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)

Su uso es opcional y cada autor podrá determinar la distribución del texto en la página, se sugiere esta presentación. En ella el autor dedica su trabajo en forma especial a personas y/o entidades.

Por ejemplo:

A mis padres

o

La preocupación por el hombre y su destino siempre debe ser el interés primordial de todo esfuerzo técnico. Nunca olvides esto entre tus diagramas y ecuaciones.

Albert Einstein

Agradecimientos

Esta sección es opcional, en ella el autor agradece a las personas o instituciones que colaboraron en la realización de la tesis o trabajo de investigación. Si se incluye esta sección, deben aparecer los nombres completos, los cargos y su aporte al documento.

Resumen

El resumen es una presentación abreviada y precisa (la NTC 1486 de 2008 recomienda revisar la norma ISO 214 de 1976). Se debe usar una extensión máxima de 12 renglones. Se recomienda que este resumen sea analítico, es decir, que sea completo, con información cuantitativa y cualitativa, generalmente incluyendo los siguientes aspectos: objetivos, diseño, lugar y circunstancias, pacientes (u objetivo del estudio), intervención, mediciones y principales resultados, y conclusiones. Al final del resumen se deben usar palabras claves tomadas del texto (mínimo 3 y máximo 7 palabras), las cuales permiten la recuperación de la información.

Palabras clave: (máximo 10 palabras, preferiblemente seleccionadas de las listas internacionales que permitan el indizado cruzado).

A continuación se presentan algunos ejemplos de tesauros que se pueden consultar para asignar las palabras clave, según el área temática:

Artes: AAT: Art y Architecture Thesaurus.
Ciencias agropecuarias: 1) Agrovoc: Multilingual Agricultural Thesaurus - F.A.O. y 2)GEMET: General Multilingual Environmental Thesaurus.
Ciencias sociales y humanas: 1) Tesauro de la UNESCO y 2) Population Multilingual Thesaurus.
Ciencia y tecnología: 1) Astronomy Thesaurus Index. 2) Life Sciences Thesaurus, 3) Subject Vocabulary, Chemical Abstracts Service y 4) InterWATER: Tesauro de IRC - Centro Internacional de Agua Potable y Saneamiento.
Tecnologías y ciencias médicas: 1) MeSH: Medical Subject Headings (National Library of Medicine's USA) y 2) DECS: Descriptores en ciencias de la Salud (Biblioteca Regional de Medicina BIREME-OPS).
Multidisciplinarias: 1) LEMB - Listas de Encabezamientos de Materia y 2) LCSH- Library of Congress Subject Headings.

También se pueden encontrar listas de temas y palabras claves, consultando las distintas bases de datos disponibles a través del Portal del Sistema Nacional de Bibliotecas[1], en la sección Recursos bibliográficos.ºpción "Bases de datos".

Abstract

Es el mismo resumen pero traducido al inglés. Se debe usar una extensión máxima de 12

[1] ver: www.sinab.unal.edu.co

renglones. Al final del Abstract se deben traducir las anteriores palabras claves tomadas del texto (mínimo 3 y máximo 7 palabras), llamadas keywords. Es posible incluir el resumen en otro idioma diferente al español o al inglés, si se considera como importante dentro del tema tratado en la investigación, por ejemplo: un trabajo dedicado a problemas lingüísticos del mandarín seguramente estaría mejor con un resumen en mandarín.

Keywords: palabras clave en inglés(máximo 10 palabras, preferiblemente seleccionadas de las listas internacionales que permitan el indizado cruzado)

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

En la introducción, el autor presenta y señala la importancia, el origen (los antecedentes teóricos y prácticos), los objetivos, los alcances, las limitaciones, la metodología empleada, el significado que el estudio tiene en el avance del campo respectivo y su aplicación en el área investigada. No debe confundirse con el resumen y se recomienda que la introducción tenga una extensión de mínimo 2 páginas y máximo de 4 páginas.

La presente plantilla maneja una familia de fuentes utilizada generalmente en LaTeX, conocida como Computer Modern, específicamente LMRomanM para el texto de los párrafos y CMU Sans Serif para los títulos y subtítulos. Sin embargo, es posible sugerir otras fuentes tales como Garomond, Calibri, Cambria, Arial o Times New Roman, que por claridad y forma, son adecuadas para la edición de textos académicos.

La presente plantilla tiene en cuenta aspectos importantes de la Norma Técnica Colombiana - NTC 1486, con el fin que sea usada para la presentación final de las tesis de maestría y doctorado y especializaciones y especialidades en el área de la salud, desarrolladas en la Universidad Nacional de Colombia.

Las márgenes, numeración, tamaño de las fuentes y demás aspectos de formato, deben ser conservada de acuerdo con esta plantilla, la cual esta diseñada para imprimir por lado y lado en hojas tamaño carta. Se sugiere que los encabezados cambien según la sección del documento (para lo cual esta plantilla esta construida por secciones).

Si se requiere ampliar la información sobre normas adicionales para la escritura se puede consultar la norma NTC 1486 en la Base de datos del ICONTEC (Normas Técnicas Colombianas) disponible en el portal del SINAB de la Universidad Nacional de Colombia[1], en la sección Recursos bibliográficos.ºpción "Bases de datos". Este portal también brinda la posibilidad de acceder a un instructivo para la utilización de Microsoft Word y Acrobat Professional, el cual está disponible en la sección "Servicios", opción "Trámitesz enlace .Entrega de tesis".

La redacción debe ser impersonal y genérica. La numeración de las hojas sugiere que las páginas preliminares se realicen en números romanos en mayúscula y las demás en números

[1] ver: www.sinab.unal.edu.co

arábigos, en forma consecutiva a partir de la introducción que comenzará con el número 1. La cubierta y la portada no se numeran pero si se cuentan como páginas.

Para trabajos muy extensos se recomienda publicar más de un volumen. Se debe tener en cuenta que algunas facultades tienen reglamentada la extensión máxima de las tesis o trabajo de investigación; en caso que no sea así, se sugiere que el documento no supere 120 páginas.

No se debe utilizar numeración compuesta como 13A, 14B ó 17 bis, entre otros, que indican superposición de texto en el documento. Para resaltar, puede usarse letra cursiva o negrilla. Los términos de otras lenguas que aparezcan dentro del texto se escriben en cursiva.

2. Capítulo 1: 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 and Cribari-Neto (2004) y Rigby and 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 and 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 and 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 and 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 and Ferrari (2012). Galvis and 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 and 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 and 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 artículo se encuentra organizado de la siguiente manera: primero se presentan las distribuciones más representativas para datos proporcionales, en la sección 3 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 and Cribari-Neto (2004) y Rigby and Stasinopoulos (2005)) y la distribución simplex propuesta por Jørgensen and 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 and 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 and 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 and 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 and Stasinopoulos (2005) cuando:

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

2.1.4. Distribución simplex

La distribución simplex fue introducida por Jørgensen and 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 $y \sim \text{ZOIP}(\mu, \sigma, p_0, p_1)$, 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 parametri- zación

Ferrari and 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 and Cribari-Neto (2004) y (2-12), (2-13) para el caso Rigby and 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 2 casos:

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

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

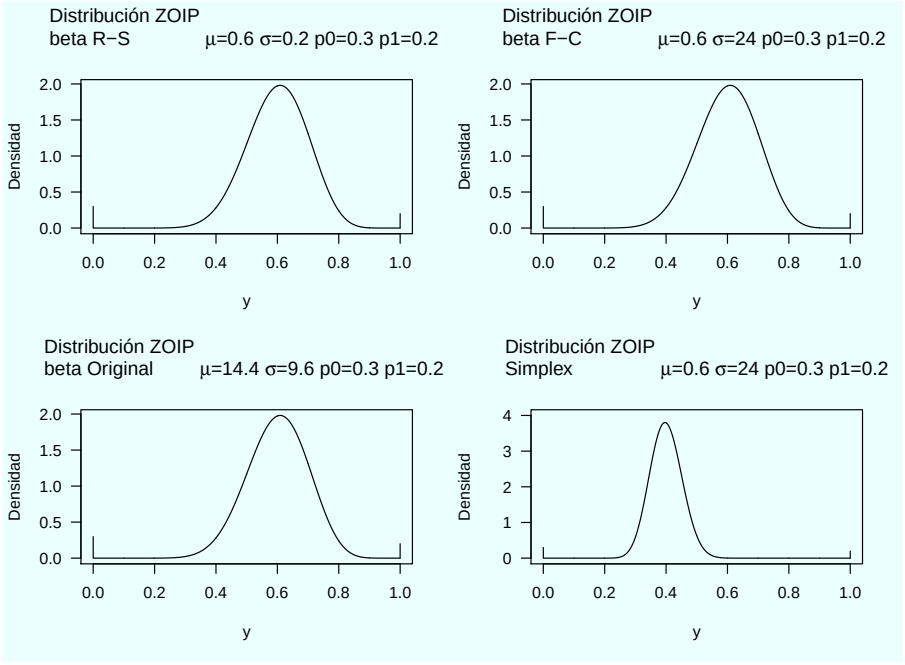

Figura 2-1.: Densidades para la distribución ZOIP para algunos valores de los parámetros, donde R-S se refiere a Rigby & Stasinopoulos (2005) y F-C es Ferrari & Cribari-Neto (2004).

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

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 and 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 and Cribari-Neto (2004) y Rigby and 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 and Cribari-Neto (2004) y (2-12), (2-13) para el caso de Rigby and 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 and 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 and 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 and Stasinopoulos (2005), si toma el valor de `'F-C'` se utilizará la distribución ZOIP-beta parametrización

Ferrari and 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 and Stasinopoulos (2005).

```
yi <- as.data.frame(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')
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
------------------------------------------------------------
------------------------------------------------------------
```

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`.

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 and 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 and 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 and 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

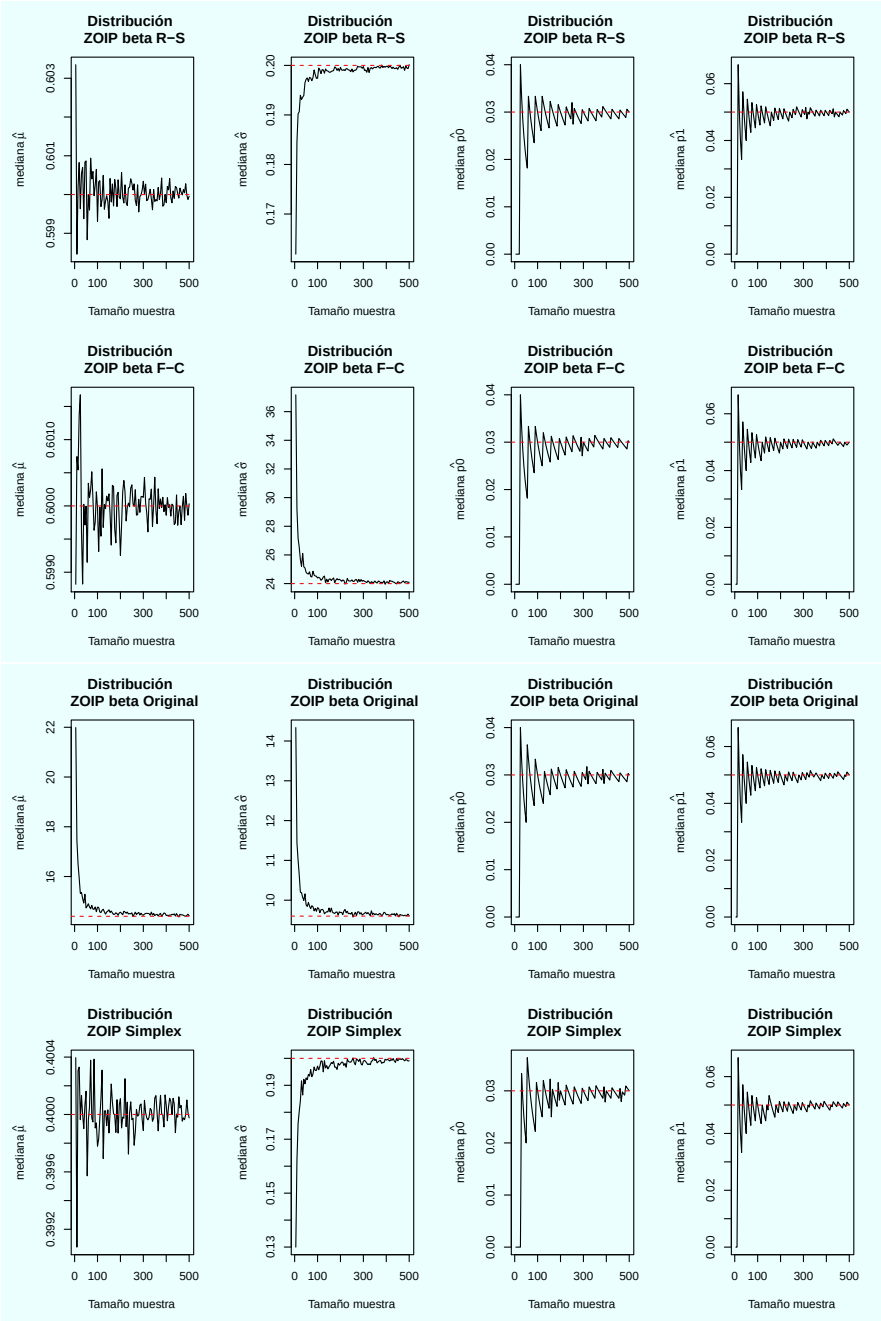

Figura 2-2.: Mediana de los parámetros estimados en el escenario 1 para distintas parametrizaciones y valores de n, las líneas rojas representan el verdadero valor del parámetro.

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	7
	Caso original	15.7	7
	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

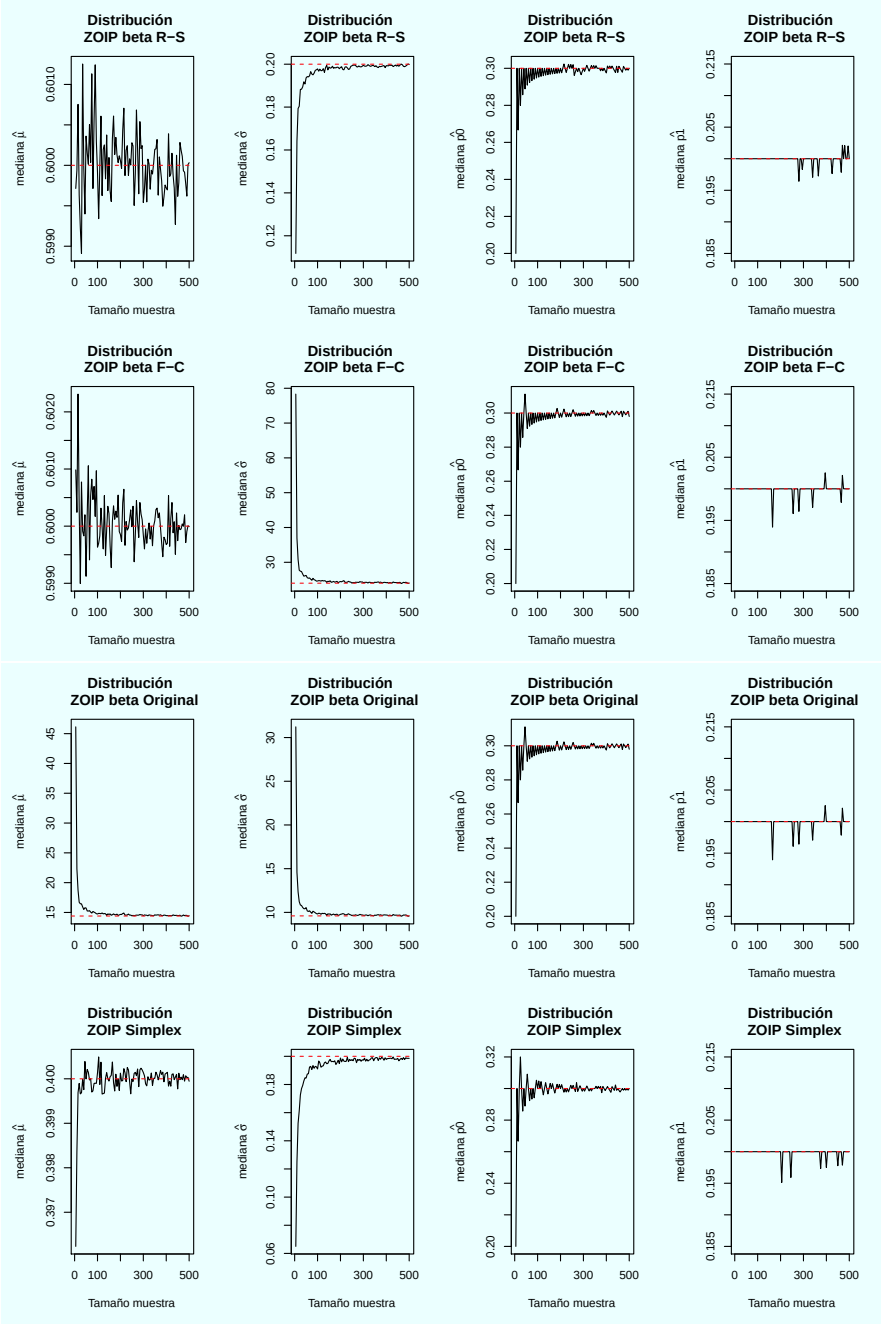

Figura 2-3.: Simulación de distribución ZOIP para distintas parámetrizaciones con parámetros de inflación grandes, distribuciones y valores de n.

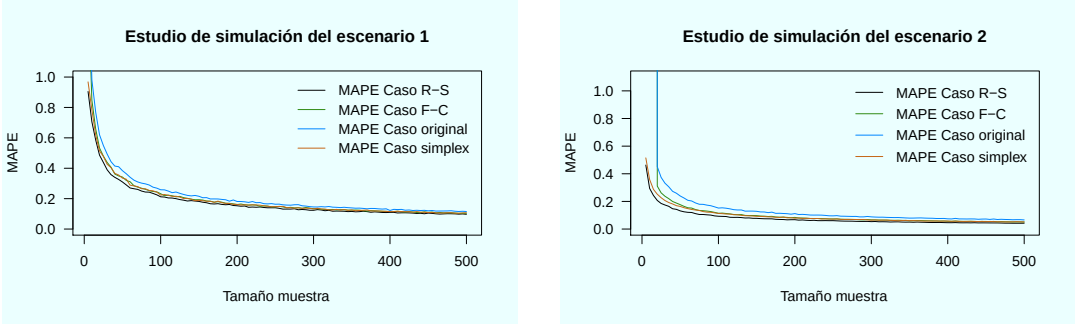

Figura 2-4.: Mape (Error porcentual absoluto medio) para los dos escenarios de simulación y para distintas parametrizaciones y valores de n.

la distribución beta con parametrización Rigby and 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

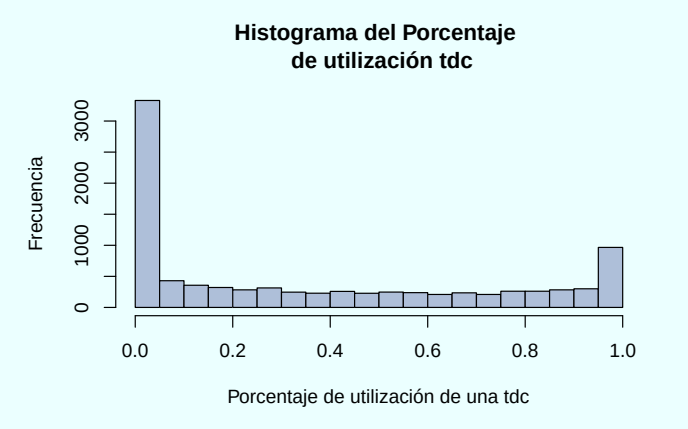

Figura 2-5.: Histograma del porcentaje de utilización de las tdc en un banco.

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	4885.4370	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 and Stasinopoulos (2005) y Ferrari and 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

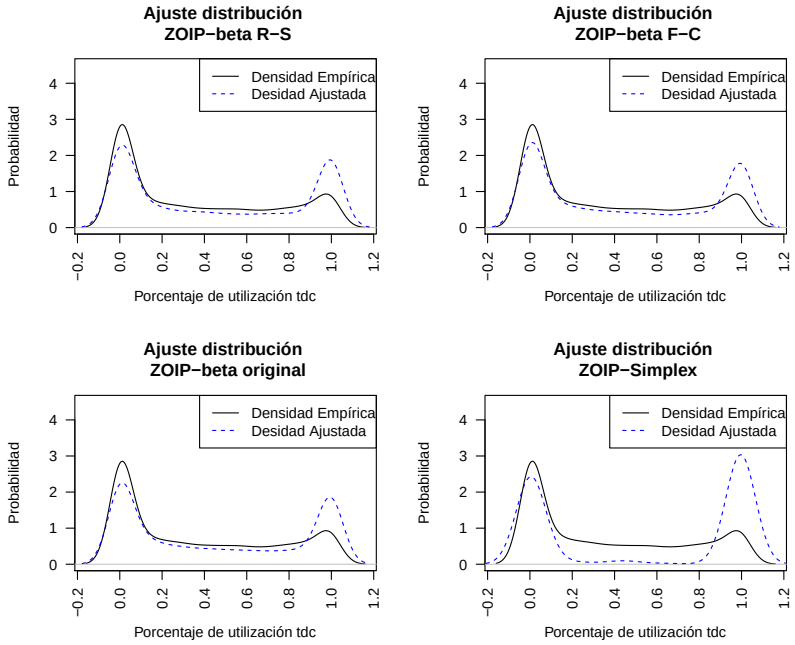

Figura 2-6.: Ajuste de diferentes distribuciones y parametrizaciones ZOIP al porcentaje de utilización de una tdc.

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 2: 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 and 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 and 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 and 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 and Ferrari (2012) realizan un modelo de regresión inflado con cero o con uno, no con ambos, bajo la distribución beta inflada de Ospina and Ferrari (2010) con parametrización Ferrari and Cribari-Neto (2004), de igual manera Rigby and 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 and 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 and 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 and 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 mediante 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 capítul 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 presenta 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 parÃ¡metro, 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 **?**.

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 and 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 and 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 and 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 regresion 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 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')
```

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 and Stasinopoulos (2005), si toma el valor de `'F-C'` se utilizara la distribución beta parametrización Ferrari and Cribari-Neto (2004), el valor de `'Original'` se utilizara la distribución beta con parametrización original, `'Simplex'` Utilizara 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 and 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.

```
devtools::install_github('jucdiaz/ZOIP', force=TRUE)
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')
summary(mod)
```

Los resultados obtenidos se muestran a continuación.

```
-------------------------------------------------------------
Fixed effects for logit(mu)
-------------------------------------------------------------
            Estimate Std. Error z value  Pr(>|z|)
(intercept)  0.35500    0.09953  3.5668 0.0003614 ***
x1          -1.52936    0.18172 -8.4158 < 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)  0.30456    0.10158  2.9983  0.002715 **
x1           2.67959    0.13954 19.2030 < 2.2e-16 ***
x2           1.03339    0.12824  8.0585 7.723e-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) 0.0710777  0.0081293  8.7434 < 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) -0.20773    0.14623 -1.4205   0.1555
x2          -3.15715    0.33901 -9.3128   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
-------------------------------------------------------------
-------------------------------------------------------------
```

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.

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \epsilon_i, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \\ h_3(p_0) &= \beta_0 + \epsilon_i, \\ h_4(p_1) &= \beta_0 + \beta_1 x_{2i} + \epsilon_i \end{aligned} \tag{3-3}$$

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 and 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**.

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 and 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 parámet-

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.

ros 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 and 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 promedio, 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

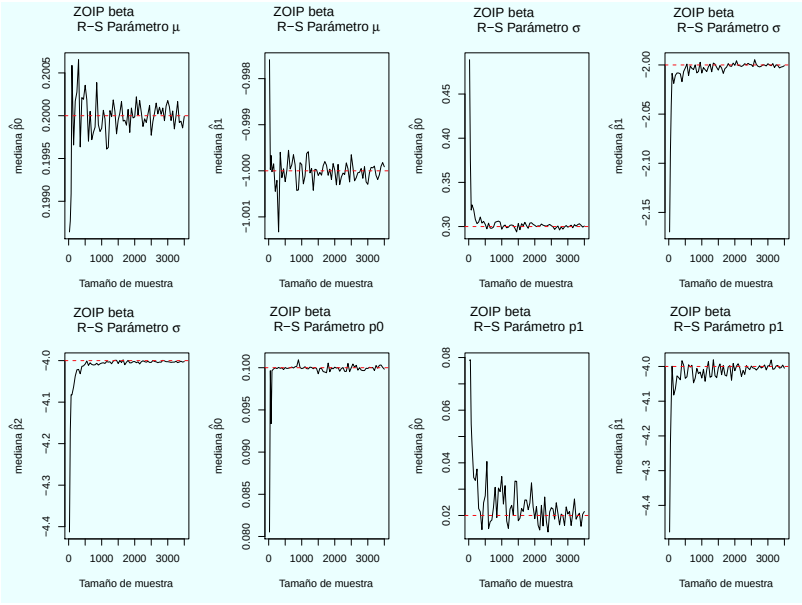

Figura 3-1.: Simulación de un modelo de regresión ZOIP-beta para la parametricazion R-S con diferentes valores de n.

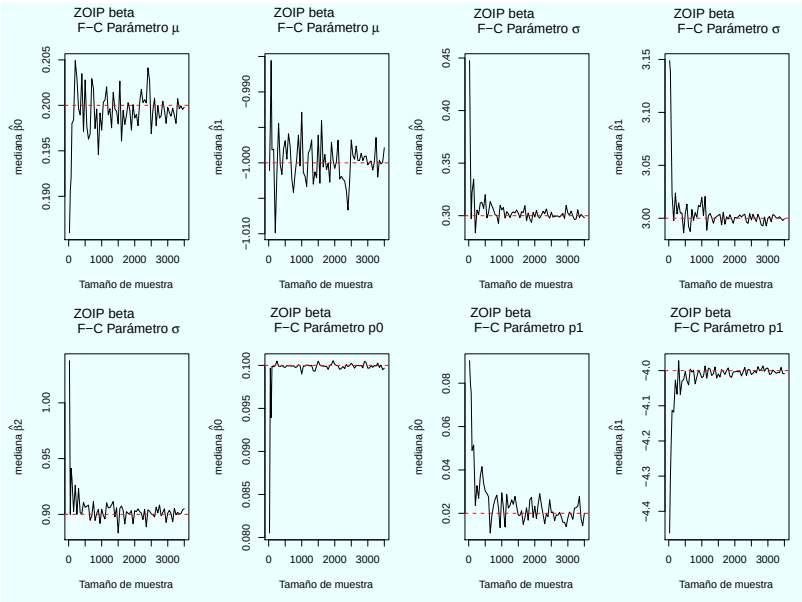

Figura 3-2.: Simulación de un modelo de regresión ZOIP-beta para la parametricazion F-C con diferentes valores de n.

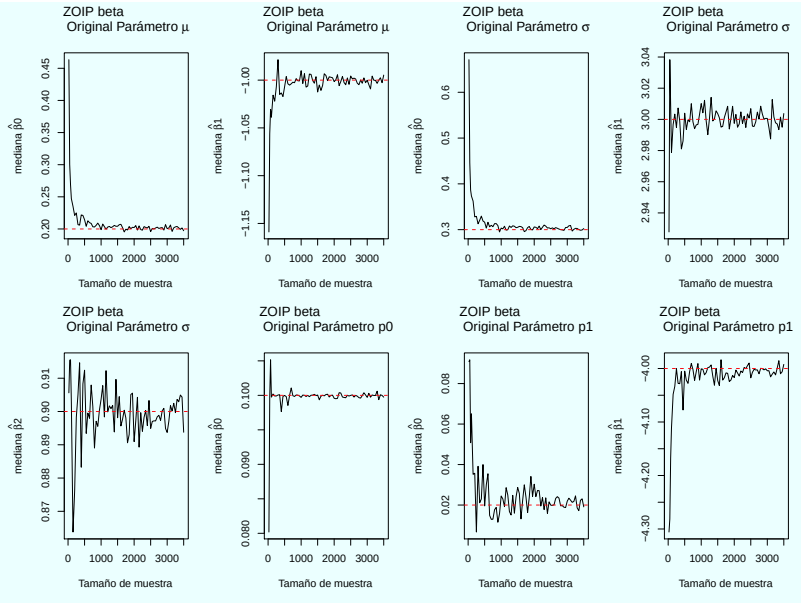

Figura 3-3.: Simulación de un modelo de regresión ZOIP-beta para la parametricazion original con diferentes valores de n.

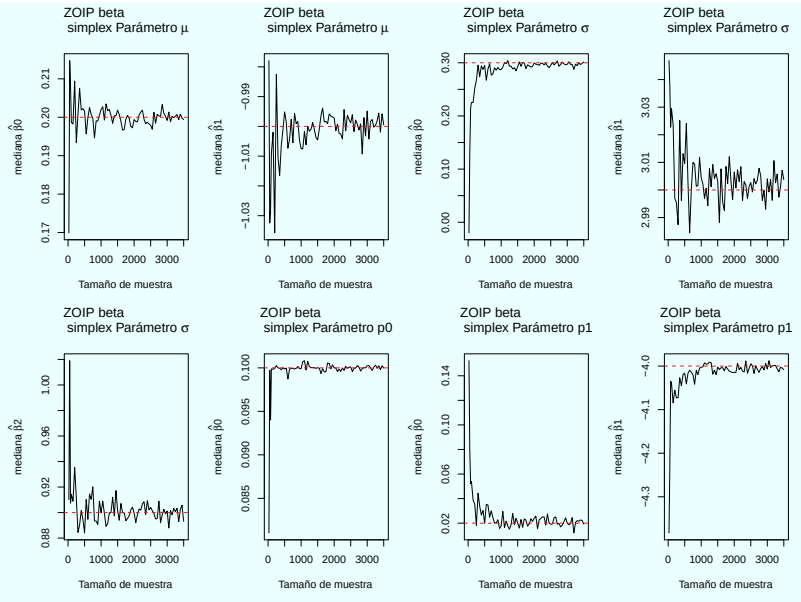

Figura 3-4.: Simulación de un modelo de regresión ZOIP-simplex con diferentes valores de n.

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 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 and 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 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

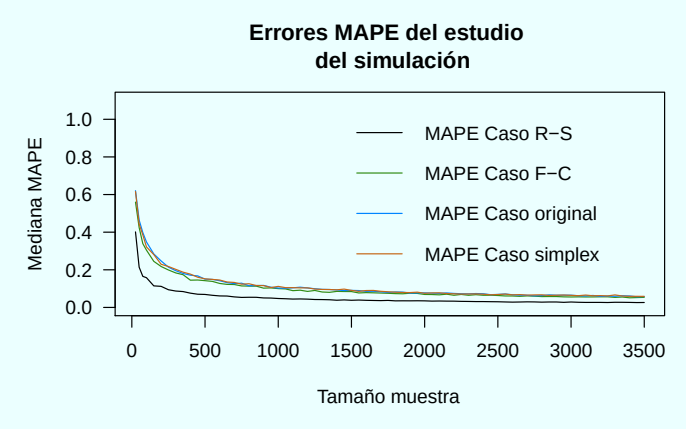

Figura 3-5.: Mape (Error porcentual absoluto medio) para modelo de regresion ZOIP simulado para distintas parametrizaciones y valores de n.

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.

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \\ h_3(p_0) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \\ h_4(p_1) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \end{aligned} \tag{3-4}$$

donde 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.

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 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

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.

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 estarán 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

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.

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 and 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 and Cribari-Neto (2004) seguido por la parametrización original y por último la parametrizacion de Rigby and 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 3: Modelo de regresión ZOIP con efectos mixtos

Introduccion.....

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:

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

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

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

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:

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

$$h_2(\sigma_{ij}) = \mathbf{x}_{ij2}^\top \boldsymbol{\beta}_2 + \gamma_{i2},$$

$$h_3(p_0) = \mathbf{x}_{ij3}^\top \boldsymbol{\beta}_3,$$

$$h_4(p_1) = \mathbf{x}_{ij4}^\top \boldsymbol{\beta}_4$$

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

La estimación de los parámetros del modelo de regresión con intercepto aleatorio 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 utilizamos 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 **?**; (**?**), 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 $\mathbf{A}$ está basado similarmente en el producto Kronecker, denotado por $\otimes$, los pesos originales $\mathbf{W}$, es dado:

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

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

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 corresponde 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 dado por:

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

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 optimizacion mas elevado, debido a la transformacion de los puntos de cuadratura, pero no todos los puntos de la AGQ aportan de manera significativa un valor sobre la solucion de la integral, es por esto que **?** desarrolla un mejoramiento de la AGQ, donde elimina los puntos de la cuadratura que no son significativos sobre la solucion de la integral, de modo que no afectan de manera significaativa los resultados finales de la integral, pero si afectan de manera positiva el tiempo de ejecucion, 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 dado por 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 minimo y la mediana de los pesos originales $\mathbf{W}$. (Ver más en **?**.)

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

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

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 dado 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 dado 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`.

5. Capítulo ...

Se deben incluir tantos capítulos como se requieran; sin embargo, se recomienda que la tesis o trabajo de investigación tenga un mínimo 3 capítulos y máximo de 6 capítulos (incluyendo las conclusiones).

6. Conclusiones y recomendaciones

6.1. Conclusiones

Las conclusiones constituyen un capítulo independiente y presentan, en forma lógica, los resultados de la tesis o trabajo de investigación. Las conclusiones deben ser la respuesta a los objetivos o propósitos planteados. Se deben titular con la palabra conclusiones en el mismo formato de los títulos de los capítulos anteriores (Títulos primer nivel), precedida por el numeral correspondiente (según la presente plantilla).

6.2. Recomendaciones

Se presentan como una serie de aspectos que se podrían realizar en un futuro para emprender investigaciones similares o fortalecer la investigación realizada. Deben contemplar las perspectivas de la investigación, las cuales son sugerencias, proyecciones o alternativas que se presentan para modificar, cambiar o incidir sobre una situación específica o una problemática encontrada. Pueden presentarse como un texto con características argumentativas, resultado de una reflexión acerca de la tesis o trabajo de investigación.

A. Anexo: Nombrar el anexo A de acuerdo con su contenido

Los Anexos son documentos o elementos que complementan el cuerpo de la tesis o trabajo de investigación y que se relacionan, directa o indirectamente, con la investigación, tales como acetatos, cd, normas, etc.

B. Anexo: Nombrar el anexo B de acuerdo con su contenido

A final del documento es opcional incluir índices o glosarios. Éstos son listas detalladas y especializadas de los términos, nombres, autores, temas, etc., que aparecen en el mismo. Sirven para facilitar su localización en el texto. Los índices pueden ser alfabéticos, cronológicos, numéricos, analíticos, entre otros. Luego de cada palabra, término, etc., se pone coma y el número de la página donde aparece esta información.

C. Anexo: Nombrar el anexo C de acuerdo con su contenido

MANEJO DE LA BIBLIOGRAFÍA: la bibliografía es la relación de las fuentes documentales consultadas por el investigador para sustentar sus trabajos. Su inclusión es obligatoria en todo trabajo de investigación. Cada referencia bibliográfica se inicia contra el margen izquierdo.

La NTC 5613 establece los requisitos para la presentación de referencias bibliográficas citas y notas de pie de página. Sin embargo, se tiene la libertad de usar cualquier norma bibliográfica de acuerdo con lo acostumbrado por cada disciplina del conocimiento. En esta medida es necesario que la norma seleccionada se aplique con rigurosidad.

Es necesario tener en cuenta que la norma ISO 690:1987 (en España, UNE 50-104-94) es el marco internacional que da las pautas mínimas para las citas bibliográficas de documentos impresos y publicados. A continuación se lista algunas instituciones que brindan parámetros para el manejo de las referencias bibliográficas:

~~Institución~~	Disciplina de aplicación
Modern Language Association (MLA)	Literatura, artes y humanidades
American Psychological Association (APA)	Ambito de la salud (psicología, medicina) y en general en todas las ciencias sociales
Universidad de Chicago/Turabian	Periodismo, historia y humanidades.
AMA (Asociación Médica de los Estados Unidos)	Ambito de la salud (psicología, medicina)
Vancouver	Todas las disciplinas
Council of Science Editors (CSE)	En la actualidad abarca diversas ciencias
National Library of Medicine (NLM) (Biblioteca Nacional de Medicina)	En el ámbito médico y, por extensión, en ciencias.
Harvard System of Referencing Guide	Todas las disciplinas
JabRef y KBibTeX	Todas las disciplinas

Para incluir las referencias dentro del texto y realizar lista de la bibliografía en la respectiva sección, puede utilizar las herramientas que Latex suministra o, revisar el instructivo desar-

rollado por el Sistema de Bibliotecas de la Universidad Nacional de Colombia[1], disponible en la sección "Servicios", opción "Trámitesz enlace .[En]trega de tesis".

[1]Ver: www.sinab.unal.edu.co

Bibliografía

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

D. Galvis and 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.

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.

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

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

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

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

R. Ospina and 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 and 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.

Z. Qiu, P. Song, and 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 and M. Stasinopoulos. Generalized additive models for location, scale and shape. *Applied Statistical*, 54(3):507–554, 2005.

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

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

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