



UNIVERSIDAD NACIONAL DE COLOMBIA

Modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos y creación del paquete **ZOIP** en **R**

Juan Camilo Díaz Zapata

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2018

Modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos y creación del paquete **ZOIP** en **R**

Juan Camilo Díaz Zapata

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

Director:
Freddy Hernández Barajas Ph.D.

Línea de Investigación:
Análisis multivariado

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2018

(Dedicatoria)

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

Agradecimientos

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

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

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

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

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

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

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

Resumen

El modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos, es un modelo de regresión donde la variable respuesta se encuentra definida a partir de una distribución para datos proporcionales, como la distribución beta o la distribución simplex, que dan resultados en el intervalo cero uno, más dos valores dados por cero y/o uno, representando la ausencia o presencia total de cierta característica. Diferentes autores han trabajado en el desarrollo de diferentes modelos y metodologías de estimación, sin embargo, no se ha desarrollado un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos, que reúna los principales modelos de regresión de este tipo y que la estimación de los parámetros sea vía máxima verosimilitud y la cuadratura de Gauss-Hermite. En este trabajo se presenta el paquete **ZOIP** del sistema computacional R, alojado en el **CRAN** y en **GitHub**, en él que se implementa la distribución ZOIP (Zeros Ones Inflated Proporcional), el modelo de regresión para efectos fijos y mixtos, ZOIP, que reúne las distribuciones y los modelos de regresión de efectos fijos y mixtos para las distribuciones beta en tres diferentes parametrizaciones y la distribución simplex, inflada con ceros y/o unos, la estimación de los parámetros se hace vía máxima verosimilitud y la cuadratura de Gauss-Hermite utilizando diferentes alternativas, como la cuadratura de Gauss-Hermite adaptativa con o sin *pruning*. Se realizan tres estudios de simulación que muestran la convergencia de los parámetros para el ajuste de una distribución ZOIP y los diferentes casos de uso de los modelos de regresión ZOIP con efectos fijos y mixtos, a través de los estudios de simulación se presenta las ventajas y desventajas de los métodos de estimación propuestos y la alternativa de estimación que obtiene mejor desempeño, para cada modelo presentado. Además, se ajustan diferentes modelos de regresión ZOIP a un caso real sobre el porcentaje de uso de las tarjetas de crédito en una entidad bancaria colombiana y la dependencia a diferentes variables de negocio que fueron definidas previamente como efectos fijos y aleatorios, dependiendo del caso de estudio.

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

Abstract

The mixed regression model for proportional data inflated with zeros and/or ones is a regression model where the response variable is defined by a proportional data distribution. A proportional data distribution could be the beta distribution or the simplex distribution, which its results are given in the interval zero one, and in other two values given by zero and/or one that represent the absence or total presence of a given characteristic. Although several authors have worked on the development of different models and estimation methodologies, a mixed regression model for proportional data inflated with zeros and/or ones that aggregates the main regression models and the estimation of the parameters through the

maximum likelihood and Gauss-Hermite quadrature has not been developed. In this paper, it is presented the ZOIP package for the computational system R, hosted in the **CRAN** and in **GitHub**, in which the ZOIP (Zeros Ones Inflated Proportional) distribution is implemented. In the ZOIP package is used a regression model for fixed and mixed effects, which unifies the distributions and these types of regressions for beta distribution with three different cases of parameters and simplex distribution, inflated with zeros and/or ones, and the estimation of the parameters solved by the maximum likelihood and Gauss-Hermite quadrature using different choices, like the Gauss-Hermite adaptive quadrature with *pruning* or without *pruning*. Three simulation studies are performed showing the convergence of the parameters for the fit of a ZOIP distribution and the different cases of use of the ZOIP regression model with fixed and mixed effects, with this simulation studies show the advantages and disadvantages of the proposed estimation methods and the estimation alternative that presents the best performance. Furthermore, different ZOIP regression models are fitted to a real case about percentage of use of credit cards in a Colombian bank and your dependence with different business variables that were previously defined as fixed and random effects, depending on the case of study.

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

Contenido

| | |
|--|------------|
| Agradecimientos | vii |
| Resumen | ix |
| 1. Introducción | 2 |
| 2. Capítulo 2: Distribución ZOIP | 5 |
| 2.1. Distribución para datos proporcionales | 6 |
| 2.1.1. Distribución beta original | 6 |
| 2.1.2. Distribución beta con la parametrización de Ferrari y Cribari-Neto (2004) | 7 |
| 2.1.3. Distribución beta con la parametrización de Rigby y Stasinopoulos (2005) | 7 |
| 2.1.4. Distribución simplex | 8 |
| 2.2. Distribución ZOIP (Zeros Ones Inflated Proporcional) | 8 |
| 2.3. Inferencia estadística | 10 |
| 2.4. Distribución ZOIP en el paquete ZOIP | 14 |
| 2.4.1. Instalación | 14 |
| 2.4.2. Funciones sobre distribución ZOIP | 15 |
| 2.4.3. Función RM.ZOIP del paquete ZOIP | 16 |
| 2.5. Aplicación | 19 |
| 2.5.1. Datos simulados | 20 |
| 2.5.2. Datos reales | 24 |
| 2.6. Conclusiones | 26 |
| 3. Capítulo 3: Modelo de regresión ZOIP con efectos fijos | 28 |
| 3.1. Modelo de regresión ZOIP | 29 |
| 3.2. Modelo de regresión ZOIP en el paquete ZOIP | 31 |
| 3.3. Aplicación | 36 |
| 3.3.1. Datos simulados | 36 |
| 3.3.2. Datos reales | 42 |
| 3.4. Conclusión | 46 |

| | |
|--|-----------|
| 4. Capítulo 4: Modelo de regresión ZOIP con efectos mixtos | 47 |
| 4.1. Modelo de regresión ZOIP mixto | 49 |
| 4.1.1. Inferencia estadística | 50 |
| 4.1.2. Cuadratura de Gauss-Hermite | 51 |
| 4.1.3. Aproximación de la función de verosimilitud vía cuadratura de Gauss-Hermite | 53 |
| 4.2. Modelo de regresión ZOIP mixto en el paquete ZOIP | 55 |
| 4.3. Aplicación | 62 |
| 4.3.1. Datos reales | 62 |
| 4.3.2. Datos simulados | 64 |
| 4.4. Conclusión | 73 |
| 5. Conclusiones y recomendaciones | 76 |
| 5.1. Conclusiones | 76 |
| 5.2. Recomendaciones | 77 |
| Bibliografía | 80 |

Lista de Figuras

| | | |
|------|--|----|
| 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). | 10 |
| 2-2. | Mediana de los parámetros estimados en el escenario 1 para distintas parametrizaciones y valores de n , las líneas punteadas representan el verdadero valor del parámetro. | 21 |
| 2-3. | Simulación de distribución ZOIP para distintas parametrizaciones con parámetros de inflación grandes, distribuciones y valores de n | 23 |
| 2-4. | MAPE (Error porcentual absoluto medio) para los dos escenarios de simulación y para distintas parametrizaciones y valores de n | 24 |
| 2-5. | Histograma del porcentaje de utilización de las tdc en un banco. | 25 |
| 2-6. | Ajuste de diferentes distribuciones y parametrizaciones ZOIP al porcentaje de utilización de una tdc. | 26 |
| 3-1. | Simulación de un modelo de regresión ZOIP-beta para la parametrización R-S con diferentes valores de n | 38 |
| 3-2. | Simulación de un modelo de regresión ZOIP-beta para la parametrización F-C con diferentes valores de n | 39 |
| 3-3. | Simulación de un modelo de regresión ZOIP-beta para la parametrización original con diferentes valores de n | 39 |
| 3-4. | Simulación de un modelo de regresión ZOIP-simplex con diferentes valores de n | 40 |
| 3-5. | Mape (Error porcentual absoluto medio) para modelo de regresion ZOIP simulado para distintas parametrizaciones y valores de n | 41 |
| 4-1. | Ajuste de la distribución exponencial a la variable <i>total mora</i> | 65 |
| 4-2. | Mediana del error relativo para la estimación del parámetro β_{10} asociado a μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 67 |
| 4-3. | Mediana del error relativo para la estimación del parámetro β_{11} asociado a μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 68 |

| | | |
|-------|---|----|
| 4-4. | Mediana del error relativo para la estimación del parámetro λ_1 desviación estándar del intercepto aleatorio asociado a la μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 69 |
| 4-5. | Mediana del error relativo para la estimación del parámetro β_{20} asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 70 |
| 4-6. | Mediana del error relativo para la estimación del parámetro β_{21} asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 71 |
| 4-7. | Mediana del error relativo para la estimación del parámetro λ_2 desviación estándar del intercepto aleatorio asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 72 |
| 4-8. | Mediana del error relativo para la estimación del parámetro β_{30} asociado al parámetro de inflación de ceros, variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 73 |
| 4-9. | Mediana del error relativo para la estimación del parámetro β_{40} asociado al parámetro de inflación de unos, variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de <i>pruning</i> | 74 |
| 4-10. | Mediana del tiempo de ejecución del modelo de regresión ZOIP mixto, bajo la función de RMM.ZOIP del paquete ZOIP de R. | 74 |

Lista de Tablas

| | |
|---|----|
| 2-1. MAPE de las estimaciones para cada parámetro en diferentes parametrizaciones en los dos estudios de simulación. | 22 |
| 2-2. Ajuste de diferentes distribuciones ZOIP en el porcentaje de utilización de una tdc, en un banco colombiano. Fuente: Cortesía del banco. | 25 |
| 3-1. Funciones de enlace adecuadas para cada parámetro, según su distribución y/o parametrización. | 37 |
| 3-2. Mediana del MAPE (Error porcentual absoluto medio) en porcentaje para los diferentes Parámetros en las diferentes parametrizaciones. | 41 |
| 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. | 43 |
| 3-4. Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$ | 43 |
| 3-5. Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Ferrari y Cribari-Neto (2004) en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$ | 44 |
| 3-6. Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización original en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$ | 44 |
| 3-7. Parámetros regresores estimados de un modelo de regresión ZOIP-simplex en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $2e^{-10}$ | 45 |
| 4-1. Estimación de los efectos fijos y los componentes de varianza λ_1 y λ_2 del modelo de regresión ZOIP mixto para el porcentaje utilización de una tdc. | 64 |
| 4-2. Mediana de los parámetros estimados en el modelo ZOIP mixto para tres tamaños de muestra y con la estrategia de “con y sin <i>pruning</i> ” y para todos los valores de Q | 66 |
| 4-3. Mediana de los parámetros estimados en el modelo ZOIP mixto para tres diferentes números de puntos de la cuadratura de Gauss-Hermite y con la estrategia de “con y sin <i>pruning</i> ” y para todos los tamaños de muestra. | 66 |

1 Introducción

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

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

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

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

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

Teniendo en cuenta lo anterior, surge la pregunta de cómo se podría realizar un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos, sobre las distribuciones beta o simplex, donde la estimación de sus parámetros sea vía máxima verosimilitud por medio de diferentes variantes de la cuadratura de Gauss-Hermite adaptativa. Debido a esto en este trabajo se incluye en una sola distribución y en un solo modelo de regresión de efectos fijos y mixtos, las principales distribuciones y modelos de regresión para datos proporcionales inflados con ceros y/o unos, dando como resultado la distribución ZOIP (Zeros Ones Inflated Proporcional), los modelos de regresión ZOIP para efectos fijos y mixtos. La estimación de los parámetros regresores de los modelos de regresión ZOIP, son realizados vía máxima verosimilitud a través de las diferentes variaciones de la cuadratura de Gauss-Hermite adaptativa, para esta distribución y los diferentes modelos de regresión propuestos, se le realizan estudios de simulación que demuestran la convergencia satisfactoria de sus parámetros. Adicional a esto, cómo no existe un paquete en R que reúna las principales distribuciones y modelos de regresión de efectos fijos y mixtos para modelar los datos proporcionales inflados con ceros y/o unos, se implementa el paquete **ZOIP**, que permite generar y ajustar distribuciones y modelos de regresión para efectos fijos y mixtos para datos proporcionales inflados con ceros y/o unos por medio de la metodología propuesta.

Organización del trabajo

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

Un modelo de regresión de efectos fijos para datos proporcionales es planteado en el capítulo

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

En el capítulo 4 se muestra la implementación y el desarrollo analítico de un modelo de regresión para datos proporcionales inflados con ceros y/o unos, teniendo en cuenta efectos fijos e interceptos aleatorios en los parámetros de la media y la varianza, se muestra la derivación de la función de verosimilitud, necesaria para estimar los parámetros regresores asociados al modelo, para dicha estimación es necesario utilizar una aproximación, la aproximación utilizada en este trabajo es la cuadratura de Gauss-Hermite, en este capítulo también se muestran las diferentes variaciones que existen en la cuadratura de Gauss-Hermite, seguido de la aproximación de la función de verosimilitud por medio de la cuadratura de Gauss-Hermite adaptativa (Liu y Pierce, 1994); (Pinheiro y Bates, 1995). Se ilustra la forma de utilizar el paquete **ZOIP** para el ajuste de un modelo de regresión ZOIP mixto. Luego se muestra el ajuste de un modelo de regresión ZOIP mixto a datos reales y un estudio de simulación con varios escenarios que nos permite concluir, cuál es la mejor metodología de estimación de los parámetros regresores, en un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos.

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

2 Capítulo 2: Distribución ZOIP

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

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

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

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

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

2.1. Distribución para datos proporcionales

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

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, \quad y \in (0, 1) \quad (2-1)$$

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:

$$E(Y) = \frac{p}{p+q} \quad (2-2)$$

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

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

Ferrari y Cribari-Neto (2004) propusieron otra parametrización para la distribución beta en función de los parámetros 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:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad y \in (0, 1) \quad (2-4)$$

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

$$E(Y) = \mu \quad (2-5)$$

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

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

$$p = \mu\phi \quad (2-7)$$

$$q = (1-\mu)\phi \quad (2-8)$$

2.1.3. Distribución beta con la parametrización de Rigby y Stasinopoulos (2005)

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

$$f(y; \mu, \sigma) = B(\mu, \sigma) y^{\mu((1-\sigma^2)/\sigma^2)-1} (1-y)^{(1-\mu)((1-\sigma^2)/\sigma^2)-1}, \quad y \in (0, 1) \quad (2-9)$$

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

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

$$E(Y) = \mu \quad (2-10)$$

$$Var(Y) = \sigma^2 \mu(1 - \mu) \quad (2-11)$$

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

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

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

2.1.4. Distribución simplex

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

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

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

$$E(Y) = \mu \quad (2-15)$$

$$Var(Y) = \mu(1 - \mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1 - \mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1 - \mu)^2}\right\} \quad (2-16)$$

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

2.2. Distribución ZOIP (Zeros Ones Inflated Proporcional)

En las distribuciones vistas en la sección 2.1, se evidenciaron ciertas distribuciones que se ajustan al comportamiento de datos proporcionales, porcentajes o tasas que están en el intervalo (0,1), sin embargo, es común que estos datos tomen valores en cero y/o uno que representarían la ausencia o presencia total de cierta característica, por lo que no sería posible

ajustar los datos a las distribuciones vistas anteriormente y es por eso que en este trabajo se propone la distribución ZOIP, como un conjunto de distribuciones para datos proporcionales inflados con ceros y/o unos.

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

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

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

$$E(Y) = p_1 + (1 - p_0 - p_1)E^*(Y) \quad (2-18)$$

$$Var(Y) = p_1(1 - p_1) + (1 - p_0 - p_1) [Var^*(Y) + (p_0 + p_1)[E^*(Y)]^2 - 2E^*(Y)p_1] \quad (2-19)$$

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

Si para la distribución $ZOIP(\mu, \sigma, p_0, p_1)$ se elige la distribución beta con parametrización de Ferrari y Cribari-Neto (2004) entonces el parámetro σ tomará el rol del parámetro ϕ de la distribución, si la parametrización es beta original, μ y σ tomarán el rol de p y q respectivamente. En las demás distribuciones y parametrizaciones μ y σ tomaran los valores y dominios correspondientes a su distribución.

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

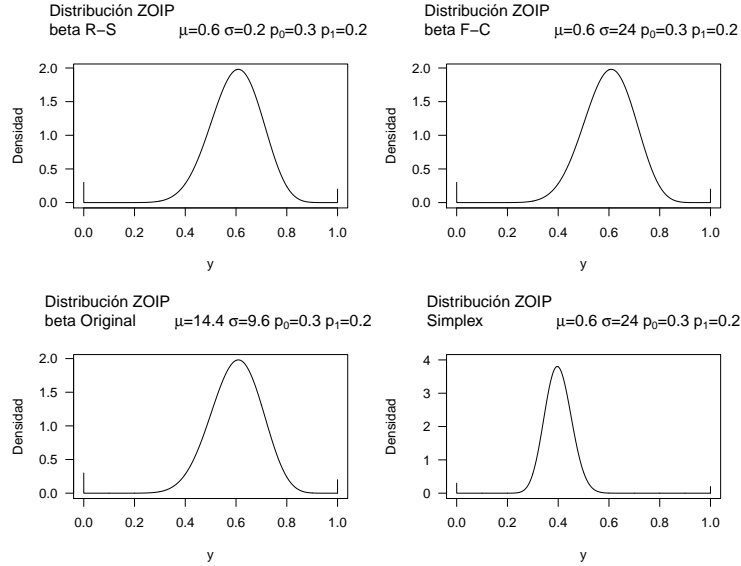


Figura 2-1: Densidades para la distribución ZOIP para algunos valores de los parámetros, donde R-S se refiere a Rigby & Stasinopoulos (2005) y F-C es Ferrari & Cribari-Neto (2004).

En la Figura 2-1 se muestran las densidades para algunas distribuciones ZOIP-Beta en sus diferentes parametrizaciones y ZOIP-simplex, es de aclarar que, en las Figuras ubicadas en la esquina superior izquierda, esquina superior derecha y esquina inferior izquierda, los valores de los parámetros son diferentes, pero dan como resultado la misma distribución gracias a las ecuaciones descritas en (2-7), (2-8) para el caso Ferrari y Cribari-Neto (2004) y (2-12), (2-13) para el caso Rigby y Stasinopoulos (2005). Además se puede observar en la Figura inferior derecha como la distribución ZOIP-simplex tiene un comportamiento distinto a las demás distribuciones, a pesar de que contiene los mismo valores en los parámetros.

2.3. Inferencia estadística

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

$$L(\theta) = \prod_{i=1}^n g(y_i; \mu, \sigma, p_0, p_1) \quad (2-20)$$

donde $g(\cdot)$ es la función de densidad de probabilidad de la distribución ZOIP, vista en la

sección 2.2.

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

1. ZOIP-beta original

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

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

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

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

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

donde:

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

Ahora, tomando logaritmo natural a la función de verosimilitud,

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

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

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

por lo que,

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

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

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

entonces

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

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

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

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

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

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

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

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

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

ver más en Owen (2008).

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

2. ZOIP-simplex

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

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

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

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

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

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

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

por lo que,

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

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

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

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

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 el CRAN y en GitHub y creado por nosotros durante este trabajo, 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 puede ingresar al repositorio CRAN, donde se encuentra disponible. Otra opción es instalar la versión en desarrollo desde 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** desde GitHub.

```
#Para instalar la version disponible en el CRAN
install.packages("ZOIP")

#Para instalar la version en desarrollo desde GitHub
if (!require("devtools")) install.packages("devtools")
devtools::install_github("jucdiaz/ZOIP", force = TRUE)
library(ZOIP) # Carga el paquete
```

2.4.2. Funciones sobre distribución ZOIP

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

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

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

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

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

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

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

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

2.4.3. Función RM.ZOIP del paquete 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(f~'{o}rmula.mu, f~'{o}rmula.sigma = ~1, f~'{o}rmula.p0 = ~1,
        f~'{o}rmula.p1 = ~1, data, link = c("identity", "identity",
        "identity", "identity"), family = "R-S", optimizer = "nlminb")
```

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

- **fórmula.mu**: fórmula que define la función de regresión para el parámetro μ , Para ajustar una distribución ZOIP debe tomar el valor de $y \sim 1$, donde y es la variable a ajustar.
- **fórmula.sigma**: fórmula que define la función de regresión para el parámetro σ , Para ajustar una distribución ZOIP debe tomar el valor de ~ 1 .
- **fórmula.p0**: fórmula que define la función de regresión para el parámetro p_0 , Para ajustar una distribución ZOIP debe tomar el valor de ~ 1 .
- **fórmula.p1**: fórmula que define la función de regresión para el parámetro p_1 , Para ajustar una distribución ZOIP debe tomar el valor de ~ 1 .
- **data**: es el conjunto de datos en formato `data.frame` donde debe contener los datos de la variable a ajustar y el nombre debe ser la tal cual como está en las fórmula para el parámetro μ .
- **family**: elección de la distribución ZOIP deseada para ajustar, si toma el valor de ‘`R-S`’ se utilizará la distribución ZOIP-beta con parametrización Rigby y Stasinopoulos (2005), si toma el valor de ‘`F-C`’ se utilizará la distribución ZOIP-beta parametrización Ferrari y Cribari-Neto (2004), si el valor es ‘`Original`’ se utilizará la distribución ZOIP-beta con parametrización original y si es ‘`Simplex`’ se 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 fórmula. Para ajustar una distribución ZOIP se debe utilizar como función enlace la opción `identity` en sus cuatro parámetros, independientemente de la distribución 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 mostrará el ajuste de una distribución ZOIP, para ello se mostrará la salida de la función `RM.ZOIP` de 1000 observaciones simuladas para la distribución ZOIP-beta según la parametrización de Rigby y Stasinopoulos (2005).

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

mod <- RM.ZOIP(f\'{o}rmula.mu = yi ~ 1, f\'{o}rmula.sigma = ~1,
  f\'{o}rmula.p0 = ~1, f\'{o}rmula.p1 = ~1, data = yi, family = "R-S")
mod

## Call:
## RM.ZOIP(f\'{o}rmula.mu = yi ~ 1, f\'{o}rmula.sigma = ~1, f\'{o}rmula.p0 = ~1,
##      f\'{o}rmula.p1 = ~1, data = yi, family = "R-S")
##
## Results:
##
## Estimated coefficients for h(mu):
## (Intercept)
##      0.605174
##
## Estimated coefficients for h(sigma):
## (Intercept)
##      0.2038938
##
## Estimated coefficients for h(p0):
## (Intercept)
##      0.03000002
##
## Estimated coefficients for h(p1):
## (Intercept)
##      0.05
##
## Convergence
## [1] 0
```

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

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

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

```
summary(mod)
## -----
## Fixed effects for identity(mu)
## -----
##           Estimate Std. Error z value Pr(>|z|)
## (intercept) 0.6066914  0.0031636  191.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Fixed effects for identity(sigma)
## -----
##           Estimate Std. Error z value Pr(>|z|)
## (intercept) 0.196643  0.004322  45.498 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```



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

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

En el resultado anterior se obtienen los valores de $\hat{\mu} = 0.6066914$, $\hat{\sigma} = 0.196643$, $\hat{p}_0 = 0.0339992$ y $\hat{p}_1 = 0.0450005$, que son los parámetros con los que se simuló 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 búsqueda de los parámetros, ya que estos son introducidas automáticamente de acuerdo al valor tomado en `family`.

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

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 parametrizaciones con tamaños de muestra n de: 5, 10, 15 y así sucesivamente hasta 500, y se realizaron 1000 réplicas para cada tamaño de muestra, posteriormente se calculó la mediana de cada una de las estimaciones de los parámetros, y así poder analizar la capacidad de convergencia de las metodologías implementadas en la distribución ZOIP y en el paquete **ZOIP**.

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

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

Como medida global del proceso de estimación se eligió el MAPE (Error porcentual absoluto medio $\sum_{i=1}^n |y_i - \hat{y}_i|/y_i$)/ n debido a los cambios de escala entre los diferentes parámetros de las diferentes distribuciones y parametrizaciones. Esta media se realizó como un promedio de los MAPES generados por cada uno de los parámetros de la distribución ZOIP en cada tamaño de muestra. En la Figura **2-4** de la izquierda 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

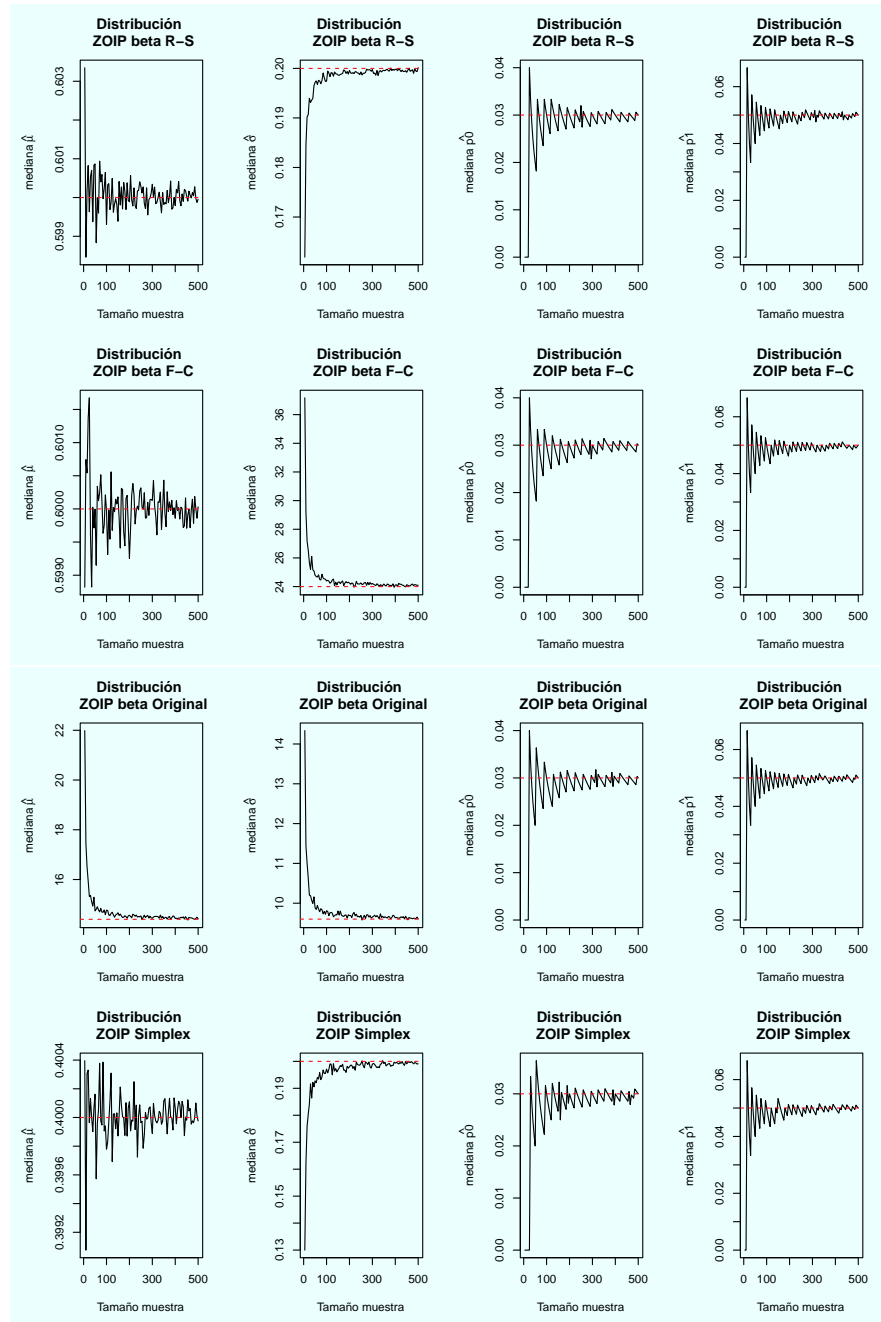


Figura 2-2: Mediana de los parámetros estimados en el escenario 1 para distintas parametrizaciones y valores de n , las líneas punteadas representan el verdadero valor del parámetro.

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 alejadas del valor real.

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

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

En la Figura **2-4** de la derecha 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 menor error sobre la estimación de los parámetros de la distribución beta con la parametrización de Rigby y Stasinopoulos (2005). En la Tabla **2-1** se presenta el MAPE para cada parámetro de cada parametrización para ambos estudios de simulación, se observa como; en general, el estudio de simulación

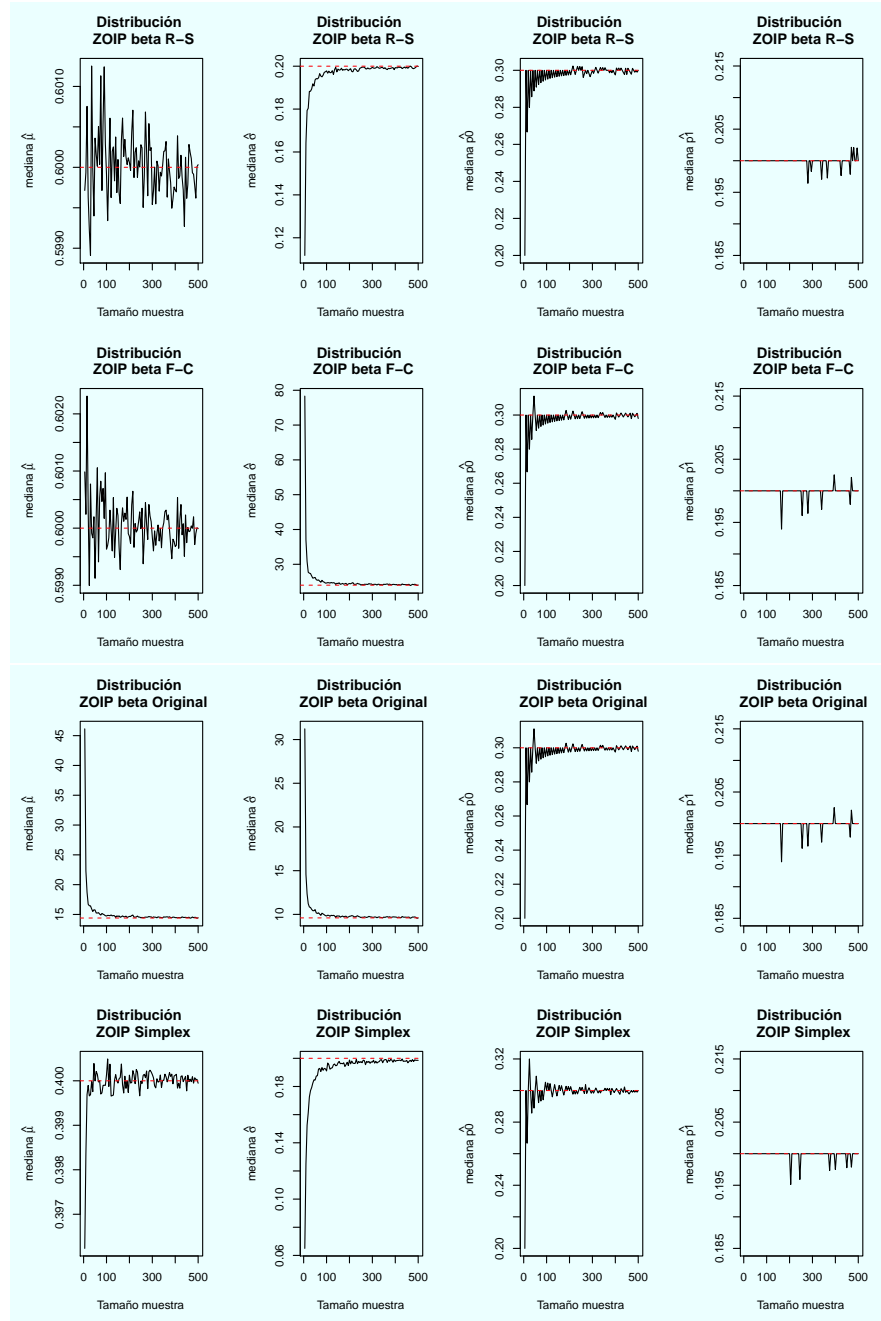


Figura 2-3: Simulación de distribución ZOIP para distintas parametrizaciones con parámetros de inflación grandes, distribuciones y valores de n .

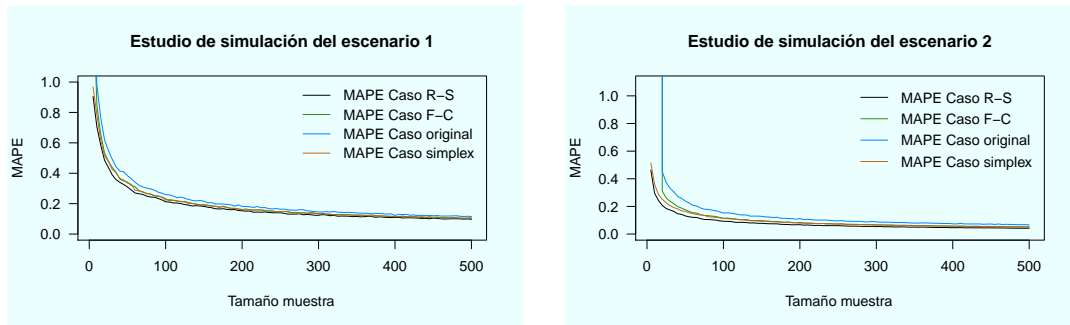


Figura 2-4: MAPE (Error porcentual absoluto medio) para los dos escenarios de simulación y para distintas parametrizaciones y valores de n .

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 promueve 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 de Colombia, que por cuestiones de confidencialidad no se revela el nombre del banco. Para una entidad financiera 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 en cuestión. Se quiere estudiar el ajuste de una distribución ZOIP y para ello se utiliza el paquete de R de nuestra creación 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 se ve cómo cambian los valores de los parámetros según la parametrización escogida, los valores de log-verosimilitud indican 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

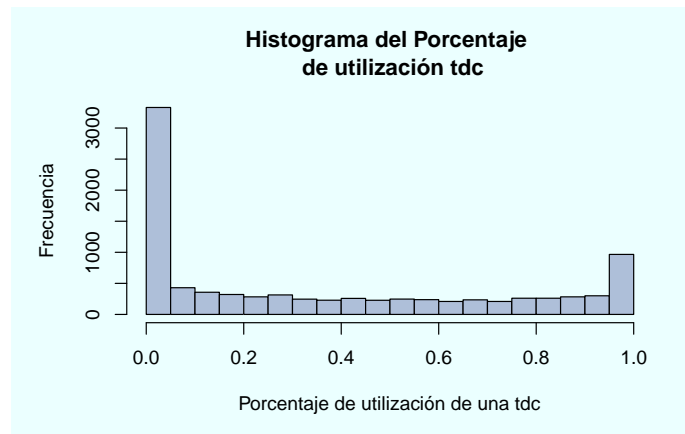


Figura 2-5: Histograma del porcentaje de utilización de las tdc en un banco.

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

Tabla 2-2: Ajuste de diferentes distribuciones ZOIP en el porcentaje de utilización de una tdc, en un banco colombiano. Fuente: Cortesía del banco.

parámetros de la distribución ZOIP-simplex no se tuvo una convergencia, por lo tanto los valores son muy distintos para el parámetro de dispersión a los vistos en la distribución ZOIP-beta, inclusive muy elevados. Además, el valor de μ es mayor que las de la parametrización en Rigby y Stasinopoulos (2005) y Ferrari y Cribari-Neto (2004), de hecho un 17 % más. Los modelos propuestos toman un tiempo promedio de 5 segundos para encontrar su convergencia, mediante el paquete **ZOIP**.

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 punteada que representa la distribución ZOIP ajustada. Es de notar que dicha línea punteada es exactamente igual en las tres ocasiones que se ajusta la distribución ZOIP-beta y se ve como sigue el com-

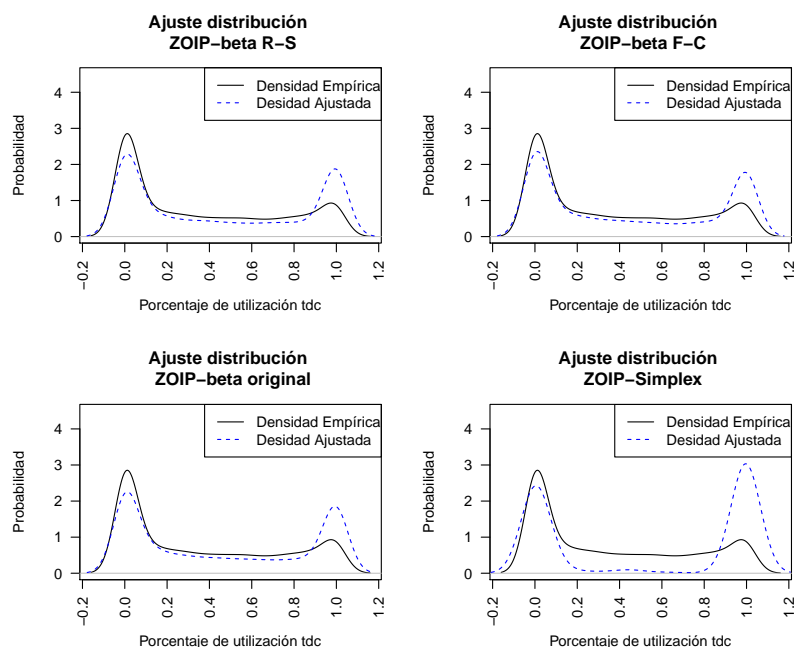


Figura 2-6: Ajuste de diferentes distribuciones y parametrizaciones ZOIP al porcentaje de utilización de una tdc.

portamiento original del porcentaje de utilización de las tdc. Es bueno resaltar también que la distribución ZOIP-simplex no tiene un buen ajuste para los valores entre cero y uno y en uno, específicamente. Por todo lo 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 realizado por nosotros, 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 colombiano.

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

3 Capítulo 3: Modelo de regresión ZOIP con efectos fijos

En muchos casos de estudio es razonable preguntarse cómo puede ser explicada una variable aleatoria proveniente de datos proporcionales a partir de diferentes covariables, como por ejemplo el porcentaje de uso de una tarjeta de crédito a partir de la cantidad de compras hechas en el mes inmediatamente anterior. el modelo más conocido en la literatura para abordar este problema es la regresión beta, Paolino (2001) estima mediante el método de máxima verosimilitud, los parámetros para una regresión beta con parametrización original; Ferrari y Cribari-Neto (2004) reparametrizan la distribución y proponen el modelo regresión beta bajo esta nueva parametrización, luego Zeileis et al. (2010) crean el paquete **betareg** de R en el cual se implementa dicha regresión. Por otro parte, Rigby y Stasinopoulos (2005) también realizaron otra reparametrización de la distribución beta original, en términos de 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. Existe otro modelo de regresión basado en la distribución simplex, propuesta por Jørgensen y Barndorff-Nielsen (1991), este nuevo modelo de regresión fue realizado por Qiu et al. (2008) e implementado en el paquete **simplexreg**, Zhang et al. (2016).

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

Es por esto que en este trabajo se implementa de manera teórica y de forma práctica mediante el paquete **ZOIP** en el sistema de computación R (R Core Team, 2017) y disponible en el CRAN de R y 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, sin tener un límite de covariables y manejando cualquier tipo de variables explicativas, numéricas y categóricas.

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

3.1. Modelo de regresión ZOIP

Una clase general de modelos de regresión ZOIP puede definirse como sigue. Sea y_1, y_2, \dots, y_n variables aleatorias independientes tal que cada y_i , para $i = 1, \dots, n$, tiene función de densidad de probabilidad definida en (2-17) y dada por:

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

cada y_i se distribuye ZOIP con parámetros $\mu = \mu_i$, $\sigma = \sigma_i$, $p_0 = p_{0i}$, y $p_1 = p_{1i}$. 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 $\mathbf{x}_{i1} = (x_{i11}, x_{i12}, \dots, x_{i1k_1})^\top$, $\mathbf{x}_{i2} = (x_{i21}, x_{i22}, \dots, x_{i2k_2})^\top$, $\mathbf{x}_{i3} = (x_{i31}, x_{i32}, \dots, x_{i3k_3})^\top$ y $\mathbf{x}_{i4} = (x_{i41}, x_{i42}, \dots, x_{i4k_4})^\top$, son vectores de covariables conocidos de dimensión k_1 , k_2 , k_3 y k_4 respectivamente, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1k_1})^\top$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2k_2})^\top$, $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \dots, \beta_{3k_3})^\top$ y $\boldsymbol{\beta}_4 = (\beta_{41}, \beta_{42}, \dots, \beta_{4k_4})^\top$ son vectores de parámetros de regresión desconocidos. Además, se asume que las funciones de enlace $h_1(\cdot)$, $h_2(\cdot)$, $h_3(\cdot)$ y $h_4(\cdot)$ son conocidas y apropiadas para mapear de los reales a los valores admisibles del parámetro, además son funciones estrictamente monótonas y doblemente

diferenciables, para conservar así el efecto de las variables sobre la variable respuesta original. Las posibles funciones de enlace 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 logit, probit y clog-log. Estudios sobre funciones enlace mal especificadas sobre modelos de regresión beta se encuentran en de Andrade (2007).

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:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n g(\mathbf{y}_i; \mu_i, \sigma_i, p_{0i}, p_{1i}) = \prod_{i=1}^n f_0(\mathbf{y}_i; p_{0i}) \prod_{i=1}^n f_1(\mathbf{y}_i; p_{1i}) \prod_{i=1}^n f_2(\mathbf{y}_i; \mu_i, \sigma_i) \quad (3-2)$$

donde las funciones $f_0(\mathbf{y}_i; p_{0i})$ y $f_1(\mathbf{y}_i; p_{1i})$ son explicadas por:

$$f_j(\mathbf{y}_i; p_{ji}) = p_{ji}^{S_j(y_i)} (1 - p_{ji})^{1-S_j(y_i)} \quad ; \quad j = 0, 1$$

Con

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

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

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

$$\ell(\boldsymbol{\theta}) = \ell_0(\boldsymbol{\beta}_3) + \ell_1(\boldsymbol{\beta}_4) + \ell_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

Note que la función de verosimilitud es factorizada en tres términos, dos de ellos del componente discreto y uno compuesto por $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ del componente continuo, por tanto los parámetros son separables (Pace y Salvani, 1997), así la máxima verosimilitud puede ser tratada por separado y por lo tanto:

$$\begin{aligned}\ell_0(\beta_3) &= \sum_{i=1}^n p_{0i}^{S_0(y_i)} (1 - p_{0i})^{1-S_0(y_i)} \\ \ell_1(\beta_4) &= \sum_{i=1}^n p_{1i}^{S_1(y_i)} (1 - p_{1i})^{1-S_1(y_i)} \\ \ell_2(\beta_1, \beta_2) &= \sum_{i=1: y_i \in (0,1)}^n f_3(y_i; \mu_i, \sigma_i)\end{aligned}$$

Con $p_{0i} = h_3^{-1}(\mathbf{x}_{i3}^\top \beta_3)$, $p_{1i} = h_4^{-1}(\mathbf{x}_{i4}^\top \beta_4)$, $\mu_i = h_1^{-1}(\mathbf{x}_{i1}^\top \beta_1)$ y $\sigma_i = h_2^{-1}(\mathbf{x}_{i2}^\top \beta_2)$ como se definió en (3-1). La función de verosimilitud depende de tres términos, el primero depende de β_3 (componente discreto para inflación en cero), el segundo de β_4 (componente discreto para explicar la inflación en uno) y el tercero depende de (β_1, β_2) (Componentes para explicar la parte continua entre cero y uno), por lo tanto los parámetros son separables y la inferencia de máxima verosimilitud para β_1 y β_2 se puede hacer por separado de la de β_3 y β_4 , como si conociera a β_3 y β_4 y viceversa. (Ospina y Ferrari, 2012).

No existen expresiones que den una solución cerrada analíticamente para encontrar los máximos de las funciones de las 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 se usará el algoritmo de optimización dado por la función `nlminb` o `optim` del paquete **stats** de R e implementado en el paquete **ZOIP** de R para el modelo de regresión ZOIP.

3.2. Modelo de regresión ZOIP en el paquete ZOIP

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

Función **RM.ZOIP** del paquete **ZOIP**

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

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

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

- **formula.mu**: fórmula que define la variable respuesta y la estructura de covariables para modelar el parámetro μ , por ejemplo si se escribe $y \sim x_1 + x_2$ significa que la variable respuesta es y y que $h(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- **formula.sigma**: fórmula que define la función de regresión de efectos fijos, para el parámetro σ , un valor posible es $\sim x_1$, por defecto ~ 1 .
- **formula.p0**: fórmula que define la función de regresión de efectos fijos, para el parámetro p_0 , un valor posible es $\sim x_1$, por defecto ~ 1 .
- **formula.p1**: fórmula que define la función de regresión de efectos fijos, para el parámetro p_1 , un valor posible es $\sim x_1$, por defecto ~ 1 .
- **data**: es el conjunto de datos en formato `data.frame` donde las variables deben tener el mismo nombre como se especificaron en el modelo.
- **family**: elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP. El valor de `‘‘R-S’’` indicará la distribución beta con la parametrización de Rigby y Stasinopoulos (2005), si toma el valor de `‘‘F-C’’` se utilizará la distribución beta con la parametrización de Ferrari y Cribari-Neto (2004), si el valor es `‘‘Original’’` se utilizará la distribución beta con la parametrización original y si es `‘‘Simplex’’` se utilizará la distribución simplex.
- **link**: es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y fórmula. Si la función de regresión no posee covariables para la explicación de los parámetros de μ , σ , p_0 o p_1 ; entonces se debe utilizar como función enlace la opción `identity`, independientemente de la parametrización deseada (family). Los posibles valores para las funciones enlace son `identity`, `logit` y `log`.
Por defecto `link=c(‘‘identity’’, ‘‘identity’’, ‘‘identity’’, ‘‘identity’’)`.
- **optimizer**: elección del optimizador, utilizado para encontrar la convergencia de la máxima verosimilitud en los parámetros de efectos fijos, se puede elegir el valor de `‘‘nlminb’’` u `‘‘optim’’`, por defecto `‘‘nlminb’’`.

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

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

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

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

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

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

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

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

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

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

mod
```

Los resultados obtenidos se muestran a continuación.

```
## Call:
## RM.ZOIP(formula.mu = y_i ~ x1, formula.sigma = ~x1 + x2, formula.p0 = ~1,
##       formula.p1 = ~x2, data = data, link = link, family = "R-S")
##
## Results:
##
## Estimated coefficients for h(mu):
## (Intercept)          x1
##  0.3395118  -1.4808985
##
## Estimated coefficients for h(sigma):
## (Intercept)          x1          x2
##  0.7006753   2.4493508   0.5328350
##
## Estimated coefficients for h(p0):
## (Intercept)
##  0.08199999
##
## Estimated coefficients for h(p1):
## (Intercept)          x2
## -0.1252807  -3.0092159
##
## Convergence
## [1] 0
##
## message
## [1] "relative convergence (4)"
##
## iterations
## [1] 39
##
## Log-likelihood
## [1] 7688.628
```

En el anterior resultado se obtienen varios aspectos importantes de la salida del modelo y leyendo de arriba hacia abajo, primero que todo nos muestra el modelo ajustado, luego para cada parámetro de la distribución ZOIP los valores estimados para cada uno de los parámetros de regresión asociados a cada covariable, luego un indicador de convergencia, donde 0 indica la convergencia, después un mensaje sobre la convergencia (resultados heredados

de la función `nlimnb`), después se encuentra el número de iteraciones que fueron necesarias para que convergiera el modelo, por último se encuentra valor de la log-verosimilitud que nos permitirá hacer comparaciones entre modelos.

En caso en que el modelo elegido no haya convergido, puede ser debido a que en ocasiones el conjunto de datos elegidos no es suficientemente grande y/o explícito sobre el modelo elegido, por lo que la función `RM.ZOIP` no es capaz encontrar el máximo de la función de log-verosimilitud. En el estudio de simulación propuesto en ocasiones ocurre que el resultado no converge por lo que era necesario realizar una nueva repetición del escenario de simulación propuesto, sin embargo, esto no afectó nunca los números de réplicas propuestos, (McCulloch y Neuhaus, 2011a). Por otra parte, hay ocasiones donde es necesario realizar exploraciones previas sobre la variable respuesta, debido a que esta puede no ser explicada a partir del modelo elegido, por lo tanto es conveniente explorar otros modelos de regresión para datos proporcionales inflados con ceros y/o unos, como modelos basados en la distribución beta-rectangular o LogitSep.

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

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

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.0820000  0.0086762  9.4512 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Fixed effects for logit(p1)
## -----
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.12528    0.15084 -0.8305  0.4062
## x2          -3.00922    0.34242 -8.7881  <2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## -----
```

Con la función `summary` aplicada al modelo de regresión ZOIP, se obtiene más detalles de los parámetros regresores estimados para cada parámetro del modelo de regresión ZOIP, se obtiene el valor estimado (Estimate), su error estándar (Std.Error), el valor Z del estimador (z value) y el valor p que indicará la significancia estadística del parámetro estimado ($\text{Pr}(>|z|)$), es de notar que para cada parámetro del modelo de regresión ZOIP le es mostrado la respectiva función enlace utilizada y definida al inicio del modelo.

3.3. Aplicación

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

de muestra que se pueden obtener en un experimento, 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.

$$\begin{aligned}
 y_i &\sim ZOIP(\mu_i, \sigma_i, p_{0i}, p_{1i}), \\
 h_1(\mu_i) &= \beta_{10} + \beta_{11}x_{1i}, \\
 h_2(\sigma_i) &= \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i}, \\
 h_3(p_{0i}) &= \beta_{30}, \\
 h_4(p_{1i}) &= \beta_{40} + \beta_{41}x_{2i},
 \end{aligned} \tag{3-3}$$

donde, para la regresión del parámetro de la media (μ): $\beta_{10} = 0.2$ y $\beta_{11} = -1$, para la regresión del parámetro σ se escogieron dos escenarios distintos, el primero si el modelo es un ZOIP-beta bajo la parametrización de Rigby y Stasinopoulos (2005), entonces: $\beta_{20} = 0.3$, $\beta_{21} = -2$, $\beta_{22} = -4$, y el segundo escenario si se tiene un modelo ZOIP-simplex o ZOIP-beta con parametrización de Ferrari y Cribari-Neto (2004) o original el escenario es $\beta_{20} = 0.3$, $\beta_{21} = 3$, $\beta_{22} = 0.9$, esto para tener una variabilidad de los datos diferente entre los modelos, ya que bajo la parametrización de Ferrari y Cribari-Neto (2004) o original y los valores del escenario uno, se obtendrán valores de dispersión grande en los datos, por lo que será mucho más difícil encontrar la convergencia de los parámetros, dicho caso de estudio no es considerado en este trabajo. Para el parámetro p_0 : $\beta_{30} = 0.07$ y para p_1 : $\beta_{40} = 0.02$ y $\beta_{41} = -4$ para todos los casos posibles de selección de la regresión ZOIP, y $x_{1i} \sim U(0, 1)$, $x_{2i} \sim U(0, 1)$. Las funciones de enlace adecuadas para cada distribución y parametrización se muestran en la Tabla 3-1.

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

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

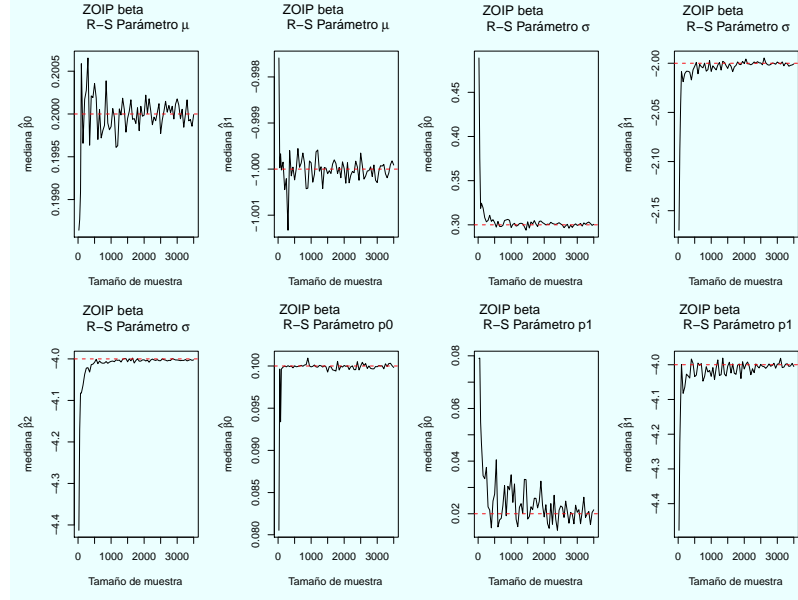


Figura 3-1: Simulación de un modelo de regresión ZOIP-beta para la parametrización R-S con diferentes valores de n .

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

En la Figura **3-2** se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización de Ferrari y Cribari-Neto (2004). En dicha figura se nota como la estimación de los parámetros asociados con la media tienen una oscilación mayor que los demás parámetros, sin embargo, en todos los parámetros se observa como a medida que el tamaño de muestra es más grande, a partir de $n = 1000$, 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

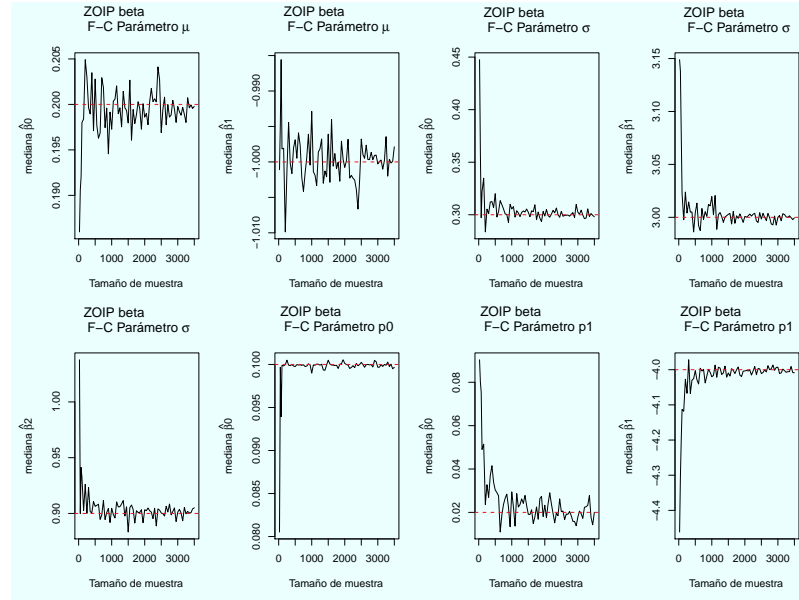


Figura 3-2: Simulación de un modelo de regresión ZOIP-beta para la parametrización F-C con diferentes valores de n .

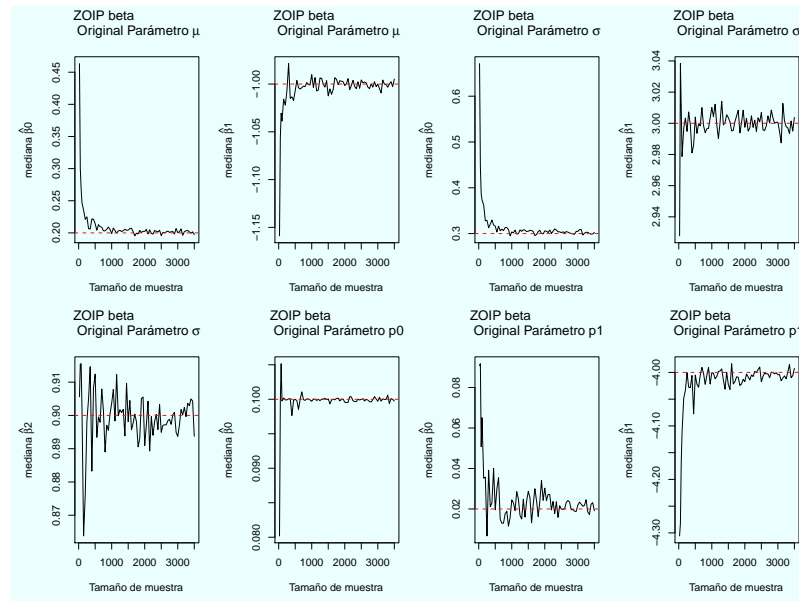


Figura 3-3: Simulación de un modelo de regresión ZOIP-beta para la parametrización original con diferentes valores de n .

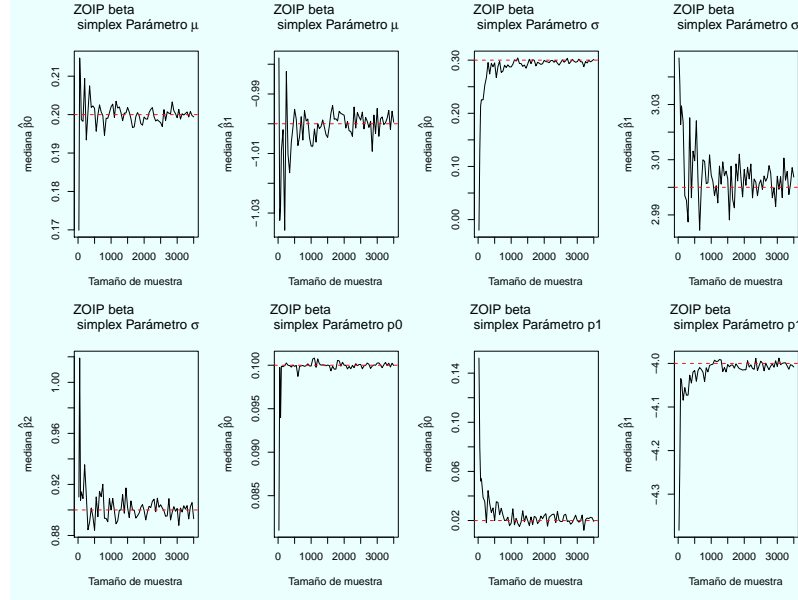


Figura 3-4: Simulación de un modelo de regresión ZOIP-simplex con diferentes valores de n .

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 medida que el tamaño de muestra es mayor la oscilación va disminuyendo, por lo que se sospecha que se necesita un tamaño de muestra mayor a 3000, 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 medida que el tamaño de muestra crece, sin embargo, unos parámetros tomarán un número superior a los 3500 de tamaño de muestra, para obtener convergencia más satisfactoria, esto ocurre en el parámetro β_1 asociado al parámetro de dispersión (σ).

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 obtiene un MAPE más grande en las regresiones asociadas a todos los parámetros de inflación, esto permite concluir que hallar los parámetros verdaderos en los parámetros de inflación es un poco más difícil que en los parámetros de localización

| 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 Parámetros en las diferentes parametrizaciones.

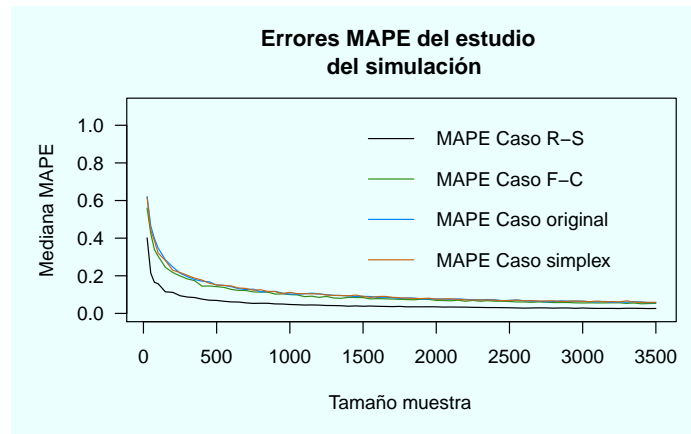


Figura 3-5: Mape (Error porcentual absoluto medio) para modelo de regresion ZOIP simulado para distintas parametrizaciones y valores de n .

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 permite concluir que a pesar de que los diferentes parámetros estimados en la simulación oscilan alrededor del valor real este todavía tiene una variabilidad muy grande por lo que hace que este MAPE sea grande y el parámetro no haya convergido con un tamaño de muestra de 3500.

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

3.3.2. Datos reales

Utilizando de manera similar información sobre la entidad financiera en Colombia descrita en la sección 2.5.2. 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 calificación 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.

$$\begin{aligned}
 y_i &\sim ZOIP(\mu_i, \sigma_i, p_{0i}, p_{1i}), \quad i = 1, 2, \dots, 9206 \\
 h_1(\mu_i) &= \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + \beta_{13}x_{3i}, \\
 h_2(\sigma_i) &= \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + \beta_{23}x_{3i}, \\
 h_3(p_{0i}) &= \beta_{30} + \beta_{31}x_{1i} + \beta_{32}x_{2i} + \beta_{33}x_{3i}, \\
 h_4(p_{1i}) &= \beta_{40} + \beta_{41}x_{1i} + \beta_{42}x_{2i} + \beta_{43}x_{3i},
 \end{aligned} \tag{3-4}$$

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

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 se puede observar cómo se explica el porcentaje de utilización de una tdc, la variabilidad de este porcentaje de utilización, el por que un cliente no llega a utilizar nunca su tdc y adicional y contrariamente a lo anterior, por que un cliente utiliza la total capacidad de su tdc.

| Familia | Parámetro | $h(\cdot)$ |
|----------|-----------|------------|
| 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.

| Familia | Parámetro | β 's | Estimación | Error estándar | Valor P | Log-Verosimilitud | Iteraciones |
|---------|-----------|--------------------|------------|----------------|----------------|-------------------|-------------|
| R-S | μ | $\hat{\beta}_{10}$ | -0.046 | 0.050 | 0.3618 | -5414.738 | 125 |
| | | $\hat{\beta}_{11}$ | -0.354 | 0.107 | 0.0009 | | |
| | | $\hat{\beta}_{12}$ | 0.022 | 0.002 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{13}$ | -0.025 | 0.009 | 0.0074 | | |
| | σ | $\hat{\beta}_{20}$ | 0.822 | 0.038 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{21}$ | -0.197 | 0.078 | 0.0114 | | |
| | | $\hat{\beta}_{22}$ | -0.006 | 0.002 | 0.0013 | | |
| | | $\hat{\beta}_{23}$ | -0.003 | 0.007 | 0.6741 | | |
| | p_0 | $\hat{\beta}_{30}$ | -1.496 | 0.101 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{31}$ | 0.724 | 0.185 | $8.87e^{-5}$ | | |
| | | $\hat{\beta}_{32}$ | -0.153 | 0.009 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{33}$ | 0.002 | 0.015 | 0.1243 | | |
| | p_1 | $\hat{\beta}_{40}$ | -1.480 | 0.095 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{41}$ | -0.630 | 0.254 | 0.0132 | | |
| | | $\hat{\beta}_{42}$ | 0.011 | 0.006 | 0.0666 | | |
| | | $\hat{\beta}_{43}$ | -0.069 | 0.022 | 0.0022 | | |

Tabla 3-4: Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Rigby y Stasinopoulos (2005) en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$

En las Tablas **3-4**, **3-5**, **3-6**, **3-7** se tiene información sobre cómo 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 los clientes con historial crediticio malo utilizan más su tdc, lo que es lógico porque para llegar a estar en mora es necesario utilizar el producto, sobre el parámetro del promedio de cuotas diferidas vemos como el parámetro

| Familia | Parámetro | β 's | Estimación | Error estándar | Valor P | Log-Verosimilitud | Iteraciones |
|---------|-----------|--------------------|------------|----------------|----------------|-------------------|-------------|
| F-C | μ | $\hat{\beta}_{10}$ | -0.045 | 0.050 | 0.3667 | -5414.605 | 105 |
| | | $\hat{\beta}_{11}$ | -0.354 | 0.107 | 0.0009 | | |
| | | $\hat{\beta}_{12}$ | 0.022 | 0.002 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{13}$ | -0.025 | 0.009 | 0.0074 | | |
| | σ | $\hat{\beta}_{20}$ | 0.068 | 0.045 | 0.1286 | | |
| | | $\hat{\beta}_{21}$ | 0.238 | 0.094 | 0.0117 | | |
| | | $\hat{\beta}_{22}$ | 0.007 | 0.002 | 0.0012 | | |
| | | $\hat{\beta}_{23}$ | 0.003 | 0.008 | 0.6854 | | |
| | p_0 | $\hat{\beta}_{30}$ | -1.496 | 0.101 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{31}$ | 0.724 | 0.185 | $8.87e^{-5}$ | | |
| | | $\hat{\beta}_{32}$ | -0.153 | 0.009 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{33}$ | 0.002 | 0.015 | 0.1243 | | |
| | p_1 | $\hat{\beta}_{40}$ | -1.480 | 0.095 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{41}$ | -0.630 | 0.254 | 0.0132 | | |
| | | $\hat{\beta}_{42}$ | 0.011 | 0.006 | 0.0666 | | |
| | | $\hat{\beta}_{43}$ | -0.069 | 0.022 | 0.0022 | | |

Tabla 3-5: Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización Ferrari y Cribari-Neto (2004) en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$

| Familia | Parámetro | β 's | Estimación | Error estándar | Valor P | Log-Verosimilitud | Iteraciones |
|----------|-----------|--------------------|------------|----------------|----------------|-------------------|-------------|
| original | μ | $\hat{\beta}_{10}$ | -0.649 | 0.048 | $< 2.2e^{-16}$ | -5415.386 | 121 |
| | | $\hat{\beta}_{11}$ | 0.035 | 0.103 | 0.7311 | | |
| | | $\hat{\beta}_{12}$ | 0.019 | 0.002 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{13}$ | -0.011 | 0.009 | 0.2267 | | |
| | σ | $\hat{\beta}_{20}$ | -0.611 | 0.054 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{21}$ | 0.397 | 0.111 | 0.0003 | | |
| | | $\hat{\beta}_{22}$ | -0.002 | 0.003 | 0.4724 | | |
| | | $\hat{\beta}_{23}$ | 0.015 | 0.010 | 0.1456 | | |
| | p_0 | $\hat{\beta}_{30}$ | -1.496 | 0.101 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{31}$ | 0.724 | 0.185 | $8.87e^{-5}$ | | |
| | | $\hat{\beta}_{32}$ | -0.153 | 0.009 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{33}$ | 0.002 | 0.015 | 0.1243 | | |
| | p_1 | $\hat{\beta}_{40}$ | -1.480 | 0.095 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{41}$ | -0.630 | 0.254 | 0.0132 | | |
| | | $\hat{\beta}_{42}$ | 0.011 | 0.006 | 0.0666 | | |
| | | $\hat{\beta}_{43}$ | -0.069 | 0.022 | 0.0022 | | |

Tabla 3-6: Parámetros regresores estimados de un modelo de regresión ZOIP-beta con parametrización original en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $1e^{-10}$

β_{12} 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_{13} asociado al cupo de su tdc tiene un efecto negativo aunque no muy significativo sobre la variable respuesta, lo cual indicaría que a mayor cupo en su tdc un poco menos de utilización de la tdc estará acompañado.

| Familia | Parámetro | β 's | Estimación | Error estándar | Valor P | Log-Verosimilitud | Iteraciones |
|---------|-----------|--------------------|------------|----------------|----------------|-------------------|-------------|
| simplex | μ | $\hat{\beta}_{10}$ | 0.180 | 0.050 | 0.0003 | -22385.78 | 117 |
| | | $\hat{\beta}_{11}$ | -3.890 | 0.206 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{12}$ | 0.166 | 0.004 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{13}$ | -0.128 | 0.019 | $5.85e^{-12}$ | | |
| | σ | $\hat{\beta}_{20}$ | 11.59 | 0.062 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{21}$ | 0.063 | 0.240 | 0.7918 | | |
| | | $\hat{\beta}_{22}$ | 0.134 | 0.004 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{23}$ | 0.344 | 0.023 | $< 2.2e^{-16}$ | | |
| | p_0 | $\hat{\beta}_{30}$ | -1.496 | 0.101 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{31}$ | 0.724 | 0.185 | $8.87e^{-5}$ | | |
| | | $\hat{\beta}_{32}$ | -0.153 | 0.009 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{33}$ | 0.002 | 0.015 | 0.1243 | | |
| | p_1 | $\hat{\beta}_{40}$ | -1.480 | 0.095 | $< 2.2e^{-16}$ | | |
| | | $\hat{\beta}_{41}$ | -0.630 | 0.254 | 0.0132 | | |
| | | $\hat{\beta}_{42}$ | 0.011 | 0.006 | 0.0666 | | |
| | | $\hat{\beta}_{43}$ | -0.069 | 0.022 | 0.0022 | | |

Tabla 3-7: Parámetros regresores estimados de un modelo de regresión ZOIP-simplex en el porcentaje de utilización de una tdc. Con un criterio de convergencia de $2e^{-10}$

Al analizar el efecto de la variabilidad de la utilización de las tdc, se observa como el parámetro de score tiene un efecto positivo sobre la precisión de la utilización de la tdc, en el modelo ZOIP-beta parametrización Rigby y Stasinopoulos (2005) este efecto parece ser negativo, pero σ , al estar representando la dispersión y no la precisión, estaría dando un efecto positivo sobre la precisión, por lo tanto a mejor comportamiento de pago la utilización de la tdc serán más parecidas entre los individuos, cabe resaltar que β_{21} no es significativo en el modelo simplex. Además el parámetro β_{23} asociado al cupo de la tdc no influye sobre la variabilidad en ninguno de los modelos propuestos, excepto en el 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 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 una tdc, lo que indica que a mayor cupo menor probabilidad de utilizar por completo una tdc (efecto que se había evidenciado

en la explicación de la utilización media de las tdc), el mismo efecto se observa sobre la variable score que indica que a peor comportamiento de pago más utilización la tdc, sin embargo, si el cliente difiere a grandes cuotas su compras este tenderá a tener una mayor probabilidad de utilizar por completo su tdc. Es importante 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 con parametrización de Ferrari y Cribari-Neto (2004) seguido por la parametrización original y por último la parametrización de Rigby y Stasinopoulos (2005).

3.4. Conclusión

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

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

4 Capítulo 4: Modelo de regresión ZOIP con efectos mixtos

El modelo de regresión clásico se ha caracterizado por medir el efecto de una variable, llamada covariable sobre otra, llamada variable respuesta. Cuando los valores que puede tener la covariable son informativos, conocidos con anterioridad e independientes entre sí, se denomina efecto fijo. Pero cuando la recolección de los datos dependen de una variable categórica o el estudio se puede ver afectado por una variable categórica y dicha variable es netamente identificativa, donde podrían encontrarse otros valores diferentes si el estudio se repitiera en otras circunstancia y además esta variable posee muchos niveles y las observaciones pertenecientes a un mismo nivel están correlacionadas entre sí, entonces hablamos de una variable con efecto aleatorio. Este tipo de variables se evidencian fácilmente en el estudio de datos longitudinales, en el que un mismo individuo obtiene diferentes observaciones a través del tiempo. La utilidad de considerar las variables de este tipo como un efecto aleatorio y no como un efecto fijo, radica primero en contemplar de manera correcta el problema debido a que hay correlaciones entre las observaciones, segundo en la repetibilidad y reproductibilidad del modelo y tercero en la manera de estimar el efecto de la variable, ya que si se estimara una variable de este tipo como un efecto fijo, se estimaría un parámetro por cada nivel de la variable, en cambio cuando la variable es considerada como un efecto aleatorio, esta es caracterizada por una distribución, por lo general normal, y basta con estimar el componente de variación de la distribución. Ver más en Seoane (2014).

Los modelos de regresión mixto introducidos por Laird y Ware (1982), son aquellos modelos que combinan variables de efecto fijo y efecto aleatorio. El efecto aleatorio puede considerarse de diferentes maneras en el modelo, una de ellas es solo en el intercepto, otra sería solo en la pendiente y la última sería en ambas partes (efecto aleatorio en intercepto y pendiente), estos modelos son útiles porque son capaces de caracterizarse por efectos fijos y considerar de manera correcta el efecto de la variable agrupadora por medio del efecto aleatorio, esto garantiza que el modelo pueda ser repetido y reproducido. Además, los efectos aleatorios permiten ver el efecto de los diferentes niveles de una variable categórica o agrupadora, estimando solo un parámetro.

Los modelos de regresión mixtos también han sido implementados cuando la variable respuesta es una variable aleatoria perteneciente a datos proporcionales, tal es el caso, que

Rigby y Stasinopoulos (2005) en los modelos aditivos generalizados para localización, escala y forma (Gamlss) implementan el modelo de regresión beta con intercepto aleatorio normal, así otros autores como Verkuilen y Smithson (2012) y Bonat et al. (2012) proponen modelos de regresión beta con efectos aleatorios normales, estimados a partir de máxima verosimilitud marginal y metodologías bayesianas. Figueroa-Zuñiga et al. (2013) extienden el modelo propuesto por Ferrari y Cribari-Neto (2004) a un modelo con efectos fijos y aleatorios bajo la distribución normal y bajo la distribución t en estructuras de regresión tanto para el parámetro de la media, como del parámetro de precisión, la estimación de los parámetros del modelo de regresión fue realizado bajo una perspectiva bayesiana, mediante implementaciones computacionales del muestreador de Gibbs. Usuga y Giampaoli (2013) desarrollan el modelo de regresión beta mixto para datos proporcionales longitudinales, bajo intercepto y pendiente aleatoria normal y no normal, la estimación de los parámetros es realizada vía máxima verosimilitud y la cuadratura de Gauss-Hermite adaptativa. Otros autores como Song et al. (2004) implementan un modelo de regresión mixto para una variable respuesta bajo la distribución simplex, Bonat et al. (2015) también realiza un análisis de verosimilitud del modelo beta mixto, donde la estimación de los parámetros de regresión es realizada bajo algoritmos MCMC.

Como se pudo notar anteriormente, la estimación de los parámetros de los modelos son metodologías de aproximación numérica, esto debido a que la estimación de los parámetros del modelo de regresión mixto no tienen una solución cerrada analíticamente y es aún más complicado cuando se trata de una variable respuesta perteneciente a datos proporcionales, por lo que se utilizan ciertas aproximaciones o en la mayoría de los casos metodologías y algoritmos bayesianos para la estimación de los parámetros regresores. Una de las metodologías utilizadas es la aproximación de la función de verosimilitud vía la cuadratura de Gauss-Hermite adaptativa, utilizada e implementada en los modelos de regresión beta mixtos por Usuga y Giampaoli (2013) y es utilizada para la estimación del modelo de regresión ZOIP mixto. Dicha cuadratura fue implementada anteriormente por Fahrmeir y Tutz (2001) sobre los modelos lineales generalizados. Diversos estudios para la estimación de parámetros sobre modelos estadísticos han sido implementados mediante esta técnica, por ejemplo, el trabajo realizado por Liu y Pierce (1994) y Smithson y Verkuilen (2006) que estima los parámetros del modelo de regresión beta bajo la cuadratura de Gauss-Hermite. Por otra parte, se han realizado diversas modificaciones sobre la cuadratura de Gauss-Hermite original, tales como la cuadratura de Gauss-Hermite adaptativa y algunas mejoras sobre esta como la cuadratura de Gauss-Hermite adaptativa con *pruning* (Hernández et al., 2014).

Los modelos de regresión mixtos que se han mencionado anteriormente sobre datos proporcionales, no se encuentran con presencia de datos en cero y/o uno, es decir inflados con ceros y/o unos, por lo que otros autores como Ospina y Ferrari (2010) presentaron una distribución beta inflada con ceros o con unos, mediante una combinación de una distribución

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

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

4.1. Modelo de regresión ZOIP mixto

Sea y_{ij} la j -ésima medida del i -ésimo grupo, una formulación matemática para el modelo es la siguiente:

$$\begin{aligned}
 y_{ij} | \gamma_{1i}, \gamma_{2i} &\stackrel{\text{ind}}{\sim} \text{ZOIP}(\mu_{ij}, \sigma_{ij}, p_{0ij}, p_{1ij}), \\
 h_1(\mu_{ij}) &= \mathbf{x}_{ij1}^\top \boldsymbol{\beta}_1 + \gamma_{1i}, \\
 h_2(\sigma_{ij}) &= \mathbf{x}_{ij2}^\top \boldsymbol{\beta}_2 + \gamma_{2i}, \\
 h_3(p_{0ij}) &= \mathbf{x}_{ij3}^\top \boldsymbol{\beta}_3, \\
 h_4(p_{1ij}) &= \mathbf{x}_{ij4}^\top \boldsymbol{\beta}_4, \\
 \gamma_{1i} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_1^2), \\
 \gamma_{2i} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_2^2),
 \end{aligned} \tag{4-1}$$

con $i = 1, 2, \dots, N$ y $j = 1, 2, \dots, n_i$. Los parámetros μ , σ , p_0 , p_1 son modelados en función de un conjunto de covariables tales que los \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. Los $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$ son vectores de parámetros desconocidos fijos asociados a las covariables y γ_{1i} , γ_{2i} son los interceptos aleato-

rios asociados al i -ésimo grupo y cada uno distribuido normal con media cero y desviación estándar λ_1 y λ_2 , respectivamente, llamados componentes de varianza; se asume también que los interceptos aleatorios γ_{1i} y γ_{2i} son independientes entre sí. 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, además 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 usada, para los parámetros de inflación p_0 y p_1 son posibles funciones de enlace logit, probit y clog-log.

Los interceptos aleatorios se asumen bajo una distribución normal, en la práctica no existe una forma aun conocida de probar si es conveniente utilizar una distribución normal para describir los interceptos aleatorios antes de realizar el modelo, es decir, netamente con los datos, posteriormente si es posible saberlo mediante el pronóstico de los interceptos aleatorios, aunque dicha validación no fue incluida en este trabajo. Ver más en McCulloch y Neuhaus (2011b) y en McCulloch y Neuhaus (2011a).

4.1.1. Inferencia estadística

Para la estimación de los parámetros del modelo de regresión dado por la expresión (4-1), por medio de máxima verosimilitud, es necesario definir el vector de parámetros y la función de verosimilitud.

El vector de parámetros para el modelo (4-1) es $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top, \lambda_1, \lambda_2)^\top$ y pertenece al espacio:

$$\Theta = \{ \boldsymbol{\theta} \in \mathbb{R}^k | \boldsymbol{\beta}_1 \in \mathbb{R}^{k_1}, \boldsymbol{\beta}_2 \in \mathbb{R}^{k_2}, \boldsymbol{\beta}_3 \in \mathbb{R}^{k_3}, \boldsymbol{\beta}_4 \in \mathbb{R}^{k_4}, \lambda_1 \in \mathbb{R}^+, \lambda_2 \in \mathbb{R}^+ \},$$

en el que $k = k_1 + k_2 + k_3 + k_4 + 2$. La distribución marginal de $\mathbf{y}_i = (y_{1i}, \dots, y_{n_i i})^\top$ es dada por:

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

Entonces la función de verosimilitud $L(\boldsymbol{\theta})$:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^N f_y(\mathbf{y}_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^N \int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f_y(y_{ij} | \gamma_{1i}, \gamma_{2i}) \cdot f(\gamma_{1i} | \lambda_1) f(\gamma_{2i} | \lambda_2) d\gamma_{1i} d\gamma_{2i}, \end{aligned}$$

Así la función de log-verosimilitud $\ell(\boldsymbol{\theta})$ esta dado por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \log \left[\int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_{1i}, \gamma_{2i}) \cdot f(\gamma_{1i}|\lambda_1) f(\gamma_{2i}|\lambda_2) d\gamma_{1i} d\gamma_{2i} \right], \quad (4-2)$$

donde $f(y_{ij}|\gamma_{1i}, \gamma_{2i})$ es la función de densidad de probabilidad ZOIP y $f(\gamma_{1i}|\lambda_1)$ y $f(\gamma_{2i}|\lambda_2)$ son funciones de densidades de probabilidad normales con desviaciones estándar λ_1 y λ_2 , respectivamente.

Para encontrar el $\boldsymbol{\theta}$ que maximiza la función $\ell(\boldsymbol{\theta})$ es necesario solucionar la integral N veces en \mathbb{R}^2 , sin embargo esta integral no tiene forma cerrada, 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 en este trabajo, se utilizó 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 \mathbb{R} 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 .

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

El conjunto de los n puntos de la cuadratura $\mathbf{P} = \{p_1, p_2, \dots, p_n\}$ 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 $\mathbf{W} = \{w_1, w_2, \dots, w_n\}$ dados por

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

Cuadratura de Gauss-Hermite adaptativa

Unidimensional

La cuadratura de Gauss-Hermite adaptativa (AGQ) es propuesta por Liu y Pierce (1994); (Pinheiro y Bates, 1995), es básicamente una transformación de los puntos asociados a la cuadratura, centrando y extendiendo alrededor del valor máximo (\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 \mathbb{R} está dada por:

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

Multidimensional

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

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

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

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

$$\int_{\mathbb{R}^q} g(x) dx \approx |\hat{Q}|^{1/2} 2^{q/2} \sum_{i=1}^{n^q} g(z_i^*) \exp(z_i^\top z_i) a_i, \quad (4-3)$$

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

de la matriz \hat{Q} , que se encuentra dada por:

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

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 optimización más elevado, debido a la transformación de los puntos de cuadratura, pero no todos los puntos de la AGQ aportan de manera significativa un valor sobre la solución de la integral, es por esto que Hernández et al. (2014) desarrolla un mejoramiento de la AGQ, donde elimina los puntos de la cuadratura que no son significativos sobre la solución de la integral, de modo que no afectan de manera significativa los resultados finales de la integral, pero si afectan de manera positiva el tiempo de ejecución, dicho mejoramiento es llamado cuadratura de Gauss-Hermite con *pruning*.

La cuadratura de Gauss-Hermite adaptativa con *pruning* consiste en eliminar puntos de la cuadratura, tales que el peso a_i asociado al punto es menor que un valor de referencia θ , estos puntos tienen la característica de estar en los extremos de la región de integración, de este modo, al ser una cuadratura adaptativa, los puntos de los extremos no influyen de manera significativa sobre el resultado de la integración numérica aproximada, la referencia θ está dado por:

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

donde $w_{[1]}$ y $w_{[\frac{n+1}{2}]}$ corresponden respectivamente, a el valor mínimo y la mediana de los pesos originales \mathbf{W} . Ver más detalles en Hernández et al. (2014).

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

En la función de log-verosimilitud definida en (4-2) se tiene que para cada i -ésimo grupo se debe resolver la siguiente integral:

$$\begin{aligned}
I_i &= \int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_{1i}, \gamma_{2i}) \cdot f(\gamma_{1i}|\lambda_1) f(\gamma_{2i}|\lambda_2) d\gamma_{1i} d\gamma_{2i} \\
&= \int_{\mathbb{R}^2} \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_{1i}, \gamma_{2i}) \cdot \frac{\exp(\gamma_{1i}^2/2\lambda_1^2)}{\lambda_1\sqrt{2\pi}} \cdot \frac{\exp(\gamma_{2i}^2/2\lambda_2^2)}{\lambda_2\sqrt{2\pi}} d\gamma_{1i} d\gamma_{2i}
\end{aligned}$$

Si se realiza el siguiente cambio de variables

$$\begin{aligned}
b_{1i} &= \frac{\gamma_{1i}}{\sqrt{2}\lambda_1} & b_{2i} &= \frac{\gamma_{2i}}{\sqrt{2}\lambda_2} \\
\therefore b_{1i}^2 &= \frac{\gamma_{1i}^2}{2\lambda_1^2} & b_{2i}^2 &= \frac{\gamma_{2i}^2}{2\lambda_2^2} \\
\therefore \gamma_{1i} &= \sqrt{2}\lambda_1 b_{1i} & \gamma_{2i} &= \sqrt{2}\lambda_2 b_{2i}
\end{aligned}$$

La integral I_i se convierte en:

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

La integral definida en (4-4) 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, debido a que tiene una forma similar al resultado de la ecuación propuesta en (4-3), de este modo la integral I_i es aproximada por:

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

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:

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

y la función de log-verosimilitud está dada por:

$$\ell(\boldsymbol{\theta}) \approx \sum_{i=1}^N \log \left[\sum_{k_1=1}^{Q_1} \sum_{k_2=1}^{Q_2} \prod_{j=1}^{n_i} f(y_{ij} | \sqrt{2}\lambda_1 z_{k_1}, \sqrt{2}\lambda_2 z_{k_2}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) \cdot \frac{w_{k_1} w_{k_2}}{\pi} \right] \quad (4-5)$$

Con la función de log-verosimilitud anterior se pueden utilizar algoritmos de optimización, tales como las funciones de R, `nlminb` u `optim` de R para hallar los estimadores máximos verosímiles.

La estimación de parámetros del modelo de regresión ZOIP con intercepto aleatorio en μ y σ en esta investigación, se hizo utilizando máxima verosimilitud vía cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*, que fue implementado en el paquete **ZOIP** de R, por medio de la función `RMM.ZOIP`.

4.2. Modelo de regresión ZOIP mixto en el paquete ZOIP

En esta sección se mostrará cómo ajustar un modelo de regresión ZOIP con interceptos aleatorios en los parámetros de μ y σ , mediante el paquete **ZOIP** de R, utilizando el método de máxima verosimilitud y su aproximación mediante la cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*.

Función `RMM.ZOIP` del paquete **ZOIP**

La función `RMM.ZOIP` estima los parámetros de un modelo de regresión ZOIP con y sin covariables y con interceptos aleatorios en los parámetros de μ y σ ; dicha estimación se realiza vía máxima verosimilitud utilizando la cuadratura de Gauss-Hermite adaptativa multidimensional con o sin *pruning*. La función `RMM.ZOIP` usa los optimizadores `nlminb` u `optim` para la estimación de los efectos fijos, así mismo y con la ayuda de la aproximación de la cuadratura de Gauss-Hermite se estiman los componentes de varianza de λ_1 y λ_2 . La estructura de la función `RMM.ZOIP` es la siguiente:

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

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

- `formula.mu`: fórmula que define la variable respuesta y la estructura de covariables para modelar el parámetro μ , por ejemplo si se escribe $y \sim x_1 + x_2$ significa que la variable respuesta es y y que $h(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

- **formula.sigma**: fórmula que define la función de regresión de efectos fijos, para el parámetro σ , un valor posible es $\sim \mathbf{x1}$, por defecto ~ 1 .
- **formula.p0**: fórmula que define la función de regresión de efectos fijos, para el parámetro p_0 , un valor posible es $\sim \mathbf{x1}$, por defecto ~ 1 .
- **formula.p1**: fórmula que define la función de regresión de efectos fijos, para el parámetro p_1 , un valor posible es $\sim \mathbf{x1}$, por defecto ~ 1 .
- **data**: es el conjunto de datos en formato **data.frame** donde las variables deben tener el mismo nombre como se especificaron en el modelo.
- **formula.random**: fórmula que define el efecto mixto dentro del modelo. Debe ser solo el intercepto aleatorio que se tendrá en cuenta en el parámetro de la media y la dispersión, la estructura admisible es la siguiente **formula.random = 1 | G1**, donde **G1** es la variable que indica los grupos o sujetos en el modelo, siempre debe ser definido.
- **family**: elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP. El valor de ‘**R-S**’ indicará la distribución beta con parametrización Rigby y Stasinopoulos (2005), si toma el valor de ‘**F-C**’ se utilizará la distribución beta parametrización Ferrari y Cribari-Neto (2004), si el valor es ‘**Original**’ se utilizará la distribución beta con parametrización original y si es ‘**Simplex**’ se utilizará la distribución simplex.
- **link**: es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y fórmula. Si la función de regresión no posee covariables para la explicación de los parámetros de μ , σ , p_0 o p_1 ; entonces se debe utilizar como función enlace la opción **identity**, independientemente de la parametrización deseada (familia). Los posibles valores para las funciones enlace son **identity**, **logit** y **log**. Por defecto
`link=c(‘identity’, ‘identity’, ‘identity’, ‘identity’)`.
- **optimizer**: elección del optimizador, utilizado para encontrar la convergencia de la máxima verosimilitud en los parámetros de efectos fijos, se puede elegir el valor de ‘**nlminb**’ u ‘**optim**’, por defecto ‘**nlminb**’.
- **n.points**: número de puntos a utilizar en la aproximación de la función de verosimilitud por medio de la cuadratura de Gauss-Hermite adaptativa multidimensional, por defecto es 11, se recomienda dar números impares para así dejar un punto de la cuadratura en el centro de la función y no dar un valor muy grande a este parámetro, por que afectará de manera significativamente los tiempos de convergencia del modelo.
- **pruning**: es un valor booleano que indica si se utiliza *pruning* o no, para la cuadratura de Gauss-Hermite adaptativa multidimensional. Por defecto es TRUE.

En el siguiente ejemplo se mostrará el ajuste de un modelo de regresión ZOIP mixto, usando un conjunto de datos simulados. primero se muestra el código utilizado para generar los datos, luego se muestra cómo usar la función **RMM.ZOIP** para ajustar el modelo y por último la salida del modelo. Los datos simulados se generan de una distribución ZOIP con la parametrización de Rigby y Stasinopoulos (2005), los parámetros μ y σ tendrán un intercepto aleatorio y una

covariable discreta definida como el logaritmo natural de la cantidad de días (2, 10, 20, 40), los parámetros de inflación son fijados como $p_0 = p_1 = 0.1$ (Si se usan valores más altos para estos parámetros será más difícil la estimación de los parámetros de la parte continua de la distribución ZOIP, sin embargo con un gran cantidad de datos no se tendrá problema; la única restricción que existe sobre estos parámetros es que la suma de ambos debe estar entre cero y uno), el número de grupos o sujetos será $N = 21$ grupos.

$$\begin{aligned}
 y_{ij} | \gamma_{1i}, \gamma_{2i} &\stackrel{\text{ind}}{\sim} ZOIP(\mu_{ij}, \sigma_{ij}, p_0, p_1), \\
 h_1(\mu_{ij}) &= 1.6 - 1.3 \log(dias) + \gamma_{1i}, \\
 h_2(\sigma_{ij}) &= 0.1 - 0.8 \log(dias) + \gamma_{2i}, \\
 p_0 &= 0.1, \\
 p_1 &= 0.1, \\
 \gamma_{1i} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_1 = 1), \\
 \gamma_{2i} &\stackrel{\text{i.i.d}}{\sim} N(0, \lambda_2 = 0.5),
 \end{aligned}$$

De esta forma el vector de parámetros θ a estimar es $\theta = (1.6, -1.3, 0.1, -0.8, 0.1, 0.1, 1, 0.5)^\top$.

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

```

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

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

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

Days <- rep(Times, times = N)
gamma1i <- rep(rnorm(n = N, sd = 1), each = length(Times))
gamma2i <- rep(rnorm(n = N, sd = 0.5), each = length(Times))

neta1 <- 1.6 + gamma1i - 1.3 * log(Days)
neta2 <- 0.1 + gamma2i - 0.8 * log(Days)

```

```
mu <- 1/(1 + exp(-neta1))
sigma <- 1/(1 + exp(-neta2))

p0 <- 0.1
p1 <- 0.1
```

Se verifica que no hayan valores de μ y σ iguales a unos o a ceros, debido a que el dominio de los parámetros de μ y σ en la parametrización de Rigby y Stasinopoulos (2005) esta entre cero y uno, tal como se vio en la sección 2.1.3, para luego simular los valores de la variable respuesta.

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

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

family <- "R-S"

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

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

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

formula.random <- ~1 | subject

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



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

mod <- RMM.ZOIP(formula.mu = formula.mu, formula.sigma = formula.sigma,
  formula.p0 = formula.p0, formula.p1 = formula.p1, data = base,
  formula.random = formula.random, link = link, family = family,
  optimizer = optimizer, n.points = n.points, pruning = pruning)
mod #Para ver la salida del modelo
```

En el paquete **ZOIP** se tiene una función genérica (método S3) llamada `print` que sirve para mostrar los resultados de un modelo de regresión mixto para datos proporcionales, los resultados obtenidos se muestran a continuación.

```
## Call:
## RMM.ZOIP(formula.mu = formula.mu, formula.sigma = formula.sigma,
##   formula.p0 = formula.p0, formula.p1 = formula.p1, data = base,
##   formula.random = formula.random, link = link, family = family,
##   optimizer = optimizer, n.points = n.points, pruning = pruning)
##
## Results:
##
## Estimated fixed coefficients for h(mu):
## (Intercept)    log(Days)
##    1.845423    -1.331609
##
## Estimated fixed coefficients for h(sigma):
## (Intercept)    log(Days)
##    0.2795793   -0.5611416
##
## Estimated fixed coefficients for h(p0):
## (Intercept)
##    0.08319893
##
## Estimated fixed coefficients for h(p1):
## (Intercept)
##    0.1191148
```

```
## Estimated random coefficients for h(mu) and h(sigma)
##                               log(.)
## Random Intercept mu      0.7610678 -0.2730329
## Random Intercept sigma 0.5038028 -0.6855703
##
## message
## [1] "relative convergence (4)"
##
## time
## [1] 227.48
##
## iterations
## [1] 39
##
## Log-likelihood
## [1] 15.75655
```

De la salida anterior se obtienen varios resultados importantes del modelo ajustado. Leyendo de arriba hacia abajo, se observa primero los efectos fijos estimados para el parámetro μ , luego los efectos fijos estimados para el parámetro σ , luego los valores estimados del parámetro de inflación de ceros p_0 , después los efectos estimados de la regresión del parámetro de inflación de unos p_1 , seguido de la predicción de los interceptos aleatorios para μ y para σ , en ella se muestra una matriz donde la primera columna son los valores de la desviación estándar asociada a la distribución normal, es decir, el valor asociado a λ_1 y λ_2 , dichos valores generan el efecto aleatorio sobre los parámetros de μ y σ en su intercepto, respectivamente, la segunda columna nos muestra el valor del logaritmo natural de λ_1 y λ_2 , respectivamente; luego el siguiente resultado es el valor de la log-verosimilitud del modelo ajustado, para ser utilizado como posible comparación entre modelos, después se muestra un mensaje de convergencia heredado del algoritmo de optimización `nlminb` u `optim`, luego el tiempo que demoró el ajuste del modelo en segundos y por último el número de iteraciones necesarias para que el algoritmo de búsqueda convergiera en el punto máximo de la función de verosimilitud.

En el paquete **ZOIP** se tiene una función genérica de método S3 (`summary`) que permite obtener una tabla de resumen usual en modelos de regresión, esta función dará resultados más detallados de la estimación de los parámetros, tendrá como resultado para cada parámetro de regresión del modelo ZOIP mixto, el valor estimado, el error estándar asociado, el valor Z de la distribución normal y su respectivo valor p; esto ayudará al usuario del paquete **ZOIP** a concluir con más argumentos el ajuste de sus parámetros y covariables dentro del modelo de regresión ZOIP mixto ajustado. Esta tabla se muestra a continuación.

```
summary(mod)
## -----
## Fixed effects for logit(mu)
## -----
##           Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)  1.84542    0.32408   5.6944 1.238e-08 ***
## log(Days)    -1.33161    0.10306 -12.9202 < 2.2e-16 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Fixed effects for logit(sigma)
## -----
##           Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.27958    0.40333  0.6932   0.4882
## log(Days)    -0.56114    0.12949 -4.3335 1.468e-05 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Fixed effects for identity(p0)
## -----
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.083199   0.030112  2.763 0.005727 **
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Fixed effects for identity(p1)
## -----
##           Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 0.119115   0.035352  3.3694 0.0007532 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## Random effects for mu and sigma
## -----
##           Estimate Std. Error z value  Pr(>|z|)
## Random Intercept mu    0.76107    0.23334  3.2616 0.0011077 **
```

```
## Random Intercept sigma 0.50380    0.14046  3.5867 0.0003349 ***
## ---
## Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
## -----
## -----
```

De los anteriores resultados se puede observar que todos los parámetros ajustados son estadísticamente significativos y bastantes parecidos a los elementos de θ , sin embargo, se podría decir que el valor que no es significativo y que no es tan cercano al valor real es el intercepto del parámetro σ el cual se estima como 0.27 y el valor real es 0.1, no es mucha la diferencia y mucho más si se observa la desviación estándar del parámetro estimado, el cual es 0.4, pero la salida del modelo indica, con un valor p de 0.49, que no es estadísticamente significativo. La función `RMM.ZOIP` entrega siempre los valores estimados sin ninguna función enlace de por medio.

4.3. Aplicación

En esta sección se muestran diferentes resultados sobre el ajuste de un modelo de regresión ZOIP con intercepto aleatorio en el parámetro de μ y σ , por medio del paquete **ZOIP**. Primero se realizó el ajuste de un modelo de regresión ZOIP mixto con datos reales, en donde la variable respuesta fue el porcentaje de utilización de la tarjeta de crédito (tdc), valorado por el efecto de la ciudad donde vive la persona dueña de la tdc. Segundo se realizó un estudio de simulación del modelo de regresión ZOIP mixto basado en la aplicación a datos reales. Este estudio permite analizar la convergencia de la estimación de los parámetros asociados a los efectos fijos y aleatorios del modelo de regresión mixto, además poder analizar y determinar cuál es la mejor alternativa para estimar los modelos de regresión ZOIP mixto, sobre las diferentes opciones de utilizar la cuadratura de Gauss-Hermite adaptativa multidimensional.

4.3.1. Datos reales

En una entidad financiera es importante estudiar los efectos de ciertas variables sobre el porcentaje de utilización de las tarjetas de crédito de la entidad, esto dará respuestas a la activación de campañas publicitarias y estudios de mercadeo para incentivar la utilización de las tdc, de acuerdo a la situación o características de la persona dueña de la tdc. Por esta razón se analizó el efecto que tienen las variables *total mora* y *ciudad*, sobre el porcentaje de utilización de la tdc. La variable *total mora* indica el tiempo en meses que la tarjeta ha entrado en mora en toda la vida de la tdc y la *ciudad* se refiere dónde vive el cliente dueña de la tdc. Estas dos variables son importantes para la entidad financiera porque ayudan a explicar

el porcentaje de utilización de una tdc para cuando un cliente ha llegado a estar en mora varios meses y vive en cierta ciudad, por lo que indicará a la entidad financiera, como actuar sobre ciertas personas de interés. Para poder cuantificar dicho efecto se planteó un modelo de regresión ZOIP-beta con intercepto aleatorio en el parámetro de la media y la varianza, dado por la variable *ciudad* y un efecto fijo en la media y la varianza dado por la variable *total mora*, el modelo es planteado bajo la distribución ZOIP-beta con parametrización de Rigby y Stasinopoulos (2005).

La base de datos cuenta con 15 tarjetas de crédito por cada una de las 10 ciudades (Bogotá, Medellín, Cali, Barranquilla, Bucaramanga, Cartagena, Cúcuta, Ibagué, Envigado y Neiva) el modelo considerado en esta aplicación es:

$$\begin{aligned}
 y_{ij} | \gamma_{1i}, \gamma_{2i} &\stackrel{\text{ind}}{\sim} \text{ZOIP}(\mu_{ij}, \sigma_{ij}, p_0, p_1), \\
 h_1(\mu_{ij}) &= \beta_{10} + \gamma_{1i} + \beta_{11}x_{1ij}, \\
 h_2(\sigma_{ij}) &= \beta_{20} + \gamma_{2i} + \beta_{21}x_{1ij}, \\
 h_3(p_0) &= \beta_{30}, \\
 h_4(p_1) &= \beta_{40},
 \end{aligned} \tag{4-6}$$

donde y_{ij} es el porcentaje de utilización de la j -ésima tdc perteneciente a la i -ésima ciudad, $i = 1, 2, \dots, 10$, es decir, $N = 10$ asociada a las 10 ciudades; $j = 1, 2, \dots, 15$, es decir, $n_i = 15$ asociada al número de tdc por cada ciudad, en este caso todas las ciudades tienen el mismo número de observaciones, es decir, se cuenta con una muestra balanceada (si no se cuenta con una muestra balanceada el modelo funciona de igual manera); x_{1ij} : es el valor del tiempo en mora en meses de la j -ésima tdc asociada a la i -ésima ciudad; γ_{1i} y γ_{2i} son los interceptos aleatorios para los parámetros de la media y la varianza, respectivamente, asociados a la i -ésima ciudad y provenientes de la distribución normal con media cero y desviación estándar λ_1 para γ_{1i} y λ_2 para γ_{2i} . Las funciones enlace asociadas a cada regresión se tomaron como una función *logit*, esto debido a que el modelo de regresión ZOIP-beta mixto se planteó bajo la parametrización de Rigby y Stasinopoulos (2005) y este deberá utilizar dicha función enlace, tal cual como se explicó en el capítulo anterior.

En la Tabla 4-1 se muestra la estimación de los efectos fijos y componentes de varianza λ_1 y λ_2 del modelo de regresión ZOIP-beta mixto, estimado vía máxima verosimilitud y mediante la aproximación de la cuadratura de Gauss-Hermite adaptativa multidimensional utilizando 11 puntos de cuadratura con *pruning*. En esta tabla se evidencia como al tener un tiempo de mora más alto aumenta el porcentaje medio y la varianza de utilización de las tdc, además se observa que $\hat{\lambda}_1 = 0.51$ dando así que $\gamma_{1i} \sim N(0, 0.51^2)$, lo que permite ver el efecto del cambio de ciudad sobre el porcentaje medio de utilización de las tdc, el valor $\hat{\lambda}_2$ es de 0.4, evidenciando que $\gamma_{2i} \sim N(0, 0.40^2)$, lo que indicará cuantificar el efecto del cambio de ciu-

| Parámetro | β 's | Estimación | Error estándar | Valor P |
|-----------|--------------------|------------|----------------|--------------|
| μ | $\hat{\beta}_{10}$ | -1.13 | 0.24 | $5.4e^{-6}$ |
| | $\hat{\beta}_{11}$ | 0.33 | 0.13 | 0.008 |
| | $\hat{\lambda}_1$ | 0.51 | 0.304 | 0.093 |
| σ | $\hat{\beta}_{20}$ | 0.33 | 0.20 | 0.095 |
| | $\hat{\beta}_{21}$ | 0.14 | 0.09 | 0.157 |
| | $\hat{\lambda}_2$ | 0.40 | 0.31 | 0.199 |
| p_0 | $\hat{\beta}_{30}$ | 0.23 | 0.03 | $3.4e^{-11}$ |
| p_1 | $\hat{\beta}_{40}$ | 0.07 | 0.02 | 0.0011 |

Tabla 4-1: Estimación de los efectos fijos y los componentes de varianza λ_1 y λ_2 del modelo de regresión ZOIP mixto para el porcentaje utilización de una tdc.

dad sobre la variabilidad del porcentaje de utilización de las tdc, por otra parte se evidencia que alrededor de un 23 % de las tdc no se utilizan para nada ($\hat{\beta}_{30} = 0.23$) y que un 7 % tienen utilizado la totalidad de su cupo de la tarjeta de crédito ($\hat{\beta}_{40} = 0.07$). Por último, el modelo fue ajustado mediante la función `RMM.ZOIP` del paquete **ZOIP**, tardando un total de 83.5 segundos en un computador con una memoria RAM de 8GB y un procesador Intel(R) Core(TM) i5, un tiempo relativamente prudente para el ajuste un modelo complejo, como lo es el modelo de regresión ZOIP-beta mixto.

El modelo propuesto en (4-6) se puede reescribir con los parámetros estimados así:

$$\begin{aligned}
y_{ij} | \gamma_{1i}, \gamma_{2i} &\stackrel{\text{ind}}{\sim} ZOIP(\mu_{ij}, \sigma_{ij}, p_0, p_1), \\
h_1(\mu_{ij}) &= -1.13 + \gamma_{1i} + 0.33x_{1ij}, \\
h_2(\sigma_{ij}) &= 0.33 + \gamma_{2i} + 0.14x_{1ij}, \\
h_3(p_0) &= 0.23, \\
h_4(p_1) &= 0.07,
\end{aligned} \tag{4-7}$$

donde $\gamma_{1i} \sim N(0, 0.51^2)$ y $\gamma_{2i} \sim N(0, 0.40^2)$.

4.3.2. Datos simulados

En esta sección se muestran los resultados de un estudio de simulación para estudiar el proceso de estimación de parámetros en el modelo ZOIP mixto bajo la parametrización de Rigby y Stasinopoulos (2005). Para el estudio de simulación se tomó la estructura del modelo final ajustado en la aplicación a datos reales de la sección anterior, así como se describió en el modelo estimado en (4-7). En el estudio de simulación se plantearon 18 escenarios de simulación y se realizaron 1000 réplicas en cada escenario; los escenarios corresponden a todas las combinaciones entre variar el número de puntos de la cuadratura de Gauss-Hermite ($Q = 3, Q = 10, Q = 20$), variar el tamaño de muestra de cada grupo o ciudad

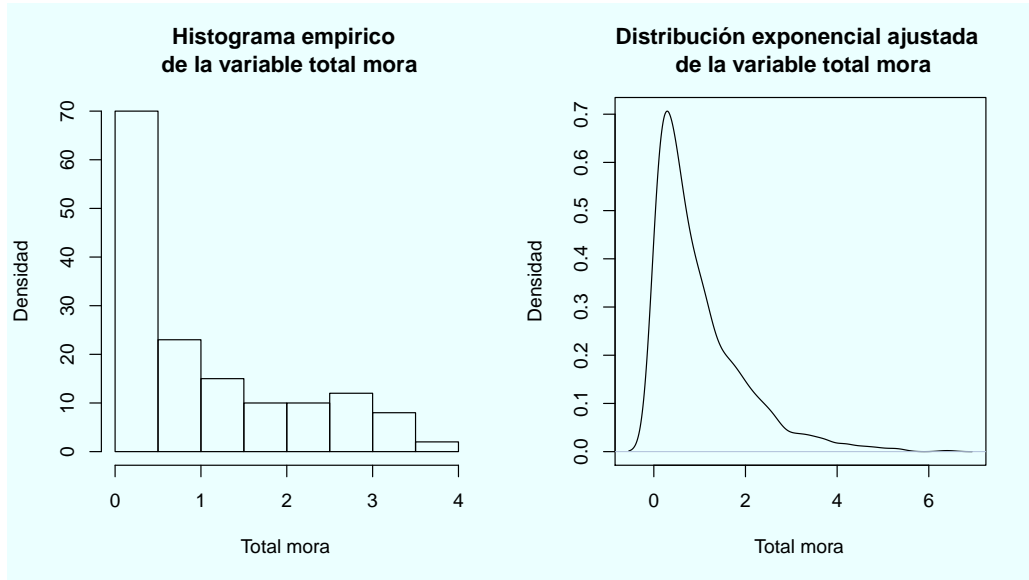


Figura 4-1: Ajuste de la distribución exponencial a la variable *total mora*

($n_i = 5, n_i = 20, n_i = 50$) y tener en cuenta si la cuadratura de Gauss-Hermite adaptativa multidimensional se realizará con o sin *pruning*.

Para poder variar el tamaño de muestra de cada ciudad fue necesario encontrar la distribución que genera la covariable *total mora*. En la Figura 4-1 se muestra como la distribución exponencial con parámetro $\lambda = 1.075$ describe de forma adecuada el comportamiento de la variable *total mora* descrita a partir de los datos reales obtenidos en la sección anterior, se eligió una distribución exponencial ya que es una distribución bastante pertinente para modelar variables de tiempo y en este caso es válido para la definición de la variable *total mora*.

A continuación se muestran los resultados de la simulación del modelo de regresión ZOIP-beta mixto, realizada bajo la función `RMM.ZOIP` del paquete **ZOIP** de R.

En la Tabla 4-2 se muestran las medianas de los valores estimados para cada uno de los parámetros del modelo simulado considerando tres tamaños de muestra n_i y si se utilizó *pruning* o no. En dicha tabla se nota como los valores de λ_1 y λ_2 asociados a los interceptos aleatorios van convergiendo a su valor verdadero a medida que el tamaño de muestra va aumentando sin importar si se hizo con o sin *pruning*; además se nota cómo las estimaciones de las pendientes β_{11} y β_{21} va convergiendo a medida que el tamaño de muestra va creciendo, este parámetro da un valor muy plausible desde tamaños de muestra pequeños, sin importar si se hizo con o sin *pruning*. Este último fenómeno se evidencia también con los parámetros

| Parámetro | Valor verdadero de β 's | Con <i>pruning</i> | | | Sin <i>pruning</i> | | |
|------------------------------|-------------------------------|--------------------|------------|------------|--------------------|------------|------------|
| | | $n_i = 5$ | $n_i = 20$ | $n_i = 50$ | $n_i = 5$ | $n_i = 20$ | $n_i = 50$ |
| μ | $\beta_{10} = -1.13$ | -1.137 | -1.110 | -1.076 | -1.128 | -1.120 | -1.080 |
| | $\beta_{11} = 0.33$ | 0.321 | 0.327 | 0.326 | 0.331 | 0.330 | 0.327 |
| | $\lambda_1 = 0.51$ | 0.879 | 0.576 | 0.507 | 0.882 | 0.568 | 0.498 |
| σ | $\beta_{20} = 0.33$ | 0.445 | 0.380 | 0.336 | 0.452 | 0.377 | 0.345 |
| | $\beta_{21} = 0.14$ | 0.072 | 0.118 | 0.132 | 0.066 | 0.121 | 0.133 |
| | $\lambda_2 = 0.4$ | 0.728 | 0.450 | 0.396 | 0.727 | 0.456 | 0.398 |
| p_0 | $\beta_{30} = 0.23$ | 0.220 | 0.230 | 0.230 | 0.220 | 0.230 | 0.230 |
| p_1 | $\beta_{40} = 0.07$ | 0.060 | 0.070 | 0.070 | 0.060 | 0.070 | 0.072 |
| Mediana del tiempo(Seg) | | 115.72 | 130.85 | 140.58 | 61.69 | 163.36 | 218.68 |
| Mediana del num. iteraciones | | 22 | 30 | 34 | 22 | 30 | 34 |

Tabla 4-2: Mediana de los parámetros estimados en el modelo ZOIP mixto para tres tamaños de muestra y con la estrategia de “con y sin *pruning*” y para todos los valores de Q .

de inflación p_0 y p_1 en los que se tiene un valor muy cercano al real desde los tamaños de muestra pequeños. En cuestión del tiempo de ejecución se ve como se reduce en aproximadamente un 50 % cuando se utiliza la metodología de la cuadratura de Gauss-Hermite adaptativa con *pruning*, con respecto a no utilizar *pruning*. Sobre el número de iteraciones se observa como al aumentar el tamaño de la muestra se requiere una cantidad mayor de iteraciones.

| Parámetro | Valor verdadero de β 's | Con <i>pruning</i> | | | Sin <i>pruning</i> | | |
|------------------------------|-------------------------------|--------------------|----------|----------|--------------------|----------|----------|
| | | $Q = 3$ | $Q = 10$ | $Q = 20$ | $Q = 3$ | $Q = 10$ | $Q = 20$ |
| μ | $\beta_{10} = -1.13$ | -1.122 | -1.069 | -1.120 | -1.117 | -1.076 | -1.129 |
| | $\beta_{11} = 0.33$ | 0.323 | 0.322 | 0.333 | 0.334 | 0.322 | 0.329 |
| | $\lambda_1 = 0.51$ | 0.632 | 0.626 | 0.634 | 0.629 | 0.616 | 0.623 |
| σ | $\beta_{20} = 0.33$ | 0.400 | 0.365 | 0.366 | 0.382 | 0.379 | 0.373 |
| | $\beta_{21} = 0.14$ | 0.117 | 0.119 | 0.120 | 0.123 | 0.117 | 0.121 |
| | $\lambda_2 = 0.4$ | 0.490 | 0.487 | 0.491 | 0.501 | 0.482 | 0.486 |
| p_0 | $\beta_{30} = 0.23$ | 0.228 | 0.228 | 0.226 | 0.226 | 0.226 | 0.226 |
| p_1 | $\beta_{40} = 0.07$ | 0.068 | 0.070 | 0.070 | 0.070 | 0.068 | 0.070 |
| Mediana del tiempo(Seg) | | 75.295 | 128.28 | 271.825 | 74.295 | 162.35 | 367.545 |
| Mediana del num. iteraciones | | 30 | 29 | 29 | 29 | 29 | 29 |

Tabla 4-3: Mediana de los parámetros estimados en el modelo ZOIP mixto para tres diferentes números de puntos de la cuadratura de Gauss-Hermite y con la estrategia de “con y sin *pruning*” y para todos los tamaños de muestra.

En la Tabla 4-3 se muestran las medianas de los valores estimados de cada uno de los parámetros del modelo simulado considerando tres diferentes números de puntos de la cuadratura de Gauss-Hermite Q y si se utilizó *pruning* o no. En dicha tabla se observa que cuando se utiliza una mayor cantidad de puntos de cuadratura Q se evidencia una pequeña mejora en la convergencia de los valores reales de los efectos fijos y los componentes de varianza λ_1 y λ_2 . Con respecto a la utilización de *pruning* o no, se observa que las estimaciones de los parámetros no tienen una diferencia significativa, por lo que es mejor uti-

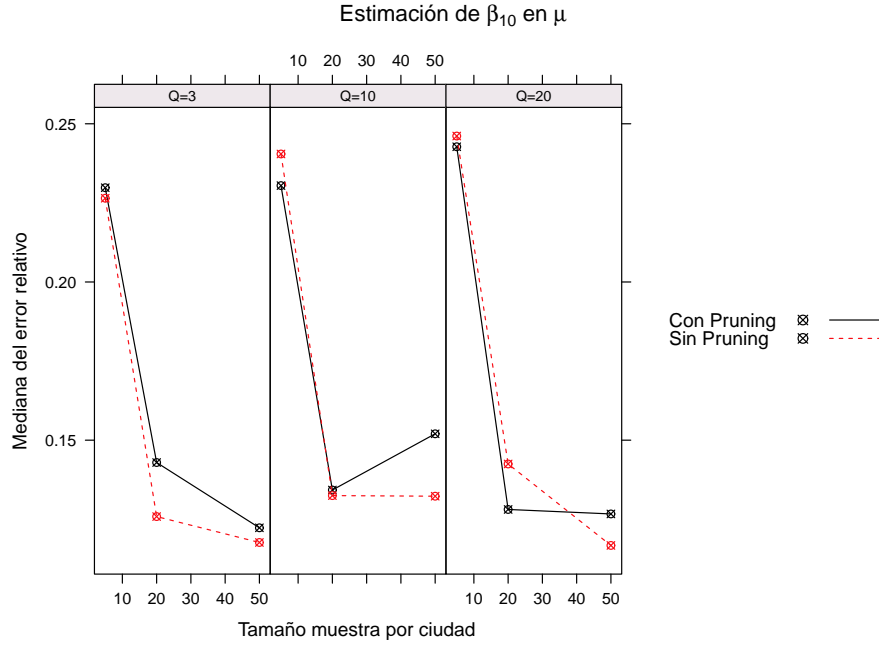


Figura 4-2: Mediana del error relativo para la estimación del parámetro β_{10} asociado a μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

lizar la cuadratura de Gauss-Hermite adaptativa con *pruning* porque al parecer se obtienen resultados similares, pero con una reducción significativa del tiempo de estimación de los parámetros. El número de iteraciones utilizadas para la estimación de los parámetros no se ve influenciado por la cantidad de puntos utilizados en la cuadratura o si se utiliza *pruning* o no.

La mediana del error relativo de un parámetro estimado, es útil para cuantificar en términos porcentuales el error cometido en la estimación sin que se vea afectado por algunos valores extremos y es calculado a partir de hallar el error relativo de las 1000 réplicas de cada escenario de simulación para cada parámetro estimado y calcular la mediana a los mil valores calculados. El error relativo se puede definir como $|(\theta - \hat{\theta})/\theta|$, donde θ representa cualquier parámetro del modelo de regresión ZOIP mixto y $\hat{\theta}$ su estimación.

En la Figura 4-2 se muestra la mediana del error relativo sobre el intercepto fijo (β_{10}) asociado a μ al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y la utilización de *pruning*. En dicha figura se observa que al aumentar el tamaño de muestra en cada ciudad se obtiene una reducción de la mediana del error relativo de β_{10} , sin embargo, al aumentar el número de puntos de la cuadratura se obtienen errores relativos muy parecidos en todos los tamaños de muestra, por lo que se podría decir que el aumento del número de

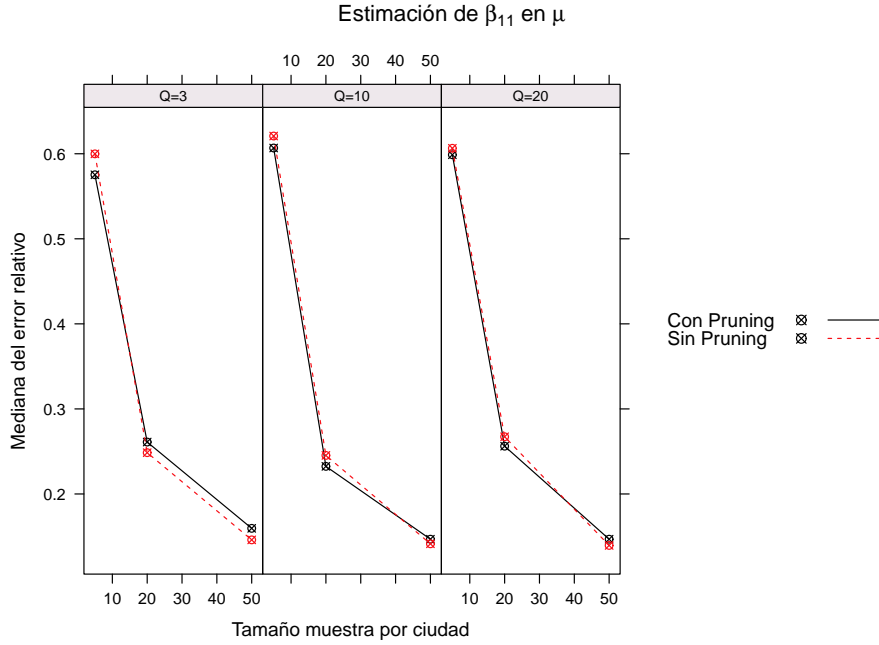


Figura 4-3: Mediana del error relativo para la estimación del parámetro β_{11} asociado a μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

puntos de cuadratura no mejora la estimación del intercepto fijo sobre el parámetro μ , por otra parte no se obtiene una diferencia en los errores relativos cuando se utiliza *pruning* y cuando no, aunque, existe un escenario de simulación cuando el tamaño de muestra es 50 y $Q = 10$ el error relativo de la utilización de *pruning* es más grande en relación al que no utiliza *pruning*.

En la Figura 4-3 se muestra la mediana del error relativo sobre el valor del efecto fijo de la variable *total mora* sobre μ , a partir de variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. En dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la mediana del error relativo, además al aumentar el número de puntos de la cuadratura se obtienen errores un poco menores, cuando se aumentan el número de puntos de la cuadratura, lo que nos permite observar que el aumento del número de puntos de la cuadratura, no afecta demasiado en la estimación de este efecto fijo, por otra parte se obtienen errores muy parecidos cuando se utiliza *pruning* y cuando no, por lo que esto no afecta la estimación del parámetro.

En la Figura 4-4 se muestra la mediana del error relativo sobre el valor de la desviación estándar de la distribución normal que genera el intercepto aleatorio (γ_1) del parámetro de μ , al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza

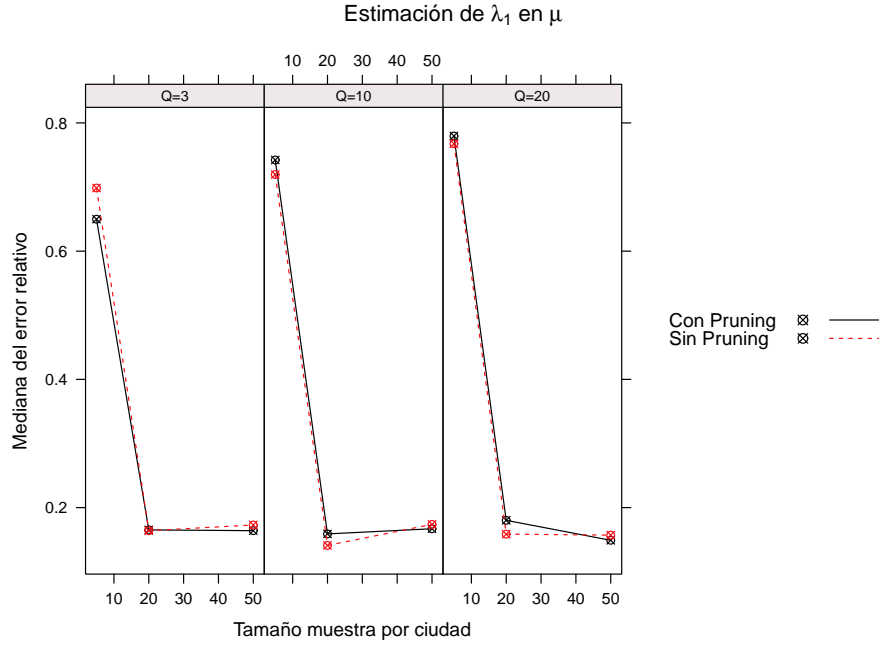


Figura 4-4: Mediana del error relativo para la estimación del parámetro λ_1 desviación estándar del intercepto aleatorio asociado a la μ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

pruning o no. En dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la mediana del error relativo, además al aumentar el número de puntos de la cuadratura se obtienen errores menores cuando se aumentan el número de puntos de la cuadratura, dicha mejora es de alrededor de un 2 % cuando el tamaño de muestra es más grande $n_i \geq 20$ y $Q = 20$, este análisis no se ve afectado por el hecho de utilizar *pruning*, por lo que se podría decir que el aumento del número de puntos de cuadratura mejora la estimación del intercepto aleatorio sobre la media cuando el tamaño de muestra es relativamente grande, por otra parte se obtienen errores relativamente parecidos cuando se utiliza *pruning* y cuando no.

En la Figura 4-5 se muestra la mediana del error relativo sobre el intercepto fijo (β_{20}) del parámetro de σ al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. En dicha figura se muestra como cuando se aumenta el tamaño de muestra en cada ciudad, se obtiene una reducción de la mediana del error relativo, por otra parte al aumentar el número de puntos de la cuadratura se obtienen errores muy parecidos en todos los tamaños de muestra, sin embargo, cuando $Q = 20$ y se realiza sin la metodología *pruning* se nota una reducción en el error, ahora si se observan las medianas de los errores relativos son relativamente parecidos cuando se utiliza *pruning* y cuando no, sin embargo, se puede rescatar que hay puntos como cuando el tamaño de muestra es 20 o 50 y $Q = 20$

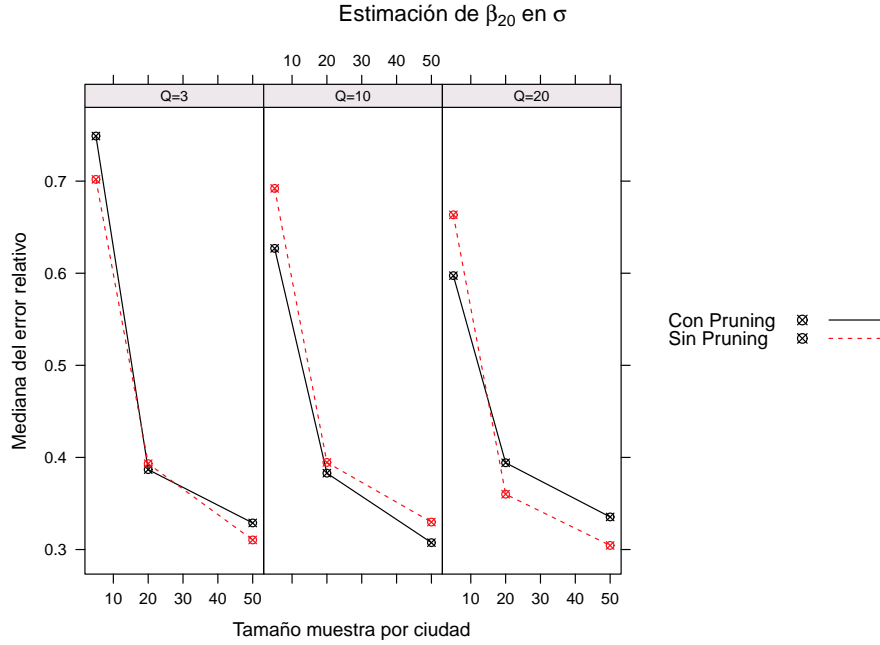


Figura 4-5: Mediana del error relativo para la estimación del parámetro β_{20} asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

el error relativo de la metodología sin utilizar *pruning* se ve reducido, mejorando así la estimación del parámetro β_{20} .

En la Figura 4-6 se muestra la mediana del error relativo sobre el valor del efecto fijo de la variable *total mora* sobre el parámetro σ al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. En dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción del error relativo, además al aumentar el número de puntos de la cuadratura no se obtiene una mejora significativa de la mediana del error relativo, otro aspecto que es importante resaltar es que las estimaciones no se ven afectadas por el hecho de utilizar la metodología *pruning*.

En la Figura 4-7 se muestra la mediana del error relativo sobre el valor de la desviación estándar de la distribución normal que genera el intercepto aleatorio (γ_2) del parámetro de σ al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. En dicha figura se muestra como el tamaño de muestra al ser aumentado obtiene una reducción de la mediana del error relativo, además al aumentar el número de puntos de la cuadratura se obtienen errores un poco menores, la mejora de alrededor de un 2% se nota a partir de $Q \geq 10$ y cuando los tamaños de muestra de cada grupo son mayores o iguales que 20, además no se nota una diferencia cuando se utiliza la metodología *pruning*.

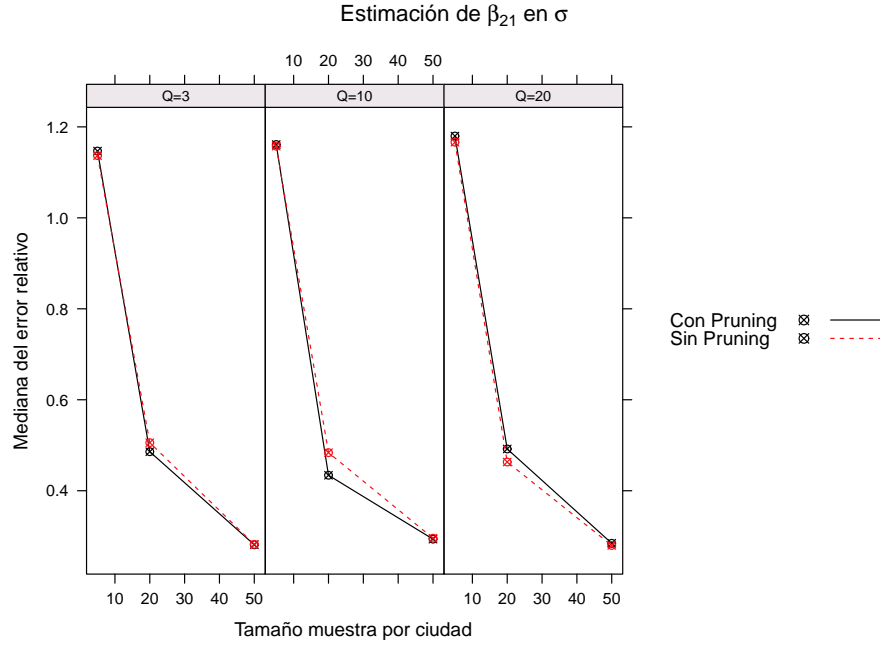


Figura 4-6: Mediana del error relativo para la estimación del parámetro β_{21} asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

ing, por lo que se podría decir que el aumento del número de puntos de cuadratura mejora la estimación del intercepto aleatorio sobre la dispersión cuando el tamaño de muestra es relativamente grande, por otra parte se obtienen errores relativamente parecidos cuando se utiliza *pruning* y cuando no, por lo que los valores de las estimaciones no se ven afectadas por la metodología *pruning*.

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

En la Figura 4-9 se muestra la mediana del error relativo sobre el valor de la estimación del porcentaje de unos dentro del modelo de regresión ZOIP mixto al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. La estimación de este parámetro se ve afectado por el tamaño de muestra elegido dentro de cada grupo, ya que se nota como al aumentar el tamaño de muestra la mediana del error relativo

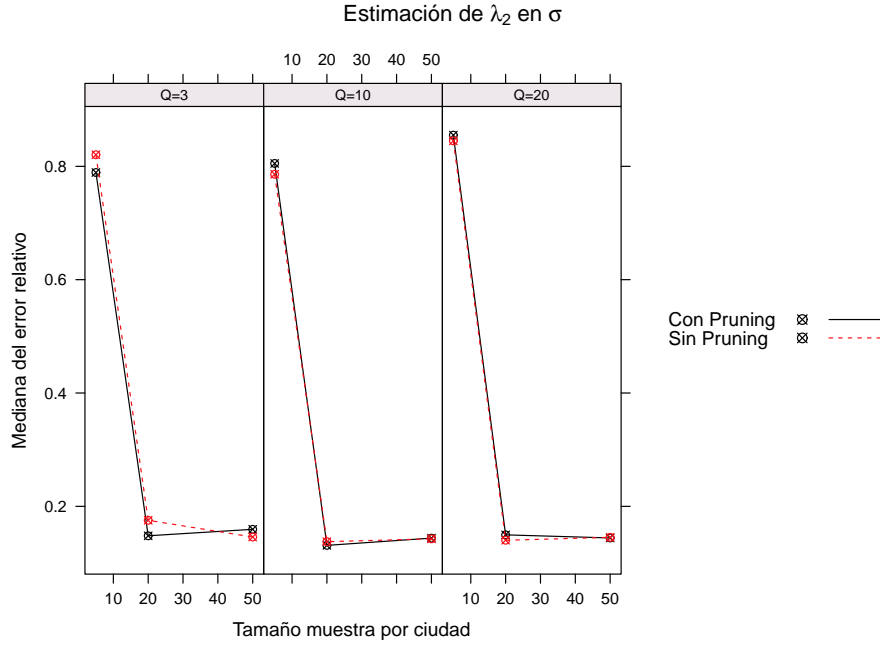


Figura 4-7: Mediana del error relativo para la estimación del parámetro λ_2 desviación estándar del intercepto aleatorio asociado a σ , variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

decrece rápidamente Al aumentar el tamaño de muestra n_i . Además se puede observar como no hay una diferencia entre la estimación del parámetro variando del número de puntos de la cuadratura ni tampoco sobre la utilización de la metodología *pruning*, sin embargo cuando $Q = 20$ y $n_i = 20$, esto no se cumple a cabalidad.

En la Figura 4-10 se muestra la mediana del tiempo de ejecución para el ajuste del modelo de regresión ZOIP mixto de ejecución al variar el tamaño de muestra n_i , el número de puntos de la cuadratura Q y si se utiliza *pruning* o no. Este modelo se ajustó utilizando la función `RMM.ZOIP` del paquete **ZOIP**, en dicha figura se nota como a medida que se va aumentando el número de puntos de la cuadratura y el tamaño de muestra la diferencia de utilizar la metodología *pruning* se hace más favorable ya que se nota una reducción en el tiempo de ejecución del modelo. Por todo el análisis previo hecho con la estimación de todos los parámetros de regresión, en el cual se nota que por lo general sin importar cualquier sea la combinación entre tamaño de muestra y el número de puntos de la cuadratura, la utilización de la metodología *pruning* no afecta la estimación de los parámetros y viendo la Figura 4-10 se ve que es más conveniente utilizar la metodología *pruning*, porque esta genera un tiempo de ejecución menor para ajustar el modelo y sin afectar de la estimación de los efectos fijos ni de los componentes de varianza del modelo de regresión ZOIP mixto. Además, se puede concluir que el hecho de aumentar el número de puntos de la cuadratura sólo beneficia a

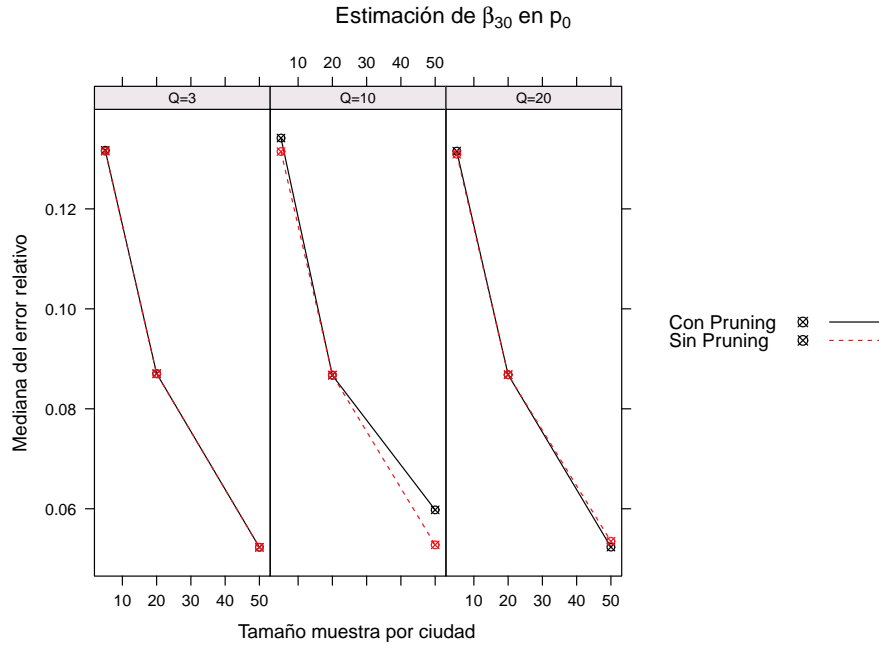


Figura 4-8: Mediana del error relativo para la estimación del parámetro β_{30} asociado al parámetro de inflación de ceros, variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

la estimación de los componentes de varianza mas no de los efectos fijos, en general lo más recomendable para mejorar las estimaciones de todos los parámetros es aumentar el tamaño muestral dentro de cada grupo.

4.4. Conclusión

El modelo de regresión ZOIP con efectos mixtos propuesto en este capítulo permite modelar los parámetros de una variable respuesta de datos proporcionales inflados con ceros y/o unos, en función de un conjunto de covariables. El modelo de regresión ZOIP mixto considera efectos fijos sobre cada uno de los parámetros de la distribución ZOIP y permite incluir interceptos aleatorios en los parámetros de μ y σ . La estimación de los parámetros del modelo se realiza por máxima verosimilitud y con la ayuda de la aproximación de la cuadratura de Gauss-Hermite adaptativa multidimensional. La estimación de los parámetros del modelo puede ser realizada usando la función `RMM.ZOIP` del paquete **ZOIP**, de una forma amigable para el usuario, en el paquete es posible realizar diferentes tipos de regresión ZOIP mixto, bajo diferentes distribuciones y parametrizaciones, además de obtener resultados del modelo mediante funciones de metodología S3 de R.

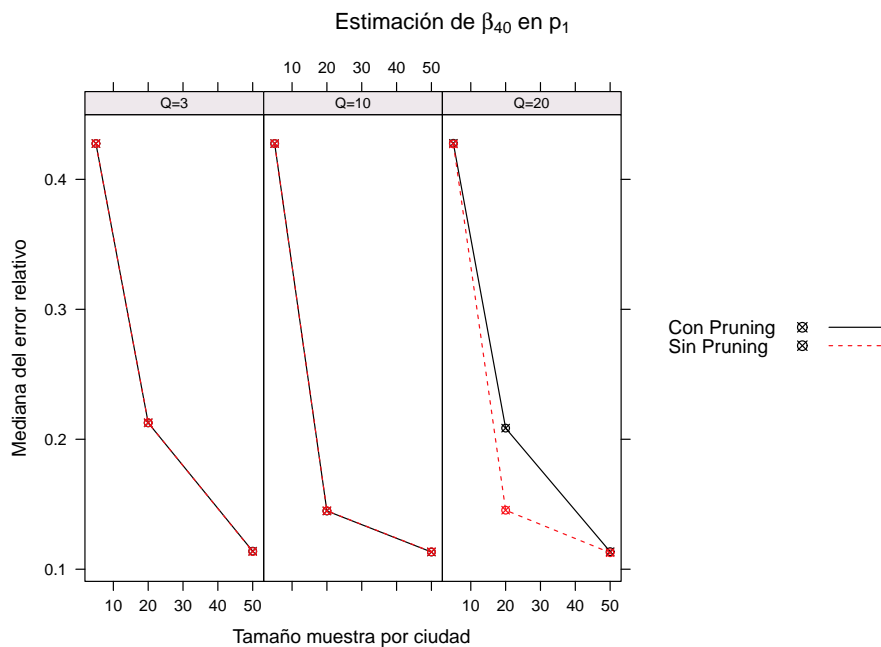


Figura 4-9: Mediana del error relativo para la estimación del parámetro β_{40} asociado al parámetro de inflación de unos, variando el tamaño de muestra, el número de puntos de la cuadratura y la utilización de *pruning*.

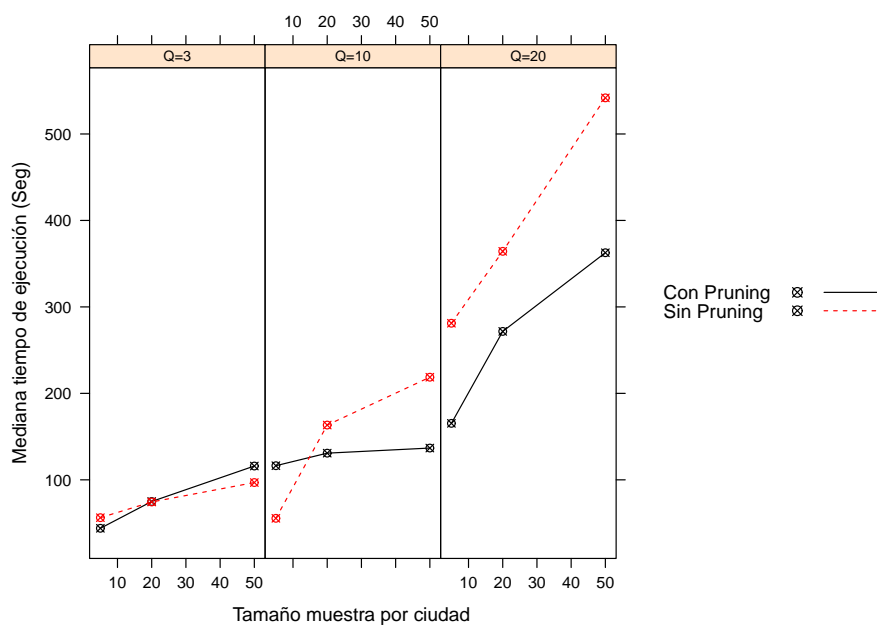


Figura 4-10: Mediana del tiempo de ejecución del modelo de regresión ZOIP mixto, bajo la función de `RMM.ZOIP` del paquete **ZOIP** de R.

El estudio de simulación realizado sobre el modelo de regresión ZOIP mixto, permite obtener varias conclusiones sobre la mejor estimación de los parámetros del modelo, una de estas conclusiones es que uno de los factores que más influyen sobre la estimación es el tamaño de muestra de cada uno de los grupos, es decir n_i , ya que este factor hace que el error relativo de la estimación de todos los parámetros se vea reducido considerablemente cuando se aumenta. Otra conclusión igual de importante es que el hecho de utilizar la metodología *pruning* hace que los valores de las estimaciones de los parámetros del modelo no cambien, pero sí que el tiempo de ejecución se vea reducido en un 50 %. Por último se puede concluir que el efecto del número de puntos de la cuadratura de Gauss-Hermite no influye demasiado en la estimación de los parámetros de efectos fijos, aunque sí afecta el aumento de este factor la estimación de los componentes de varianza de los interceptos aleatorios, un número prudente de puntos es entre cinco y quince, ya que el aumento considerado de este factor no influye en la mejora de la estimación de los parámetros, pero sí en el aumento del tiempo de ejecución.

5 Conclusiones y recomendaciones

5.1. Conclusiones

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

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

Como un resultado importante se destaca la creación del paquete **ZOIP**. Este paquete tiene varias funciones importantes como lo son la función de densidad de probabilidad, de distribución acumulada, y la función cuantil y un generador de números aleatorios. Por otra parte, el paquete permite ajustar distribuciones ZOIP o modelos de regresión ZOIP, mediante algoritmos de optimización, como `nlimnb` u `optim`. También es posible ajustar los modelos de regresión ZOIP mixtos, vía máxima verosimilitud usando la cuadratura de Gauss-Hermite adaptativa. El paquete **ZOIP** también incluye sobre el ajuste de los modelos, algunas funciones de método S3.

Se realizaron estudios de simulación que permitieron ver la convergencia de los parámetros en el ajuste de la distribución ZOIP, el modelo de regresión ZOIP y el modelo de regresión ZOIP mixto, además esto permitió concluir con que alternativas se encuentran las mejores

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

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

Para la realización de esta tesis se utilizó material bibliográfico adicional que no fue mencionado en los capítulos anteriores pero que por su valor vale la pena destacarlo. Cook et al. (2008), Houston y James (1996) y Kieschnick y McCullough (2003) sirvieron para entender aplicaciones de modelos regresión para datos proporcionales utilizando datos reales, Evans y Swartz (2000) y Lindstrom y Bates (1988) fueron útiles para entender los métodos de estimación de modelos mixtos vía algoritmos de Newton-Raphson y aproximación Monte Carlo, Además Seoane (2014) fue útil para entender a manera de introducción sobre los modelos mixtos y su utilidad, por último el software de versionamiento de código `Git` fue muy útil para mantener una versión controlada del paquete **ZOIP** y de este trabajo, para el entendimiento del funcionamiento de `Git` fue muy útil Chacon y Straub (2014).

5.2. Recomendaciones

Como posibles trabajos futuros y recomendaciones se propone:

- Extender la distribución ZOIP a otras distribuciones para datos proporcionales no tomadas en cuenta en este trabajo, tales como la distribución beta-rectangular y la distribución LogitSep.
- Realizar un estudio de análisis de residuales para los parámetros estimados de los modelos de regresión ZOIP fijos y mixtos.

- Incluir un análisis de selección de variables en los modelos de regresión ZOIP fijos y mixtos.
- Incluir interceptos aleatorios en los parámetros asociados a la inflación de la distribución ZOIP, en el modelo de regresión ZOIP mixto propuesto en este trabajo.
- Incluir pendientes aleatorias en los diferentes parámetros de la regresión ZOIP.
- Extender los interceptos aleatorios en los parámetros de la media y la dispersión, a interceptos aleatorios no normales, o normales correlacionados.
- Incluir más funciones de método S3 en el paquete **ZOIP**, tales como `plot`, `predict`, `AIC`, `BIC`, entre otras.
- Realizar una comparación entre la estimación de los parámetros del modelo de regresión ZOIP mixto, por metodologías bayesianas y utilizando máxima verosimilitud y la cuadratura de Gauss-Hermite adaptativa.
- Mejorar la eficiencia de los algoritmos implementados en el paquete **ZOIP**, mediante la integración del paquete **ZOIP** por las metodologías implementadas en el paquete **RCPP**.
- Estudiar e implementar otras metodologías de estimación de los parámetros, nunca exploradas sobre los modelos de regresión mixtos para datos proporcionales inflados con ceros y/o con unos, tales como la metodología bayesiana INLA o los algoritmos propuestos por Ogden (2015), sobre métodos de reducción secuencial.
- Evaluar más a fondo la simulación del modelo de regresión ZOIP mixto propuesto, para las demás parametrizaciones de la distribución ZOIP-beta y la distribución ZOIP-simplex.
- Optimizar los algoritmos propuestos en el paquete **ZOIP**, para que el tiempo de estimación de los parámetros sean más corto.
- Estudiar tópicos de los datos proporcionales inflados para datos proporcionales inflados con ceros y/o unos, como cuando más del 50 % de la muestra se encuentra inflada por ceros o por unos. se podría evaluar la eficacia y convergencia de las funciones implementadas, en el paquete **ZOIP** mediante estudios de simulación.
- Realizar mejoras continuas del paquete **ZOIP**, a medida que se vayan adquiriendo retroalimentaciones de los usuarios que utilizan el paquete.
- Extender los estudios de simulación del modelo de regresión ZOIP mixto, cuando se tienen muestras no balanceadas.

-
- Ampliar el tipo de funciones enlace vinculadas en el paquete **ZOIP**, para realizar los modelos de regresión, tales como funciones *probit* o *cloglog*.

Bibliografía

- W. Bonat, P. Ribeiro, y W. Zeviani. Modelo beta com efeitos aleatórios. *Computational Statistics & Data Analysis*, 2012.
- W. Bonat, P. Ribeiro, y W. Zeviani. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics online*: doi:10.1080/02664763.2014.947248, 42(2):252–266, 2015.
- S. Chacon y B. Straub. *Pro Git*. Apress, 2014.
- D. Cook, R. Kieschnick, y B. McCullough. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, 15:860–867, 2008.
- A.C.G de Andrade. Efeitos da especificação incorreta da função deligação no modelo der-regressão beta. dissertação de mestrado. *Instituto de Matemática e Estatística*, 2007.
- M. Evans y T. Swartz. *Approximating integrals via Monte Carlo and deterministic methods*. Oxford University Press Inc., 2000.
- L. Fahrmeir y G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer, 2001.
- S. Ferrari y F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.
- S. Ferrari y E. Pinheiro. Improved likelihood inference in beta regression. *Journal of statistical computation and simulation available online*:doi:10.1080/00949650903389993, 2010.
- Z. Figueroa-Zuñiga, R. arellano, y S. Ferrari. Mixed beta regression: a bayesian perspective. *Computational Statistics & Data Analysis*, 61(Supplement C):137 – 147, 2013.
- D. Galvis y V. Lachos. *Bayesian analysis of regression models for proportional data in the presence of zeros and ones*. PhD thesis, Estatal University of Campinas, 2014.
- E. Hahn. Mixture densities for project management activity times: A robust approach to pert. *European Journal of operational research*, 188(2):450–459, 2008.
- F. Hernández, O. Usuga, y V. Giampaoli. Improving the adaptive gaussian quadrature. *Journal of Statistical Software*, submitting, 2014.

- A. Hossain. Modelling a proportion response variable using generalized additive models for location scale and shape. *International Workshop on Statistical Modelling*, 2:137–140, 2015.
- J. Houston y C. James. Bank information monopolies and the mix of private and public debt claims. *Journal of Finance*, 51:1863–1889, 1996.
- B. Jørgensen. The theory of dispersion models. *Computational statistics and Data analysis*, 76, 1997.
- B. Jørgensen y Barndorff-Nielsen. Some parametric models on the simplex. *Journal of multivariate analysis*, 39(1):106–116, 1991.
- R. Kieschnick y B. McCullough. Regression analysis of variates observed on (0,1):percentages, proportions, and fractions. *Statistical Modelling*, 3(3):193–213, 2003.
- I. Kosmidis, A. Zou, y K. Jagan. Beyond beta regression: modelling percentages and fractions in the presence of boundary observations. *International Workshop on Statistical Modelling*, 1:267–272, 2015.
- N. Laird y J. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.
- M. Lindstrom y D. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.
- Q. Liu y D. Pierce. A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629, 1994.
- C. McCulloch y J. Neuhaus. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67:270–279, 2011a.
- C. McCulloch y J. Neuhaus. Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26:388–402, 2011b.
- H. Ogden. A sequential reduction method for inference in generalized linear mixed models. *Electronic Journal of Statistics*, 9:135–152, 2015.
- R. Ospina y S. Ferrari. Inflated beta distributions online: doi:10.1007/s00362-008-0125-4. *Statistical papers*, 2010.
- R. Ospina y S. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, 56(6):1609–1623, 2012.
- C. Owen. Parameter estimation for the beta distribution. Master’s thesis, Brigham Young University, 2008.

- L. Pace y A. Salvan. Principles of statistical inference from a neo-fisherian perspective. in: Advanced series on statistical science and applied probability. *World Scientific*, 4:128, 1997.
- P. Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Polical Analysis*, 9(4):325–346, 2001.
- J. Pinheiro y D. Bates. Apprximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Grapchical Statistics*, 4(1):12–35, 1995.
- Z. Qiu, P. Song, y M. Tan. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, 35(4):577–596, 2008.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <http://www.R-project.org/>.
- B. Rigby y M. Stasinopoulos. Generalized additive models for location, scale and shape. *Applied Statistical*, 54(3):507–554, 2005.
- B. Rigby y M. Stasinopoulos. Instructions on how to use the gamlss package in r. *Computational statistics and Data analysis*, 2:194–195, 2008.
- J. Seoane. ¿modelos mixtos (lineales)? una introducción para el usuario temeroso. *Temas actuales de Etología*, 2014.
- A. Simas, W. Barrero-Souza, y A. Rocha. Improved estimators for a general class of beta regression models. *Computational statistics and Data analysis*, 54:348–366, 2010.
- M. Smithson y J. Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.
- P. Song, Z. Qi, y M. Tan. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, 46(5):540–533, 2004.
- O. Usuga y V. Giampaoli. *Modelos de regressão beta com efeitos aleatórios normais e não normais para dados longitudinais*. PhD thesis, Universidade de São Paulo, 2013.
- J. Verkuilen y M. Smithson. Mixed and mixture regression model for continuos bounded responses using the beta distribution. *Journal of education and behavioral statistics*, 37(1):82–113, 2012.
- A. Zeileis, F. Cribari-Neto, y B. Grün. Beta regression in R. *Journal of statistical software*, 34(2):1–24, 2010.
- P. Zhang, Z. Qiu, y C. Shi. simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Sftware*, 71(11), 2016.