con los parámetros de inflación en los que se tiene un valor muy cercano al real desde los tamaños de muestra pequeños. En cuestión del tiempo de ejecución se ve como es reducido en aproximadamente un 50 %, cuando se utiliza la metodología de la cuadratura de Gauss-Hermite adaptativa con *pruning*, con respecto a no utilizar *pruning*. Sobre el número

de iteraciones se observa como al aumentar el tamaño de la muestra se requiere una cantidad de iteraciones más elevada.

		Con *pruning*			Sin *pruning*		
Parámetro	β's	$Q=3$	$Q=10$	$Q=20$	$Q=3$	$Q=10$	$Q=20$
μ	$\hat{\beta}_0$	-1.122	-1.069	-1.120	-1.117	-1.076	-1.129
	$\hat{\beta}_1$	0.323	0.322	0.333	0.334	0.322	0.329
	$\hat{\lambda}_1$	0.632	0.626	0.634	0.629	0.616	0.623
σ	$\hat{\beta}_0$	0.400	0.365	0.366	0.382	0.379	0.373
	$\hat{\beta}_1$	0.117	0.119	0.120	0.123	0.117	0.121
	$\hat{\lambda}_2$	0.490	0.487	0.491	0.501	0.482	0.486
p_0	$\hat{\beta}_0$	0.228	0.228	0.226	0.226	0.226	0.226
p_1	$\hat{\beta}_0$	0.068	0.070	0.070	0.070	0.068	0.070
Tiempo(Seg)		75.295	128.28	271.825	74.295	162.35	367.545
Nro Iteraciones		30	29	29	29	29	29

Tabla 4-3: Estimación de los parámetros de regresión y los interceptos aleatorios de un modelo de regresión simulado ZOIP mixto, visto a partir del número de puntos de la cuadratura y si se utiliza *pruning* o no.

En la tabla **4-3** se muestran los valores estimados de cada uno de los parámetros de las regresiones ajustadas con respecto al número de puntos utilizados en la cuadratura de Gauss-Hermite adaptativa Q y si se utilizó *pruning* o no, los valores estimados de cada parámetro regresor y desviación estándar de las distribuciones normales asociadas a los interceptos aleatorios fueron hallados, a partir de la mediana de los valores estimados de las réplicas de los escenarios de simulación que cumplen con cada combinación entre los diferentes números de puntos de cuadratura y si fue con *pruning* o no. En dicha tabla se observa como en general cuando se utiliza una mayor cantidad de puntos de cuadratura se evidencia una pequeña mejora en la convergencia de los valores reales de los efectos fijos y aleatorios. Con respecto a la utilización de *pruning* o no, se observa como las estimaciones de parámetros no tienen una diferencia significativa en la utilización de *pruning* o no, por lo que es mejor utilizar la cuadratura de Gauss-Hermite adaptativa con *pruning* por que al parecer se obtienen resultados similares, pero con una reducción significativa del tiempo de estimación de los parámetros. El número de iteraciones utilizadas para la estimación de los parámetros no se ve influenciado por la cantidad de puntos utilizados en la cuadratura o si se utiliza *pruning* o no.

La mediana del error relativo de un parámetro estimado es calculado a partir de hallar el error relativo de las 1000 réplicas de cada escenario de simulación para cada parámetro estimado y calcularle la mediana a los mil valores calculados, el error relativo se puede definir como $|(\theta-\hat{\theta})/\theta|$, donde θ representa cualquier parámetro estimado en el modelo de regresión ZOIP mixto.

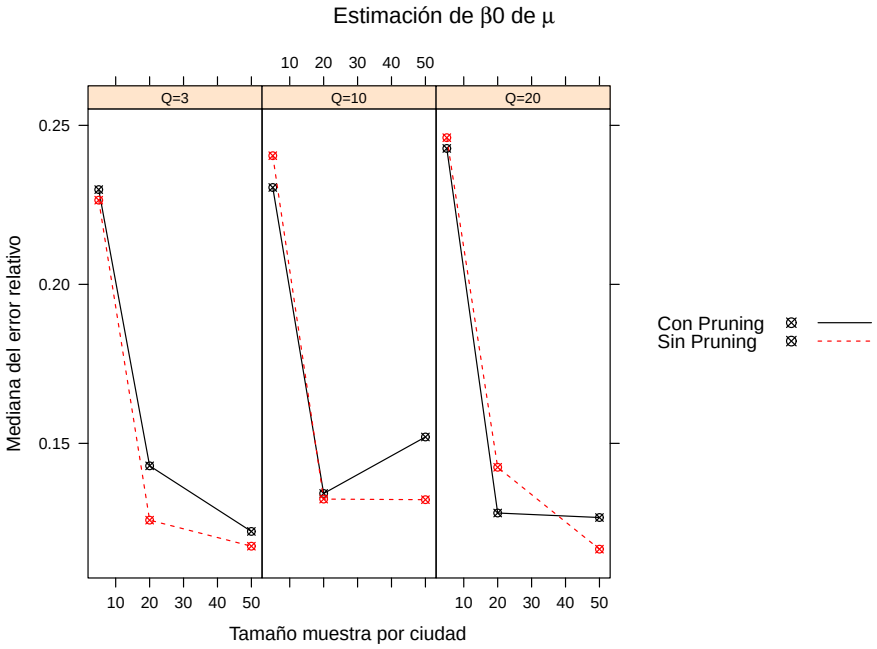

Figura 4-2: Mediana del error relativo de la estimación del parámetro $\beta 0$ asociado a la media, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

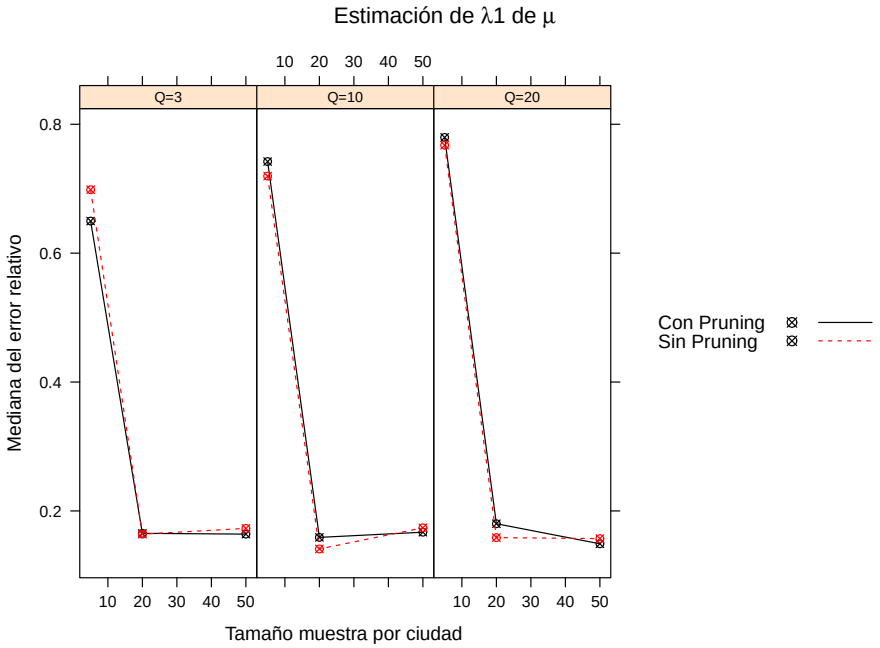

Figura 4-3: Mediana del error relativo de la estimación del parámetro λ_1 desviación estándar del intercepto aleatorio asociado a la media, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

En la figura **4-2** se muestra la mediana del error relativo de variar el tamaño de muesta n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el intercepto fijo $(\hat{\beta}_0)$ asociado a μ, en dicha figura se muestra como el aumentar el tamaño de muestra en cada ciudad se obtiene una reducción de la mediana del error relativo de $\hat{\beta}_0$, sin embargo, al aumentar el número de puntos de la cuadratura se obtienen errores muy parecidos en todos los tamaños de muestra, por lo que se podría decir que el aumento del número de puntos de cuadratura no mejora la estimación del intercepto fijo sobre la media, por otra parte se obtienen errores relativos parecidos cuando se utiliza *pruning* que cuando no, sin embargo, se puede rescatar que hay puntos como cuando el tamaño de muestra es 50 y $Q = 10$ el error relativo de con *pruning* es más grande del que no utiliza *pruning*.

En la figura **4-3** se muestra la mediana del error relativo de variar el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el valor de la desviación estándar de la distribución normal que genera el intercepto aleatorio $(\hat{\gamma}_{1i})$ del parametro de μ, en dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción significativa sobre la mediana del error relativo, además al aumentar el número de puntos de la cuadratura se obtienen errores menores cuando se aumentan el número de puntos de la cuadratura, dicha mejora es de alrededor de un 2 % cuando el tamaño de muestra es más grande y $Q = 20$, esto para cuando se utiliza *pruning* y cuando

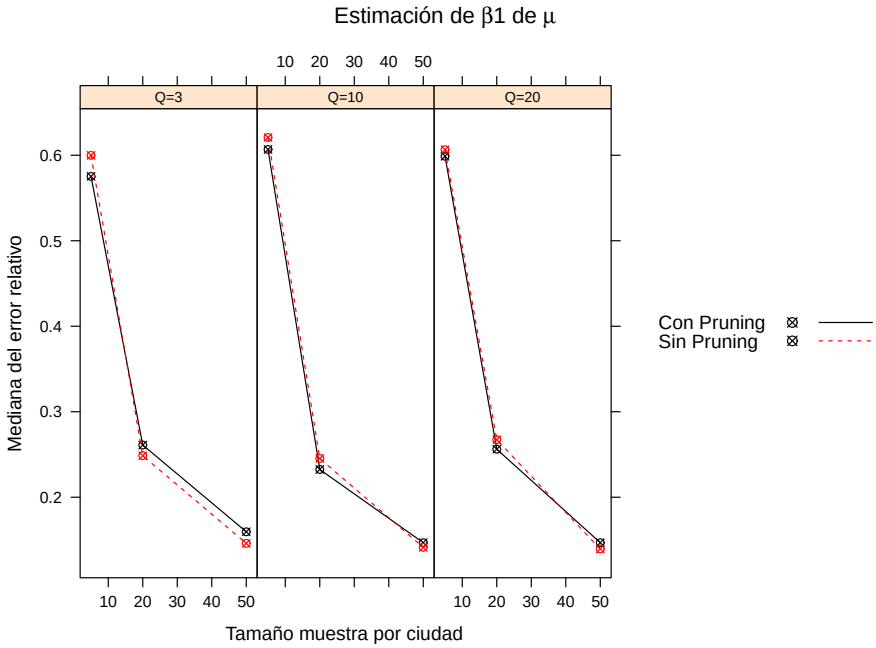

Figura 4-4: Mediana del error relativo de la estimación del parámetro $\beta 1$ asociado a la media, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

no, por lo que se podría decir que el aumento del número de puntos de cuadratura mejora la estimación del intercepto aleatorio sobre la media cuando el tamaño de muestra es relativamente grande, por otra parte se obtienen errores relativamente parecidos cuando se utiliza *pruning* que cuando no.

En la figura **4-4** se muestra la mediana del error relativo de variar el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el valor del efecto fijo de la variable *Total mora* sobre la media, en dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la mediana del error relativo, además al aumentar el número de puntos de la cuadratura se obtienen errores un poco menores, pero no significativo, cuando se aumentan el número de puntos de la cuadratura, lo que nos permite observar que el aumento del número de puntos de la cuadratura, no afecta demasiado en la estimación de este efecto fijo, esto sin importar si se utiliza *pruning* o no, por otra parte se obtienen errores muy parecidos cuando se utiliza *pruning* y cuando no, esto no parece afectar la estimación del parámetro.

En la figura **4-5** se muestra la mediana del error relativo de variar el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el intercepto fijo $(\hat{\beta}_0)$ del parámetro de σ, en dicha figura se muestra como cuando se aumenta el tamaño de

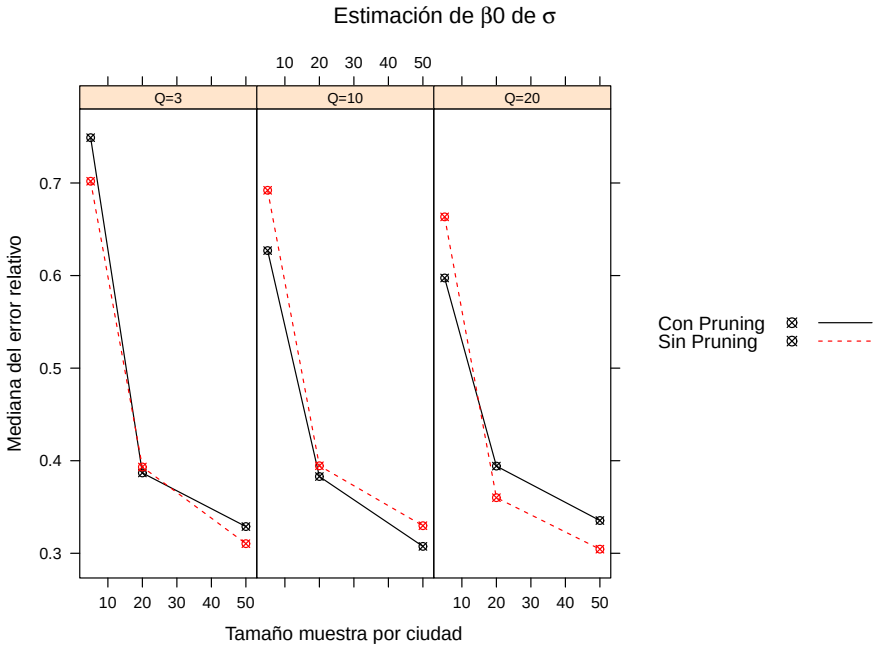

Figura 4-5: Mediana del error relativo de la estimación del parámetro $\beta 0$ asociado a la dispersión, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

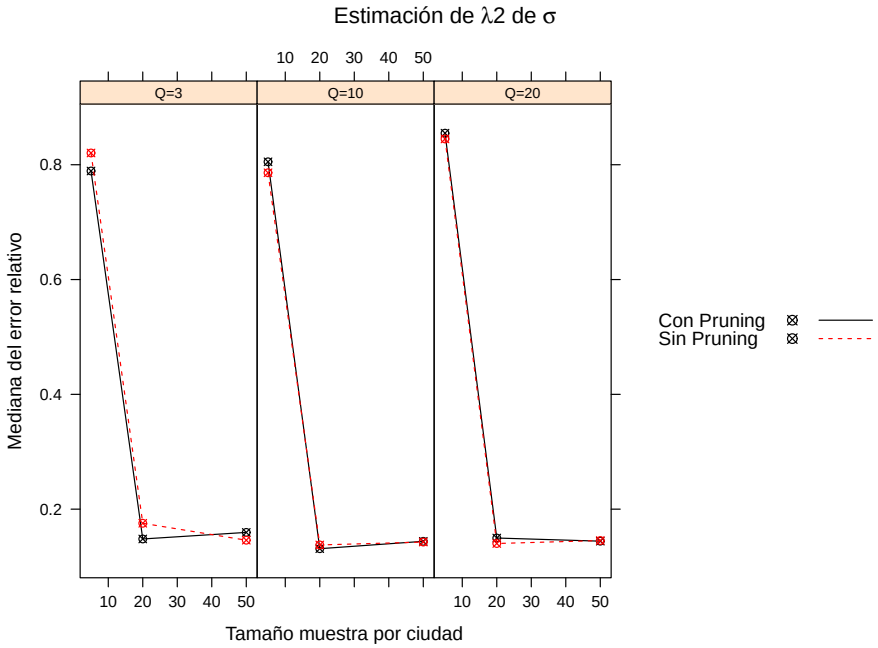

Figura 4-6: Mediana del error relativo de la estimación del parámetro λ_2 desviación estándar del intercepto aleatorio asociado a la dispersión, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

muestra en cada ciudad se obtiene una reducción de la mediana del error relativo, por otra parte al aumentar el número de puntos de la cuadratura se obtienen errores muy parecidos en todos los tamaños de muestra, sin embargo, cuando $Q = 20$ y se realiza sin *pruning* se nota una reducción en el error, Ahora si observamos las medianas de los errores relativos son relativamente parecidos cuando se utiliza *pruning* que cuando no, sin embargo, se puede rescatar que hay puntos como cuando el tamaño de muestra es 20 o 50 y el $Q = 20$ el error relativo, de la metodología sin utilizar *pruning* se ve reducido la mediana del error relativo, mejorando así la estimación del parámetro.

En la figura **4-6** se muestra la mediana del error relativo de variar, el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el valor de la desviación estándar de la distribución normal que genera el intercepto aleatorio ($\hat{\gamma}_2$) del parámetro de σ, en dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la mediana del error relativo de forma significativa, además al aumentar el número de puntos de la cuadratura se obtienen errores un poco menores, la mejora se nota a partir de $Q \geq 10$, dicha mejora es de alrededor un 2 % cuando los tamaños de muestra de cada grupo es mayor que 10, no se nota una diferencia cuando se utiliza *pruning* o cuando no, por lo que se podría decir que el aumento del número de puntos de cuadratura mejora la estimación del intercepto aleatorio sobre la dispersión cuando el tamaño de

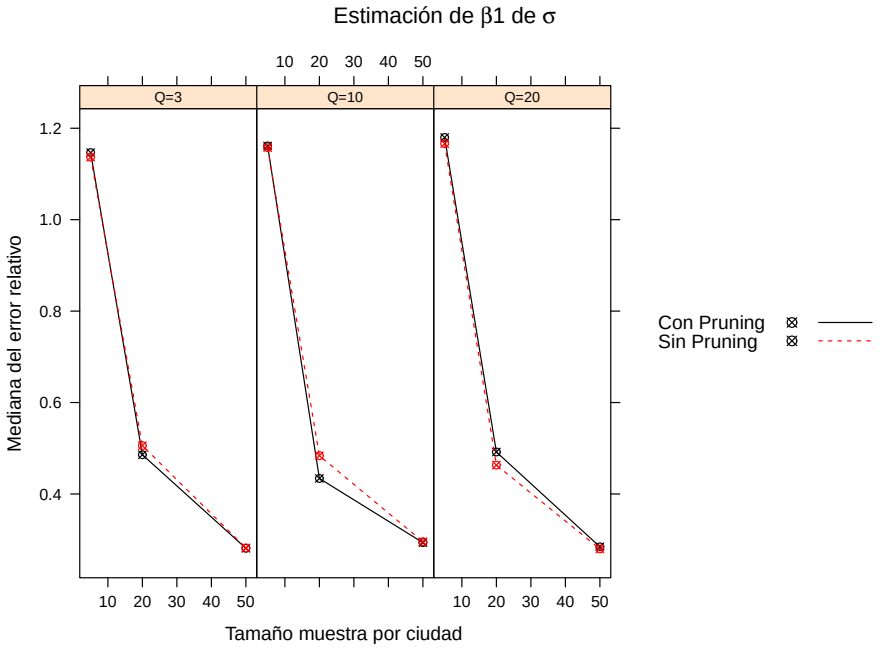

Figura 4-7: Mediana del error relativo de la estimación del parámetro $\beta 1$ asociado a la dispersión, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

muestra es relativamente grande, por otra parte se obtienen errores relativamente parecidos cuando se utiliza *pruning* que cuando no, por lo que los valores de las estimaciones no se ven afectadas por la metodología *pruning*.

En la figura **4-7** se muestra la mediana del error relativo de variar, el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el valor del efecto fijo de la variable *Total mora* sobre la dispersión, en dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la median del error relativo, además al aumentar el número de puntos de la cuadratura no se obtiene una mejora significativa de la mediana del error relativo, Otro aspecto que es importante resaltar es que las estimaciones no se ven afectadas por el hecho de utilizar la metodología *pruning* o no.

En la figura **4-8** se muestra la mediana del error relativo de variar, el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no sobre el valor de la estimación del porcentaje de ceros dentro del modelo de regresión ZOIP mixto, es estimado muy bien desde valores de tamaño de muestra pequeños, con una mediana del error relativo de alrededor del 13 %, sin embargo, se nota como a medida que el tamaño de muestra aumenta dicho error decrece rápidamente; no se nota una diferencia al variar el número de puntos de la cuadratura ni sobre la utilización de la metodología *pruning*.

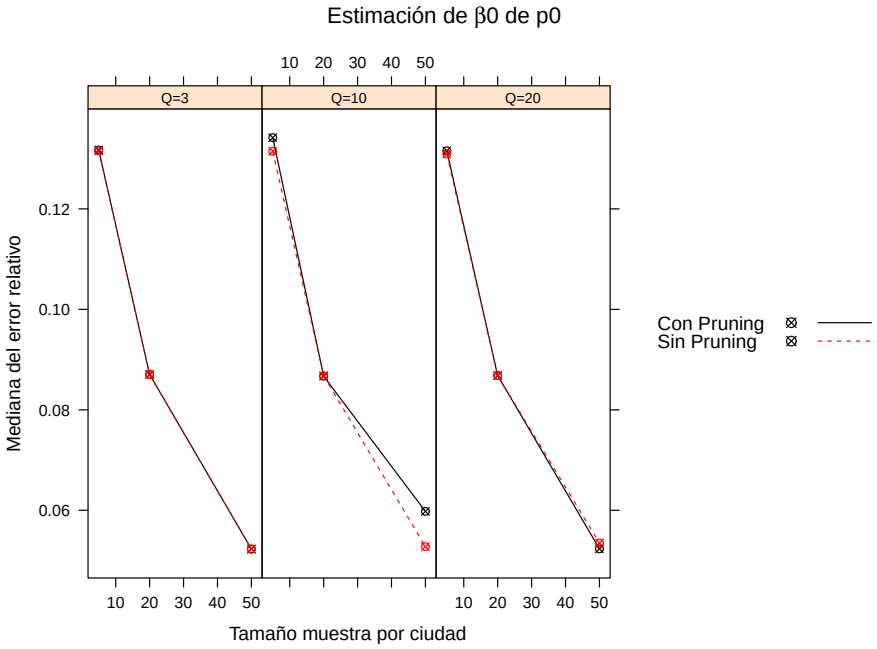

Figura 4-8: Mediana del error relativo de la estimación del parámetro $\beta 0$ asociado al parámetro de inflación de ceros, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

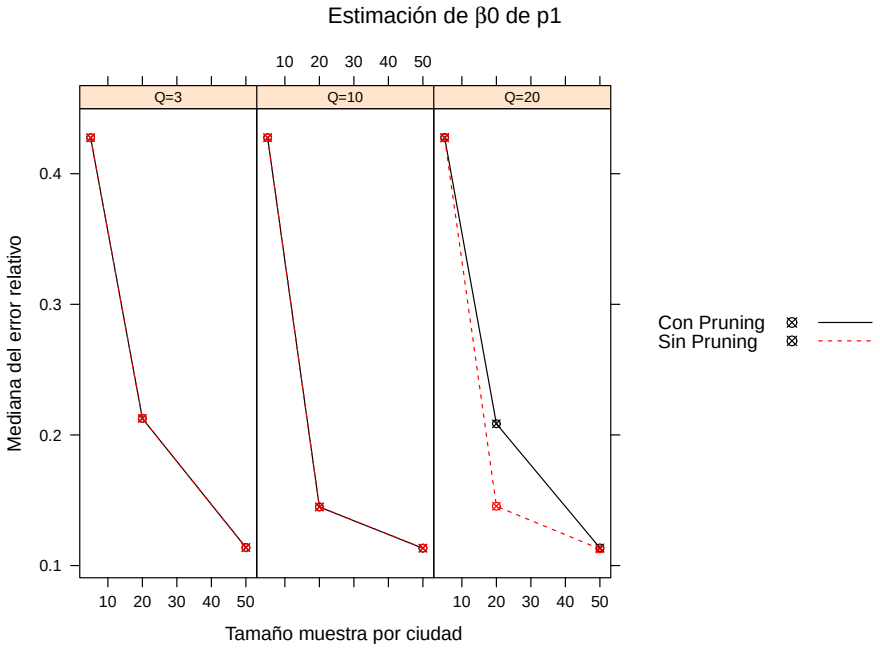

Figura 4-9: Mediana del error relativo de la estimación del parámetro $\beta 0$ asociado al parámetro de inflación de unos, variando el tamaño de muestra, el número de puntos de la cuadratura y si utiliza *pruning* o no.

En la figura **4-9** se muestra la mediana del error relativo de variar, el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el valor de la estimación del porcentaje de unos dentro del modelo de regresión ZOIP mixto, la estimación de este parámetro se ve afectado por el tamaño de muestra elegido dentro de cada grupo, ya que se nota como al aumentar el tamaño de muestra la mediana del error relativo decrece rápidamente entre el tamaño de muestra $n_i = 5$ y $n_i = 20$ y decrece medianamente rápido y entre los valores de n_i 20 y 50. Además se puede observar como no hay una diferencia entre la estimación del parámetro variando del número de puntos de la cuadratura ni tampoco sobre la utilización de la metodología *pruning*, excepto un poco $Q = 20$ y $n_i = 20$.

En la figura **4-10** se muestra la mediana del tiempo de ejecución de variar, el tamaño de muestra n_i, el número de puntos de la cuadratura Q y si se utiliza *pruning* o no, sobre el tiempo de ejecución para el ajuste del modelo de regresión ZOIP mixto, utilizando la función `RMM.ZOIP` del paquete **ZOIP**, en dicha figura se nota como a medida que se va aumentando el número de puntos de la cuadratura y el tamaño de muestra la diferencia de utilizar la metodología *pruning* se hace más favorable ya que se nota una reducción significativa en el tiempo de ejecución del modelo. Por todo el análisis previo hecho con la estimación de todos los parámetros de regresión, en el cual se notaba que por lo general sin importar cualquier

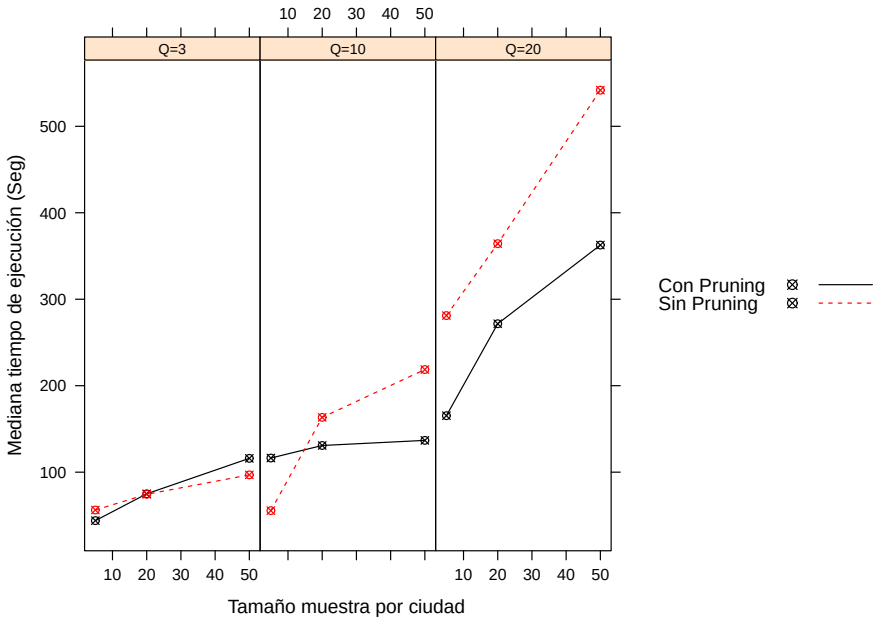

Figura 4-10: Mediana del tiempo de ejecución del modelo de regresión ZOIP mixto, bajo la función de `RMM.ZOIP` del paquete **ZOIP** de R.

combinación entre tamaño de muestra y el número de puntos de la cuadratura, la utilización de la metodología *pruning* no afecta la estimación de los parámetros de una manera significa y viendo la figura **4-10** se ve que es más conveniente utilizar la metodología *pruning*, porque esta genera un tiempo de ejecución más rápido para ajustar el modelo y no afecta de manera significativa la estimación de los efectos fijos ni de los aleatorios del modelo de regresión ZOIP mixto. Además se puede concluir que el hecho de aumentar el número de puntos de la cuadratura solo beneficia a la estimación de los efectos aleatorios mas no de los efectos fijos, en general lo más recomendable para mejorar las estimaciones de todos los parámetros es aumentar el tamaño muestra dentro de cada grupo.

4.4. Conclusión

El modelo de regresión ZOIP con efectos mixtos permite hacer ajustes a un modelo de regresión mixto, donde la variable respuesta está dada por datos proporcionales inflados con ceros y/o unos, es decir, bajo una distribución ZOIP, se considera un modelo de regresión mixto porque además de considerar efectos fijos sobre cada uno de los parámetros de la distribución ZOIP, permite incluir efectos aleatorios, tales como interceptos aleatorios sobre los parámetros de la media y la dispersión, la estimación de estos parámetros se realiza bajo máxima verosimilitud y bajo la aproximación de la cuadratura de Gauss-Hermite adaptativa

multidimensional, dicha estimación puede ser realizada en el paquete **ZOIP** de R, de una forma muy agradable para el usuario, en el paquete es posible realizar diferentes tipos de regresión ZOIP mixto, bajo diferentes distribuciones y parametrizaciones, además de obtener resultados del modelo mediante funciones de metodología S3 de R.

El modelo de regresión ZOIP mixto a manera de resultados, presenta una excelente convergencia de sus parámetros bajo los estudios de simulación realizados, además nos permite concluir que la mejor metodología para obtener un buen tiempo de ejecución y de convergencia de los parámetros del modelo de regresión ZOIP mixto, es cuando se utiliza un numero prudente de puntos de la cuadratura, es decir, entre cinco y quince puntos, utilizando la metodología *pruning* y obteniendo un buen tamaño de observaciones por cada grupo de la variable de agrupación del modelo mixto, este tamaño de muestra es el que mejor efecto obtiene sobre la mejora circunstancial de la convergencia de los parámetros regresores del modelo de regresión ZOIP mixto. Durante el estudio de simulación también se observó que el efecto de aumentar el número de puntos de la cuadratura, afecta de manera positiva a la convergencia de los parámetros asociados a los interceptos aleatorios.

5 Conclusiones y recomendaciones

5.1. Conclusiones

En este trabajo se propone una nueva clase de distribuciones y modelos de regresión para efectos fijos y mixtos para datos proporcionales inflados con ceros y/o unos, en la distribución propuesta y los dos modelos de regresión, se integran tres diferentes parametrizaciones de la distribución beta, además a manera de distribución alternativa se implementa la distribución simplex, estos modelos y distribuciones son muy flexibles, ya que en una sola distribución o modelo de regresión se obtienen diferentes opciones de modelado.

La distribución en este trabajo y en la que se basan los modelos de regresión propuestos, es la distribución ZOIP, esta distribución permite obtener diferentes distribuciones para datos proporciones inflados con ceros y/o unos, dicha distribución permite ajustarse a datos que se encuentran inflados solo con ceros o solo con unos, incluso a datos proporcionales no inflados. El modelo de regresión ZOIP es propuesto en este trabajo como el ajuste de un modelo de regresión con efectos fijos en los cuatro parámetros de la distribución ZOIP y es ajustado vía máxima verosimilitud. Por último, se propone el modelo de regresión ZOIP mixto, el cual permite ajustar un modelo de regresión con efectos fijos en todos los parámetros de la distribución ZOIP y tener en cuenta interceptos aleatorios normales en el parámetro de la media y la varianza, la estimación de los parámetros fue realizada vía máxima verosimilitud y mediante la aproximación de la función de verosimilitud, por medio de las diferentes alternativas de la cuadratura de Gauss-Hermite adaptativa multidimensional.

Es propuesto un paquete de R llamado **ZOIP**, que permite obtener valores de la función de densidad de probabilidad, funciones de distribución acumulada y función cuantil de la distribución ZOIP, además generación de valores aleatorios de dicha distribución. Por otra parte, el paquete permite ajustar distribuciones ZOIP o modelos de regresión ZOIP, mediante algoritmos de optimización, como `nlimnb` o `optim`. También, es posible ajustar los modelos de regresión ZOIP mixto, vía máxima verosimilitud y la cuadratura de Gauss-Hermite adaptativa. El paquete **ZOIP**, también incluye sobre el ajuste de los modelos, algunas funciones de método S3.

Se realizaron estudios de simulación que permitieron ver la convergencia de los parámetros en el ajuste de la distribución ZOIP, el modelo de regresión ZOIP y el modelo de regresión

ZOIP mixto, además esto permitió concluir con que alternativas se encuentran las mejores estimaciones de los parámetros regresores de los modelos. En el modelo de regresión ZOIP mixto se encontró que el hecho de utilizar la metodología *pruning* ayuda a reducir los tiempos de ajuste del modelo en un 50 % cuando se utilizan grandes tamaños de muestra y el número de puntos de la cuadratura de Gauss-Hermite es grande, además se evidencio que el aumento del número de puntos de la cuadratura afecta de manera positiva y en mayor proporción a los parámetros asociados a los interceptos aleatorios de los parámetros de la media y la varianza. Por último, se encontró en todos los escenarios de simulación para todos los modelos, que el aumento del tamaño de muestra es el mayor factor que hace que la estimación de los parámetros regresores mejoren, considerablemente.

Por ultimo se realizaron aplicaciones a datos reales, para el ajuste de distribuciones ZOIP, modelos de regresión ZOIP y modelos de regresión ZOIP mixtos sobre el porcentaje de utilización de una tarjeta de crédito asociada a una entidad financiera, en dicha aplicación se evidencio que el ajuste de los modelos fueron satisfactorios y fue posible encontrar diferentes factores que afectan al porcentaje de utilización de las tarjetas de crédito, lo que permitiría sacar conclusiones muy importantes para estrategias de mercadeo y fidelización de los clientes en la entidad financiera.

5.2. Recomendaciones

Como posibles trabajos futuros y recomendaciones se propone:

- Extender la distribución ZOIP a otras distribuciones para datos proporcionales no tomadas en cuenta en este trabajo, tales como la distribución beta-rectangular y la distribución LogitSep.
- Realizar un estudio de análisis de residuales para los parámetros estimados de los modelos de regresión ZOIP fijos y mixtos.
- Incluir un análisis de selección de variables en los modelos de regresión ZOIP fijos y mixtos.
- Incluir interceptos aleatorios en los parámetros asociados a la inflación de la distribución ZOIP, en el modelo de regresión ZOIP mixto propuesto en este trabajo.
- Incluir pendientes aleatorias en los diferentes parámetros de la regresión ZOIP.
- Extender los interceptos aleatorios en los parámetros de la media y la dispersión, a interceptos aleatorios no normales, o normales correlacionados.
- Incluir más funciones de método S3 en el paquete **ZOIP**, tales como `plot`, `predict` `AIC`, `BIC`, entre otras.

- Realizar una comparación entre la estimación de los parámetros del modelo de regresión ZOIP mixto, por metodologías bayesianas y utilizando máxima verosimilitud y la cuadratura de Gauss-Hermite adaptativa.
- Mejorar la eficiencia de los algoritmos implementados en el paquete **ZOIP**, mediante la integración del paquete **ZOIP** por las metodologías implementadas en el paquete **RCPP**.
- Estudiar e implementar otras metodologías de estimación de los parámetros, nunca exploradas sobre los modelos de regresión mixtos para datos proporcionales inflados con ceros y/o con unos, tales como la metodología bayesiana INLA o los algoritmos propuestos por Ogden (2015), sobre métodos de reducción secuencial.
- Evaluar más la simulación del modelo de regresión ZOIP mixto propuesto, para las demás parametrizaciones de la distribución ZOIP-beta y la distribución ZOIP-simplex.
- Optimizar los algoritmos propuestos en el paquete **ZOIP**, para que el tiempo de estimación los parámetros sean más cortos.
- Estudiar tópicos de los datos proporcionales inflados para datos proporcionales inflados con ceros y/o unos, como cuando más del 50 % de la muestra se encuentra inflada por ceros o por unos. se podría evaluar la eficacia y convergencia de las funciones implementadas, en el paquete **ZOIP** mediante estudios de simulación.
- Realizar mejoras continuas del paquete **ZOIP**, a medida que se vayan adquiriendo retroalimentaciones de los usuarios que utilizan el paquete.
- Extender los estudios de simulación del modelo de regresión ZOIP mixto, cuando se tienen muestras no balanceadas.
- Ampliar el tipo de funciones enlace vinculadas en el paquete **ZOIP**, para realizar los modelos de regresión, tales como funciones *probit* o *cloglog*.

Bibliografía

W. Bonat, P. Ribeiro, y W. Zeviani. Modelo beta com efeitos aleatórios. 2012.

W. Bonat, P. Ribeiro, y W. Zeviani. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics online: doi:10.1080/02664763.2014.947248*, 42(2):252–266, 2015.

S. Chacon y B. Straub. *Pro Git*. Apress, 2014.

D. Cook, R. Kieschnick, y B. McCullough. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, 15:860–867, 2008.

A.C.G de Andrade. Efeitos da especificação incorreta da função deligação no modelo der-regressão beta. dissetação de mestrado. *Institudo de Matemática e Estatística*, 2007.

M. Evans y T. Swartz. *Approximating integrals via Monte Carlo and deterministic methods*. Oxford University Press Inc., 2000.

L. Fahrmeir y G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer, 2001.

S. Ferrari y F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

S. Ferrari y E. Pinheiro. Improved likelihood inference in beta regression. *Journal of statistical computation and simulation available online:doi:10.1080/00949650903389993*, 2010.

Z. Figueroa-Zuñiga, R. arellano, y S. Ferrari. Mixed beta regression: a bayesian perspective. *Computational Statistics & Data Analysis*, 61(Supplement C):137 – 147, 2013.

D. Galvis y V. Lachos. *Bayesian analysis of regression models for proportional data in the presence of zeros and ones*. PhD thesis, Estatal University of Campinas, 2014.

E. Hahn. Mixture densities for project management activity times: A robust approach to pert. *European Journal of operational research*, 188(2):450–459, 2008.

F. Hernández, O. Usuga, y V. Giampaoli. Improving the adaptive gaussian quadrature. 2013.

A. Hossain. Modelling a proportion response variable using generalized additive models for location scale and shape. *International Workshop on Statistical Modelling*, 2:137–140, 2015.

J. Houston y C. James. Bank information monopolies and the mix of private and public debt claims. *Journal of Finance*, 51:1863–1889, 1996.

B. Jørgensen. The theory of dispersion models. *Computational statistics and Data analysis*, 76, 1997.

B. Jørgensen y Barndorff-Nielsen. Some parametric models on the simplex. *Journal of multivariate analysis*, 39(1):106–116, 1991.

R. Kieschnick y B. McCullough. Regression analysis of variates observed on (0,1):percentages, proportions, and fractions. *Statistical Modelling*, 3(3):193–213, 2003.

I. Kosmidis, A. Zou, y K. Jagan. Beyond beta regression: modelling percentages and fractions in the presence of boundary observations. *International Workshop on Statistical Modelling*, 1:267–272, 2015.

N. Laird y J. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

M. Lindstrom y D. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404): 1014–1022, 1988.

Q. Liu y D. Pierce. A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629, 1994.

H. Ogden. A sequential reduction method for inference in generalized linear mixed models. *Electronic Journal of Statistics*, 9:135–152, 2015.

R. Ospina y S. Ferrari. Inflated beta distributions online: doi:10.1007/s00362-008-0125-4. *Statitical papers*, 2010.

R. Ospina y S. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, 56(6):1609–1623, 2012.

C. Owen. Parameter estimation for the beta distribution. Master's thesis, Brigham young university, 2008.

L. Pace y A. Salvan. Principles of statistical inference from a neo-fisherian perspective. in: Advanced series on statistical science and applied probability. *World Scientific*, 4:128, 1997.

P. Paolino. Maximim likelihood estimation of models with beta-distributed dependent variables. *Polotical Analysis*, 9(4):325–346, 2001.

J. Pinheiro y D. Bates. Apprximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Grapchical Statistics*, 4(1):12–35, 1995.

Z. Qiu, P. Song, y M. Tan. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, 35(4):577–596, 2008.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL http://www.R-project.org/.

B. Rigby y M. Stasinopoulos. Generalized additive models for location, scale and shape. *Applied Statistical*, 54(3):507–554, 2005.

B. Rigby y M. Stasinopoulos. Instructions on how to use the gamlss package in r. *Computational statistics and Data analysis*, 2:194–195, 2008.

J. Seoane. ¿modelos mixtos (lineales)? una introducción para el usuario temeroso. *Temas actuales de Etología*, 2014.

A. Simas, W. Barrero-Souza, y A. Rocha. Improved estimators for a general class of beta regression models. *Computational statistics and Data analysis*, 54:348–366, 2010.

M. Smithson y J. Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.

P. Song, Z. Qi, y M. Tan. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, 46(5):540–533, 2004.

O. Usuga y V. Giampaoli. *Modelos de regressão beta com efeitos aleatórios normais e não normais para datos longitudinais*. PhD thesis, Universidade de São Paulo, 2013.

M. Venezuela. *Equacão generalizada e influencia local para modelos de regressão beta con medidas repetidas*. PhD thesis, Universidade de São Paulo, 2008.

J. Verkuilen y M. Smithson. Mixed and mixture regression model for continuos bounded responses using the beta distribution. *Journal of education and behavioral statistics*, 37 (1):82–113, 2012.

A. Zeileis, F. Cribari-Neto, y B. Grün. Beta regression in r. *Journal of statistical software*, 34(2):1–24, 2010.

P. Zhang, Z. Qiu, y C. Shi. simplexreg: An r package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Sftware*, 71(11), 2016.