Revista Colombiana de Estadística
Edición para autor(es)

Análisis de regresión para datos proporcionales inflados con ceros y/o unos utilizando el paquete ZOIP en R

Regression analysis for proportional data inflated with zeros and ones using ZOIP package in R

Juan Camilo Díaz Zapata[1,a], Freddy Hernández Barajas[1,b]

[1]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

Resumen

Los datos obtenidos a partir de variables medidas bajo porcentajes, tasas o proporciones, suelen dar resultados con valores de cero y/o uno, conociéndose como datos proporcionales inflados con ceros y/o unos, para un caso de estudio donde se quiera explicar este tipo de datos a partir de otras variables, se estará hablando de modelo de regresión para datos proporcionales inflados con ceros y/o unos, en la literatura estadística se han desarrollado diferentes modelos que dan solución a este problema y es el modelo de regresión beta inflado con ceros y/o unos que se encuentra construido bajo tres diferentes parametrizaciones, la parametrización original, la parametrización de Rigby & Stasinopoulos (2005) y la de Ferrari & Cribari-Neto (2004), además de la distribución beta existen otros tipos de modelos que tratan este problema, el modelo de regresión simplex inflado con ceros y/o unos (Galvis & Lachos 2014) es uno de ellos, sin embargo no existe un modelo de regresión basado en las principales distribuciones para datos proporcionales inflados con ceros y/o unos que lo haga en una sola función, además a pesar de que el modelo de regresión beta inflado con ceros y/o unos de Rigby & Stasinopoulos (2005) se encuentra implementado en R, y que este requiere utilizar diferentes funciones para ajustar el modelo de regresión, dependiendo si los datos son inflados solo con ceros, unos, o ambos, no existe un paquete en R que lo haga en una sola función para diferentes distribuciones y parametrizaciones. En este trabajo se busca implementar mediante el paquete **ZOIP** de R un modelo de regresión que se base en las principales distribuciones para datos proporcionales inflados con ceros y/o unos (Modelo de regresión ZOIP), mediante la distribución ZOIP (Zeros Ones Inflated Proportional), estimando los parámetros vía máxima verosimilitud, en una sola función. Se

[a]Estudiante de maestría en estadística. E-mail: jucdiazza@unal.edu.co
[b]Profesor asistente. E-mail: fhernanb@unal.edu.co

realiza un estudio de simulación que muestra la convergencia satisfactoria de los parámetros de los diferentes modelos aplicados y se aplica el modelo de regresión ZOIP a datos reales.

***Palabras clave*:** Datos proporcionales inflados, Regresion ZOIP, Máxima verosimilitud, R.

Abstract

Los datos obtenidos a partir de variables medidas bajo porcentajes, tasas o proporciones, suelen dar resultados con valores de cero y/o uno, conociéndose como datos proporcionales inflados con ceros y/o unos, para un caso de estudio donde se quiera explicar este tipo de datos a partir de otras variables, se estará hablando de modelo de regresión para datos proporcionales inflados con ceros y/o unos, en la literatura estadística se han desarrollado diferentes modelos que dan solución a este problema y es el modelo de regresión beta inflado con ceros y/o unos que se encuentra construido bajo tres diferentes parametrizaciones, la parametrización original, la parametrización de Rigby & Stasinopoulos (2005) y la de Ferrari & Cribari-Neto (2004), además de la distribución beta existen otros tipos de modelos que tratan este problema, el modelo de regresión simplex inflado con ceros y/o unos (Galvis & Lachos 2014) es uno de ellos, sin embargo no existe un modelo de regresión basado en las principales distribuciones para datos proporcionales inflados con ceros y/o unos que lo haga en una sola función, además a pesar de que el modelo de regresión beta inflado con ceros y/o unos de Rigby & Stasinopoulos (2005) se encuentra implementado en R, y que este requiere utilizar diferentes funciones para ajustar el modelo de regresión, dependiendo si los datos son inflados solo con ceros, unos, o ambos, no existe un paquete en R que lo haga en una sola función para diferentes distribuciones y parametrizaciones. En este trabajo se busca implementar mediante el paquete **ZOIP** de R un modelo de regresión que se base en las principales distribuciones para datos proporcionales inflados con ceros y/o unos (Modelo de regresión ZOIP), mediante la distribución ZOIP (Zeros Ones Inflated Proportional), estimando los parámetros vía máxima verosimilitud, en una sola función. Se realiza un estudio de simulación que muestra la convergencia satisfactoria de los parámetros de los diferentes modelos aplicados y se aplica el modelo de regresión ZOIP a datos reales.

***Key words*:** Datos proporcionales inflados, Regresion ZOIP, Máxima verosimilitud, R.

1. Introducción

Los datos obtenidos a partir de variables medidas bajo porcentajes, tasas y proporciones, son llamados datos proporcionales y estos se encuentran ubicados por lo general en el intervalo (0,1), sin embargo existen casos de este tipo de variables pueden dar resultados de cero y/o uno, representando la ausencia o presencia total de la característica medida a partir de una variable, este tipo de datos son conocidos como datos proporcionales inflados con ceros y/o unos, existen diferentes distribuciones que explican estos datos, tales como la distribución beta inflada con

ceros o unos (Ospina & Ferrari 2010) o de una manera más general la distribución ZOIP (Zeros Ones inflated Proportional) que busca reunir la distribución simplex y beta bajo diferentes parametrizaciones en una sola distribución.

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

Sin embargo los anteriores modelos de regresión son realizados para datos proporcionales no inflados con ceros o unos, es por esto que Ospina & Ferrari (2012) realizan un modelo de regresión inflado con cero o con uno, no con ambos, bajo la distribución beta inflada de Ospina & Ferrari (2010) con parametrización Ferrari & Cribari-Neto (2004), de igual manera Rigby & Stasinopoulos (2005) implementan los modelos de regresión beta inflados en ceros y/o unos, y se encuentran implementados en el paquete **gamlss** de R (Rigby & 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 & 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 & Lachos 2014), no existe un paquete en R que permita realizar un ajuste sobre dicho modelo de regresión.

Es por esto que en este trabajo se implementa de manera teórica y de forma práctica mediante el paquete **ZOIP** en el sistema de computación R (R Core Team 2017) y disponible en el repositorio web `GitHub`, un modelo de regresión para datos proporcionales inflados con ceros y/o unos (Modelo de regresión ZOIP) que permita mediante una misma función ajustar modelos en diferentes distribuciones para datos proporcionales y en diferentes parametrizaciones.

El artículo se encuentra organizado de la siguiente manera: primero se presenta la distribución ZOIP, en la siguiente sección el modelo de regresión ZOIP y su debida estimación, en la sección número cuatro se presenta unas aplicaciones a datos simulados y a datos reales y por último se presenta la implementación del modelo de regresión ZOIP en el paquete **ZOIP** de R.

2. Distribución ZOIP (Zero One Inflated Proporcional)

Las distribuciones que se ajustan al comportamiento de datos proporcionales, porcentajes o tasas 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 una distribución para datos proporcionales como común la distribución beta o simplex, es por eso que es desarrollada 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 que es continua y además es una función de densidad de probabilidad para datos proporcionales, para este artículo se tomaran la distribución beta original, es decir, con parámetros $p > 0$ y $q > 0$, la distribución beta parametrización propuesta por Ferrari & Cribari-Neto (2004), la distribución beta con parametrización dada por Rigby & Stasinopoulos (2005) y la distribución simplex presentada por Jørgensen & Barndorff-Nielsen (1991). Si la variable aleatoria y tiene distribución ZOIP con parámetros μ, σ, p_0 y p_1, se denotará como $y \sim ZOIP(\mu, \sigma, p_0, p_1)$, la función de densidad de probabilidad está dado por:

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

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 función de densidad de probabilidad para datos proporcionales. La media y varianza de y, están dadas por

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

$$Var(y) = p_1(1 - p1) + (1 - p_0 - p_1) \left[Var^*(y) + (p_0 + p_1)[E^*(y)]^2 - 2E^*(y) p_1 \right] \tag{3}$$

donde $E^*(y)$ y $Var^*(y)$ es el valor esperado y la varianza de una distribución para datos proporcionales, respectivamente.

Si para la distribución $ZOIP(\mu, \sigma, p_0, p_1)$ se elige la distribución beta con parametrización Ferrari & Cribari-Neto (2004) entonces el parámetro σ tomará el rol del parámetro ϕ donde $\phi > 0$ (Ferrari & Cribari-Neto 2004), si la parametrización es beta original μ y σ tomarán el rol de p y q, respectivamente, donde $p > 0$ y $q > 0$. En las demás distribuciones y parametrizaciones tenidas en cuenta μ y σ tomaran valores entre cero y uno.

En general la distribución ZOIP puede estar inflada bilateralmente (con ceros y unos a la misma vez) o unilateralmente (solo con ceros o unos, no ambos), de esta forma si $p_0 = 0$ entonces tendríamos una distribución OIP (Ones Inflated Proporcional) y si $p_1 = 0$ se tendría una distribución ZIP (Zeros Inflated Proporcional), si $p_0 = p_1 = 0$ se tendrán valores clásicos de las distribuciones proporcionales.

En la figura 1 se muestra una gráfica donde se evidencia el comportamiento de las distribuciones ZOIP-beta en sus diferentes parametrizaciones y ZOIP-simplex, es de aclarar que en las figuras 1a. 1b. 1c. los valores de los parámetros son diferentes pero dan como resultado la misma distribución ZOIP-beta, ya que existen ecuaciones de equivalencias que permiten cambiar de una parametrización de la distribución beta a otra fácilmente. Además se puede observar en la figura 1d. Como la distribución ZOIP-simplex hereda el comportamiento bimodal de la distribución simplex común, con valores mínimo y máximo en cero y uno respectivamente.

3. Modelo de regresión ZOIP

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

$$\begin{aligned} h_1(\mu_i) &= \mathbf{x}_{i1}^\top \boldsymbol{\beta}_1, \\ h_2(\sigma_i) &= \mathbf{x}_{i2}^\top \boldsymbol{\beta}_2, \\ h_3(p_{0i}) &= \mathbf{x}_{i3}^\top \boldsymbol{\beta}_3, \\ h_4(p_{1i}) &= \mathbf{x}_{i4}^\top \boldsymbol{\beta}_4 \end{aligned} \tag{4}$$

donde $\mathbf{x}_{i1} = (x_{i11}, x_{i12}, \ldots, x_{i1k_1})$, $\mathbf{x}_{i2} = (x_{i21}, x_{i22}, \ldots, x_{i2k_2})$, $\mathbf{x}_{i3} = (x_{i31}, x_{i32}, \ldots, x_{i1k_3})$ y $\mathbf{x}_{i4} = (x_{i41}, x_{i42}, \ldots, x_{i1k_4})$, son vectores de covariables conocidos de dimensión k_1, k_2, k_3 y k_4 respectivamente, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1k_1})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2k_2})$, $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \ldots, \beta_{3k_3})$ y $\boldsymbol{\beta}_4 = (\beta_{41}, \beta_{42}, \ldots, \beta_{4k_4})$ son vectores de parámetros de regresión desconocidos. Además se asume que las funciones de enlace $h_1 : (0,1) \to \mathbb{R}$, $h_2 : (0,1) \to \mathbb{R}$, $h_3 : (0,1) \to \mathbb{R}$ y $h_4 : (0,1) \to \mathbb{R}$

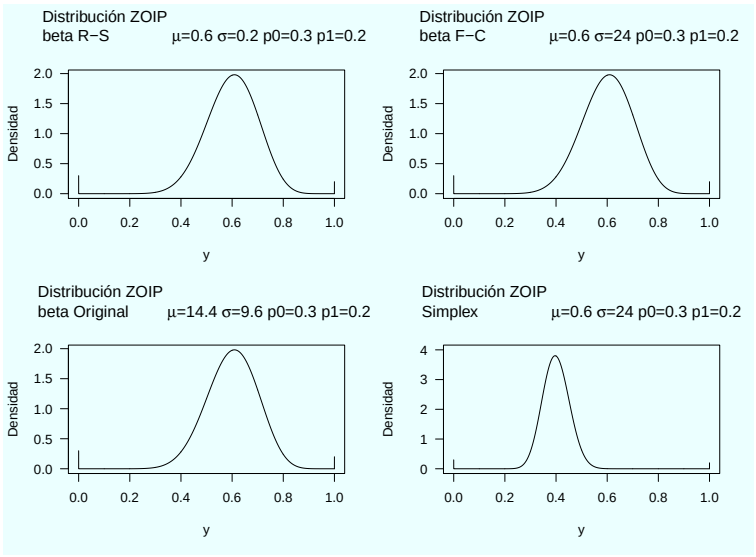

FIGURA 1: Representación de la distribución ZOIP para algunos valores de los parámetros, donde R-S se refiere a Rigby & Stasinopoulos (2008) y F-C es Ferrari & Cribari-Neto (2004)

son estrictamente monótonas y doblemente diferenciables. Las posibles funciones para el parámetro μ y σ son logit, probit, clog-log, o log dependiendo de la parametrización, para los parámetros de inflación p_0 y p_1 son posibles funciones de enlace como logit, probit, clog-log.

3.1. Inferencia estadística

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

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\mathbf{y}_i; \mu_i, \sigma_i, p_{0i}, p_{1i}) \tag{5}$$

donde para el caso de ZOIP-beta original $\mu_i = p_i$, $\sigma_i = q_i$; si la distribución ZOIP-beta fuese con parametrización de Ferrari & 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 (5) al aplicar logaritmo natural se obtiene la función de log verosimilitud definida como:

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

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

$$\ell_1(\boldsymbol{\beta_3}) = \sum_{i=1}^{n} p_{0i}^{S_0(y_i)}(1-p_{0i})^{1-S_0(y_i)}$$
$$\ell_2(\boldsymbol{\beta_4}) = \sum_{i=1}^{n} p_{1i}^{S_1(y_i)}(1-p_{1i})^{1-S_1(y_i)}$$
$$\ell_3(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = \sum_{i=1:y_i\in(0,1)}^{n} f(y_i;\mu_i,\sigma_i)$$

Con

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

Con $p_{0i} = h_3^{-1}(\mathbf{x}_{i3}^\top \boldsymbol{\beta}_3)$, $p_{1i} = h_4^{-1}(\mathbf{x}_{i4}^\top \boldsymbol{\beta}_4)$, $\mu_i = h_1^{-1}(\mathbf{x}_{i1}^\top \boldsymbol{\beta}_1)$ y $\sigma_i = h_2^{-1}(\mathbf{x}_{i2}^\top \boldsymbol{\beta}_2)$ como se definio en (4). La funcion de verosimilitud depende de tres terminos, el primero depende de $\boldsymbol{\beta}_3$ (componente discreto para inflación en cero), el segundo de $\boldsymbol{\beta}_4$ (componente discreto para explicar la inflación en uno) y el tercero depende de $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ (Componentes para explicar la parte continua), por lo tanto los parámetros son separables y la inferencia de máxima verosimilitud para $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ se puede hacer por separado de la de $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$, como si conociera a $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$ y viceversa. (Ospina & Ferrari 2012).

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

4. Paquete ZOIP

En esta sección presentaremos el paquete **ZOIP** realizado en R para ajustar un modelo de regresión ZOIP.

4.1. Instalación

La versión más actualizada del paquete **ZOIP** se encuentra ubicada en `GitHub`, el cual es un alojamiento de repositorios Git, para obtener dicha versión es necesario ejecutar el siguiente código que instala el devtools que es necesario para descargar el paquete **ZOIP**.

```
if (!require('devtools')) install.packages('devtools')
devtools::install_github('jucdiaz/ZOIP', force=TRUE)
require(ZOIP)  # Carga el paquete
```

4.2. Función RM.ZOIP

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

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

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

- `formula.mu`: Formula que define la función de regresión para el parámetro μ, Un valor posible es `y` $\sim$ `x1 + x2`, es necesario definir la variable respuesta (y).
- `formula.sigma`: Formula que define la función de regresión para el parámetro σ, Un valor posible es $\sim$ `x1`. Por defecto $\sim$ `1`.
- `formula.p0`: Formula que define la función de regresión para el parámetro p_0, Un valor posible es $\sim$`x1`. Por defecto $\sim$`1`.
- `formula.p1`: Formula que define la función de regresión para el parámetro p_1, Un valor posible es $\sim$`x1`. Por defecto $\sim$`1`.
- `data`: es el conjunto de datos en formato `data.frame` donde debe contener las nombres de las columnas tal cual como están en las fórmulas.
- `family`: Elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP, si toma el valor de `'R-S'` se utilizara la distribución beta con parametrización Rigby & Stasinopoulos (2005), si toma el valor de `'F-C'` se utilizara la distribución beta parametrización Ferrari & Cribari-Neto (2004), el valor de `'Original'` se utilizara la distribución beta con parametrización original, `'Simplex'` Utilizara la distribución simplex.
- `link`: Es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y formula. Si el modelo de regresión no posee covariables se debe utilizar como función enlace la opción `identity`, independientemente del

valor escogido en familia, opciones posibles son `logit`, `exp`. Por defecto `link=c('identity','identity','identity','identity')`.
- `optimizer`: Elección del optimizador, utilizado para la convergencia de la máxima verosimilitud.

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

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

```
devtools::install_github('jucdiaz/ZOIP', force=TRUE)
library(ZOIP)
n<-1000
x1<-runif(n);x2<-runif(n)

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

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

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

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

param<-cbind(mu_i,sigma_i,p0_i,p1_i)
y_i<-apply(param,1,function(x){rZOIP(1,mu=x[1],sigma=x[2]
                                     ,p0=x[3],p1=x[4]
                                     ,family='R-S')})
data<-as.data.frame(cbind(y_i,x1,x2))

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

Los resultados obtenidos se muestran a continuación.

```
---------------------------------------------------------------
Fixed effects for logit(mu)
---------------------------------------------------------------
            Estimate Std. Error z value  Pr(>|z|)
(intercept)  0.35500    0.09953  3.5668 0.0003614 ***
x1          -1.52936    0.18172 -8.4158 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
---------------------------------------------------------------
Fixed effects for logit(sigma)
---------------------------------------------------------------
            Estimate Std. Error z value  Pr(>|z|)
(intercept)  0.30456    0.10158  2.9983  0.002715 **
x1           2.67959    0.13954 19.2030 < 2.2e-16 ***
x2           1.03339    0.12824  8.0585 7.723e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
---------------------------------------------------------------
Fixed effects for identity(p0)
---------------------------------------------------------------
             Estimate Std. Error z value  Pr(>|z|)
(intercept) 0.0710777  0.0081293  8.7434 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
---------------------------------------------------------------
Fixed effects for logit(p1)
---------------------------------------------------------------
            Estimate Std. Error z value Pr(>|z|)
(intercept) -0.20773    0.14623 -1.4205   0.1555
x2          -3.15715    0.33901 -9.3128   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
---------------------------------------------------------------
---------------------------------------------------------------
```

5. Aplicación

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

5.1. Datos simulados

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

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \epsilon_i, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \\ h_3(p_0) &= \beta_0 + \epsilon_i, \\ h_4(p_1) &= \beta_0 + \beta_1 x_{2i} + \epsilon_i \end{aligned} \tag{6}$$

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

Familia	Parámetro	$h(\cdot)$
R-S [1]	μ	Logit
	σ	Logit
	p_0	NA
	p_1	Logit
F-C [2]	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 1: Funciones de enlace adecuadas para cada parámetro, según su distribución y/o parametrización.

En la figura 2 se describen los valores estimados para diferentes valores de tamaño de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización

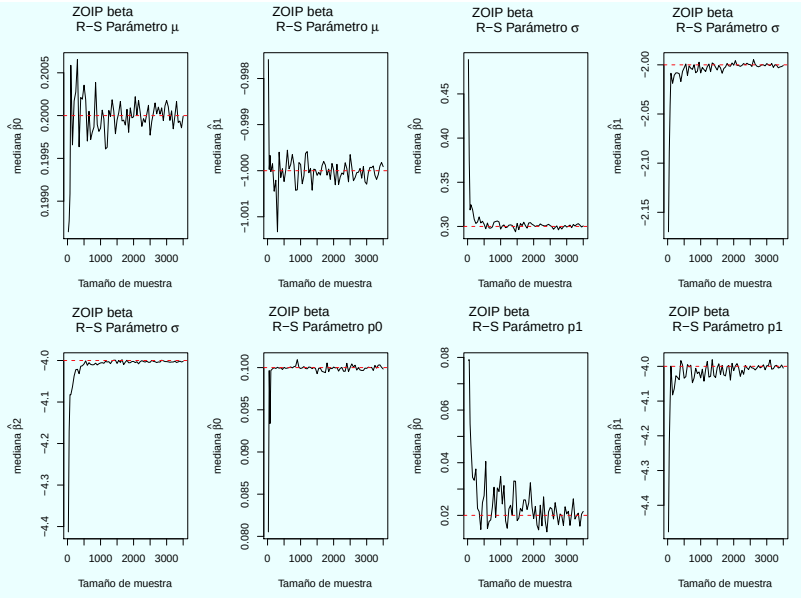

FIGURA 2: Simulación de un modelo de regresión ZOIP-beta para la parametricazion R-S con diferentes valores de n.

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

En la figura 4 se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-beta con parametrización original, se puede ver como con los valores del escenario de simulación elegidos, se obtiene una distribución ZOIP con mayor variabilidad, por lo que los valores de los parámetros asociados a σ tienen una mayor oscilación, sin embargo, este oscila solo en un 0.01 de sus unidades, lo que no es preocupante. Por otra parte, se observa como el parámetro de intercepto del parámetro de inflación de unos (p_1) si oscila mucho más ya que este tiene una desviación estándar de 0.04 en promedio, pero

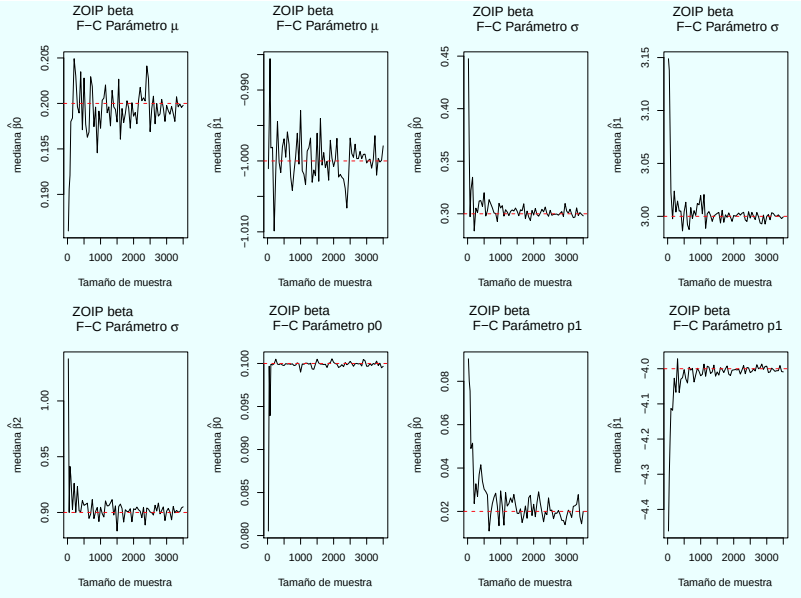

FIGURA 3: Simulación de un modelo de regresión ZOIP-beta para la parametricazion F-C con diferentes valores de n.

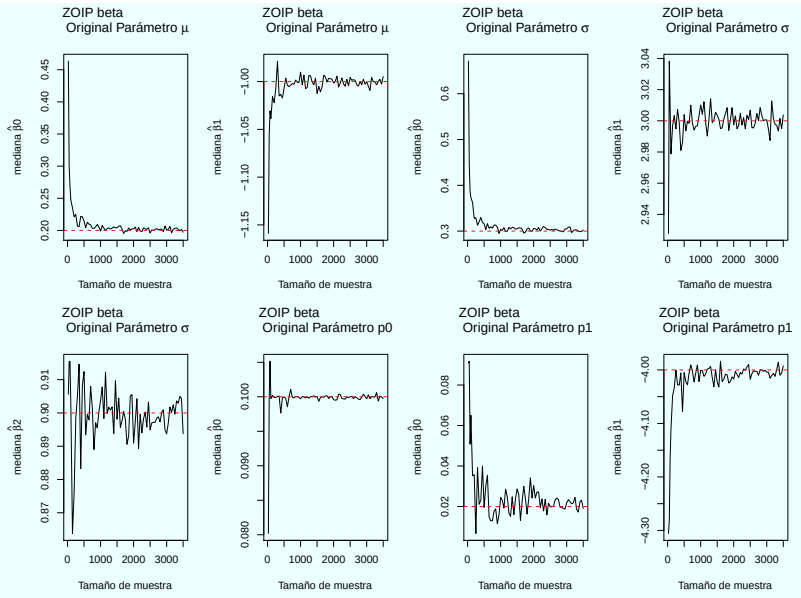

FIGURA 4: Simulación de un modelo de regresión ZOIP-beta para la parametricazion original con diferentes valores de n.

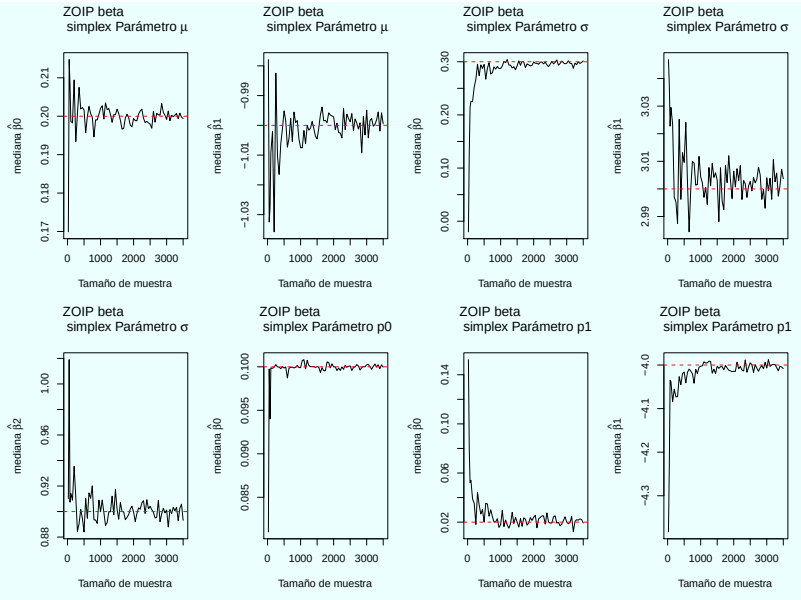

FIGURA 5: Simulación de un modelo de regresión ZOIP-simplex con diferentes valores de n.

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

En la figura 5 se describen los valores estimados para diferentes tamaños de muestra, cuando se elige realizar una regresión ZOIP-simplex, Se nota como todos los parámetros oscilan alrededor de los valores verdaderos y como estas oscilaciones se van reduciendo a través de que el tamaño de muestra crece, sin embargo, unos parámetros toman mayor tiempo de convergencia como es el parámetro β_1 asociado al parámetro de dispersión (σ).

Familia	μ		σ			p_0	p_1	
	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_0}$	$\hat{\beta_0}$	$\hat{\beta_1}$
R-S	1.25	0.32	1.45	2.55	1.38	4.86	383.09	4.88
F-C	14.22	3.96	22.21	2.9	10.14	4.86	91.21	4.88
original	22.34	8.03	22.55	3.62	8.69	4.84	90.58	4.96
simplex	13.93	5.89	24.49	3.11	11.01	4.85	91.15	4.81

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

En la tabla 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

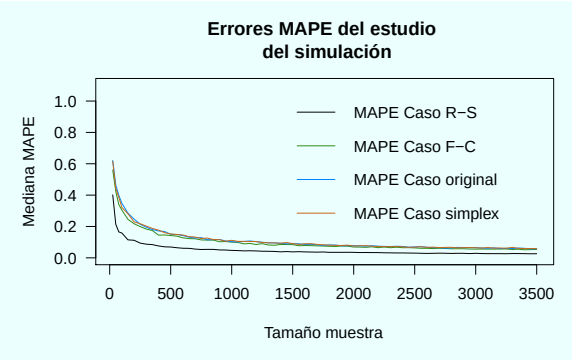

FIGURA 6: Mape (Error porcentual absoluto medio) para modelo de regresion ZOIP simulado para distintas parametrizaciones y valores de n.

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

En la figura 6 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 & 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.

5.2. Datos reales

En una entidad financiera tiene gran importancia conocer el comportamiento del porcentaje de utilización de las tarjetas de crédito (tdc), con el fin de conocer el comportamiento de la cartera de tarjeta de crédito, además de detectar los diferentes factores que pueden afectar este tipo de cartera. Se define a y como el porcentaje de uso de una tdc, claramente y se encuentra entre cero y uno, pero

adicional es normal que se tengan tdc que no son utilizadas ($y = 0$) y tdc que están utilizadas en la totalidad de su cupo asignado ($y = 1$), por lo que se trata a y como una variable aleatoria perteneciente a datos proporcionales inflados con ceros y unos. Se tiene un total de 9206 tarjetas de crédito. Se quiere estudiar el impacto de algunas variables sobre el porcentaje de utilización de una tdc, para ello se busca ajustar un modelo de regresión ZOIP mediante la función `RM.ZOIP` del paquete **ZOIP** de R, que permita explicar el comportamiento del porcentaje de utilización de una tdc mediante bajo 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 sobre el comportamiento de pago de un cliente que pertenece a la entidad financiera, donde cero es el peor valor y mil un comportamiento de pago ideal; *Prom Cuotas:* se define como el promedio al que ha diferido sus compras en los últimos seis meses; *Cupo tdc Entidad:* es el cupo de la tdc asignada, 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 que nos permitan explicar el comportamiento de los cuatro parámetros que se compone 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 (7) 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 se muestran las diferentes funciones enlaces adecuadas para cada parámetro dependiendo de la distribución escogida y/o la parametrización.

$$\begin{aligned} h_1(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \\ h_2(\sigma_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \\ h_3(p_0) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{3i} + \epsilon_i, \\ h_4(p_1) &= \beta_0 + \epsilon_i \end{aligned} \tag{7}$$

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

En las ecuaciones dadas en (7) se puede ver como los parámetros μ, σ serán explicados por las variables score, cupo de su tdc y el promedio al que difiere sus compras, el parámetro p_0 por las variables score y el cupo de su tdc, para observar como ajusta el parámetro p_1 sin covariables, no se pusieron variables regresoras.

En la tabla 4 se tiene información sobre como las variables influyen en el modelo, primero se puede ver que al modelar el parámetro de la media, es decir, el porcentaje medio de utilización de la tdc, la variable regresora score afecta de manera negativa y significativa en cada uno de los modelos ajustados, excepto en el modelo con parametrización original en el cual el parámetro no dio significativo; esto nos indica que a un mejor comportamiento de pago, menos utilización de la

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

Tabla 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.

tdc, sobre el parámetro del promedio de cuotas diferidas vemos como el parámetro β_2 sin excepción alguna en todos los modelos es la variable más significativa que permite explicar la proporción media de la utilización de una tdc, esto nos indica que a medida que los clientes difieren a mayores cuotas sus compras con la tdc estarán utilizando más su tdc, algo muy lógico, desde el punto de vista del problema; ahora el parámetro β_3 asociado al cupo de su tdc tiene un efecto negativo aunque no muy fuerte sobre la variable respuesta excepto en el modelo ZOIP-simplex donde tiene un efecto positivo mucho más fuerte.

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 variabilidad excepto en el modelo ZOIP-beta parametrización Rigby & Stasinopoulos (2005), cabe resaltar que β_1 no es significativo en el modelo simplex. Además el parámetro β_3 asociado al cupo de la tdc no influye sobre la variabilidad en ninguno de los modelos propuestos.

Por otra parte el efecto de que alguien no utilice su tdc es exactamente igual en los cuatro modelos, esto por la metodología de estimación y se puede observar como el cupo de la tdc no influye. Además como la variable score que describe el comportamiento de pago de una persona influye de manera positiva con la no utilización de la tdc.

En el porcentaje de utilización global de la tdc, vemos como sin ninguna covariable el parámetro es significativo, indicándonos como no se ve afectado su estimación del parámetro por la estimación de los parámetros regresores de los demás modelos.

Familia	Parámetro	β's	Estimación	Error estándar	Valor P	Log-Verosimilitud	Iteraciones
R-S	μ	$\hat{\beta_0}$	-0.046	0.050	0.3618	-5732.63	101
		$\hat{\beta_1}$	-0.354	0.107	0.0009		
		$\hat{\beta_2}$	0.022	0.002	$< 2.2e^{-16}$		
		$\hat{\beta_3}$	-0.025	0.009	0.0074		
	σ	$\hat{\beta_0}$	0.822	0.038	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	-0.197	0.078	0.0114		
		$\hat{\beta_2}$	-0.006	0.002	0.0013		
		$\hat{\beta_3}$	-0.003	0.007	0.6741		
	p_0	$\hat{\beta_0}$	-1.834	0.096	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.882	0.178	0.0000007		
		$\hat{\beta_2}$	-0.007	0.015	0.6480		
	p_1	$\hat{\beta_0}$	0.070	0.003	$< 2.2e^{-16}$		
F-C	μ	$\hat{\beta_0}$	-0.045	0.050	0.3667	-5732.50	84
		$\hat{\beta_1}$	-0.354	0.107	0.0009		
		$\hat{\beta_2}$	0.022	0.002	$< 2.2e^{-16}$		
		$\hat{\beta_3}$	-0.025	0.009	0.0074		
	σ	$\hat{\beta_0}$	0.068	0.045	0.1286		
		$\hat{\beta_1}$	0.238	0.094	0.0117		
		$\hat{\beta_2}$	0.007	0.002	0.0012		
		$\hat{\beta_3}$	0.003	0.008	0.6854		
	p_0	$\hat{\beta_0}$	-1.834	0.959	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.882	0.178	0.0000007		
		$\hat{\beta_2}$	-0.007	0.015	0.6481		
	p_1	$\hat{\beta_0}$	0.070	0.003	$< 2.2e^{-16}$		
original	μ	$\hat{\beta_0}$	-0.649	0.048	$< 2.2e^{-16}$	-5733.28	91
		$\hat{\beta_1}$	0.035	0.103	0.7311		
		$\hat{\beta_2}$	0.019	0.002	$< 2.2e^{-16}$		
		$\hat{\beta_3}$	-0.011	0.009	0.2267		
	σ	$\hat{\beta_0}$	-0.611	0.054	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.397	0.111	0.0003		
		$\hat{\beta_2}$	-0.002	0.003	0.4724		
		$\hat{\beta_3}$	0.015	0.010	0.1456		
	p_0	$\hat{\beta_0}$	-1.834	0.096	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.882	0.178	0.0000007		
		$\hat{\beta_2}$	-0.007	0.015	0.6480		
	p_1	$\hat{\beta_0}$	0.070	0.003	$< 2.2e^{-16}$		
simplex	μ	$\hat{\beta_0}$	0.180	0.050	0.0003	-22703.67	123
		$\hat{\beta_1}$	-3.889	0.206	$< 2.2e^{-16}$		
		$\hat{\beta_2}$	0.166	0.004	$< 2.2e^{-16}$		
		$\hat{\beta_3}$	0.128	0.019	$5.84e^{-12}$		
	σ	$\hat{\beta_0}$	11.586	0.062	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.063	0.240	0.7915		
		$\hat{\beta_2}$	0.133	0.004	$< 2.2e^{-16}$		
		$\hat{\beta_3}$	-0.344	0.023	$< 2.2e^{-16}$		
	p_0	$\hat{\beta_0}$	-1.834	0.096	$< 2.2e^{-16}$		
		$\hat{\beta_1}$	0.882	0.178	0.0000007		
		$\hat{\beta_2}$	-0.007	0.015	0.6480		
	p_1	$\hat{\beta_0}$	0.070	0.003	$< 2.2e^{-16}$		

Tabla 4: Parametros regresores estimados de un modelo de regresión ZOIP en el porcentaje de utilización de una tdc.

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

6. Conclusión

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

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

Referencias

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

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

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

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

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

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

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

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

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

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

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

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

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