Paquete ZOIP de R para modelo de regresión mixto con datos proporcionales inflados con ceros y/o unos

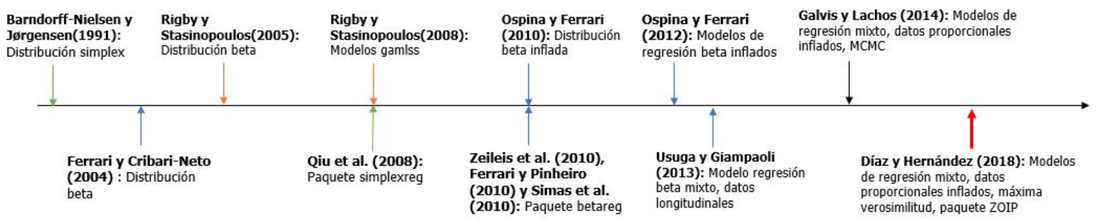
Juan Camilo Díaz Zapata
twitter: @jkmilodiaz
Freddy Hernández Barajas
twitter: @fhernanb74
Escuela de Estadística
Universidad Nacional de Colombia, Medellín

Introducción

Los modelos de regresión mixto para datos proporcionales inflados con ceros y/o unos, son útiles para determinar el comportamiento de una variable proporcional a partir de variables consideradas como efectos fijos y aleatorios.

La integración de los modelos estadísticos en un paquete del sistema computacional *R* permite que la utilización de los modelos estadísticos a problemas aplicados sea de gran facilidad para la comunidad estadística.

Estado del arte



Distribución ZOIP (Zeros Ones Inflated Proportional)

Si la variable aleatoria Y tiene distribución ZOIP con parámetros μ, σ, p_0 y p_1, se denotará como $Y \sim$ ZOIP(μ, σ, p_0, p_1), la función de densidad de probabilidad está dada por:

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

donde $p_0 \geq 0$ y $p_1 \geq 0$, además $f(y; \mu, \sigma)$ representa alguna de las funciones de densidad de probabilidad para datos proporcionales

Modelo de regresión ZOIP con efectos mixtos

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

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

con $i = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, n_i$.

Estimación vía máxima verosimilitud y AGHQ

El vector de parámetros para el modelo (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 de esa forma la funcion de log-verosimilitud $\ell(\boldsymbol{\theta})$ esta dado por:

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

Para encontrar el punto máximo de la función de log-verosimilitud anterior, es necesario realizar una aproximación por medio de la cuadratura de Gauss-Hermite adaptativa con o sin *pruning*

Paquete ZOIP de R

El paquete *ZOIP* de *R* es útil para ajustar distribución ZOIP, modelos de regresión de efectos fijos y mixtos para datos proporcionales inflados con ceros y/o unos. Estas son las principales funciones:

```
dZOIP  pZOIP  qZOIP  rZOIP  RM.ZOIP  RMM.ZOIP
```

Ejemplo:

```
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 = base,
        link = link, family = "R-S")

summary(mod)
```

```
link <- c("logit", "logit", "identity", "identity")
mod<-RMM.ZOIP(formula.mu = Y ~ log(Days), formula.sigma = ~log(Days),
         formula.p0 = ~1, formula.p1 = ~1, data = base,
         formula.random = ~1 | subject, link = link, family = "R-S")

summary(mod)
```

Aplicación modelo de regresión ZOIP mixto

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*, para la variable respuesta **porcentaje de utilización de una tarjeta de crédito (tdc)**.

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

con $\gamma_{1i} \sim N(0, \lambda_1^2)$ y $\gamma_{2i} \sim N(0, \lambda_2^2)$.

y_{ij} es el porcentaje de utilización de la j-ésima tdc perteneciente a la i-ésima ciudad, $i = 1, 2, \ldots, 10$ y $j = 1, 2, \ldots, 15$.

x_{1ij}: es el valor del tiempo en mora en meses de la j-ésima tdc asociada a la i-ésima ciudad.

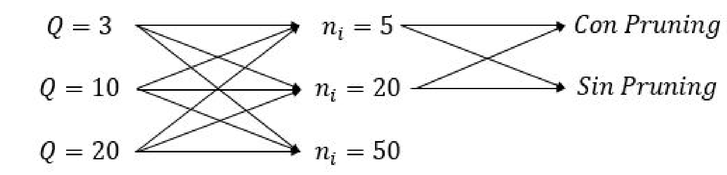
El modelo propuesto se puede reescribir con los parámetros estimados por medio del paquete `ZOIP` y la función `RMM.ZOIP`, así:

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

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

Estudio de simulación: Modelo regresión ZOIP mixto

El modelo de regresión ZOIP mixto en el cual se basa el estudio de simulación esta dado por (2) y el esquema de simulacion es:



Esquema de simulación.

En el estudio de simulación se plantearon 18 escenarios de simulación y se realizaron 1000 réplicas en cada escenario, dando como resultado:

		Con *pruning*			Sin *pruning*		
Parámetro	β's	$n_i = 5$	$n_i = 20$	$n_i = 50$	$n_i = 5$	$n_i = 20$	$n_i = 50$
μ	$\beta_{10} = -1.13$	-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
Med tiempo(Seg)		115.72	130.85	140.58	61.69	163.36	218.68
Med num. iter		22	30	34	22	30	34

Table 1: *Resultados del estudio de simulación*

Conclusiones

1. El paquete *ZOIP* se implementan diferentes funciones, con el objetivo de que sean utilizados los modelos de regresión ZOIP por un usuario. Dicho paquete muestra convergencia de sus funciones desarrollados, por medio de los estudios de simulación.
2. 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.
3. 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%.

References

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

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

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

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

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

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