Universidad Nacional de Colombia
sede medellín

MODELO DE REGRESIÓN MIXTO PARA DATOS PROPORCIONALES INFLADOS CON CEROS Y/O UNOS

Tesis

para obtener el título de

magister en ciencias estadística

presenta

juan camilo díaz zapata

Director de Tesis: freddy hernández barajas, Ph.D

Medellín. 2017

Índice

1. Resumen ejecutivo

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, tal como la distribución beta o la distribución simplex. La variable respuesta resulta de porcentajes, tasas o proporciones donde estas pueden llegar a tomar valores de cero o uno, debido a la ausencia o presencia total de una característica y es por esto que se dice que son modelos inflados con ceros y/o unos. La estimación de los parámetros de este tipo de modelos vía máxima verosimilitud no tiene una solución cerrada analíticamente, por lo que es necesario recurrir a metodologías computacionales que puedan llegar a una solución, es por esto que se quiere realizar este trabajo de investigación para estimar los parámetros del modelo vía máxima verosimilitud usando algunas variaciones de la cuadratura de Gauss-Hermite.

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

Abstract

The mixed model regression to data proportional inflated with zeros and/or ones, it is a regression model where the response can be adequately modeled by a data for a proportional distribution such as beta distribution or simplex distribution. The response comes from percents, rates and proportions, these ones can take one or zero in accordance with the absence or presence of a feature and that is why they are called inflated models with zeros and/or ones. The estimation of the parameters by mean of maximun likelihood can not be solved analitycally so it is necessary to use computational methodologies that can reach a solution, that is why we want to produce this research to estimate by means of maximum likelihood parameters of the model using some variation of the quadrature Gauss-Hermite.

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

2. Descripción del proyecto

2.1. Planteamiento del problema

En el análisis de datos es usual que se quiera estudiar la relación entre una variable de interés y un conjunto de variables explicativas, la forma usual de abordar el problema es por medio del modelo de regresión lineal múltiple, sin embargo, cuando la variable de interés se encuentra en un intervalo entre cero y uno el modelo de regresión lineal múltiple no resulta apropiado ya que no cumple el supuesto de variable respuesta con distribución normal. Los datos proporcionales se encuentran en el intervalo (0,1) y son el resultado de porcentajes, tasas, proporciones y fracciones que son comúnmente evidenciados en diversos campos tanto sociales, económicos y científicos. Por todo lo anterior existen los modelos de regresión para datos proporcionales, cuando la variable de interés sigue una distribución de datos proporcionales, tal como la distribución beta, simplex, entre otros, con los cuales se obtienen modelos de regresión beta y simplex respectivamente.

En los modelos de regresión un efecto fijo se considera como una covariable en el que su influencia sobre la variable respuesta se mide a través de distintos parámetros del modelo, además los valores que pueden tomar esta covariables son informativos y están fijados previamente. Por el contrario una covariable se considera como un efecto aleatorio cuando los valores de la covariable son solamente identificativos y podrían encontrase otros valores si el estudio se repitiera en diferentes circunstancias. Un efecto aleatorio se considera como una variable aleatoria que no se observa pero cuya distribución puede estimarse mediante la varianza de cierta distribución, p.e la distribución normal (Seoane, 2014). Un ejemplo claro de considerar una covariable como un efecto aleatorio es cuando esta es de tipo categórico, posee un número elevado de niveles y estos niveles pueden cambiar fácilmente. La inclusión de una o más covariables como un efecto aleatorio y otras covariables como efectos fijos sobre el modelo regresión para datos proporcionales, es lo que se llama modelo de regresión mixto para datos proporcionales, pero en este tipo de modelos es común que haga falta la inclusión de los valores extremos, es decir, frecuentemente al trabajar con datos proporcionales se presenten tasas o porcentajes con valores en cero o uno y los modelos descritos anteriormente no tienen en cuenta este tipo de casos, es por esto que se crearon los modelos de regresión mixto para datos proporcionales inflados con ceros y/o unos, este tipo de modelos contienen efectos aleatorios que a la hora de estimar los parámetros del modelo por medio de máxima verosimilitud causan un problema, debido a que la función de verosimilitud no tiene una solución cerrada analíticamente.

En trabajos previos para la estimación de los parámetros del modelo de regresión mixto para datos proporcionales sin presencia de ceros y unos se han utilizado dis-

tintas técnicas, tales como el enfoque bayesiano por medio de MCMC y la máxima verosimilitud vía cuadratura de Gauss-Hermite, sin embargo no se ha realizado una adaptación de esta última metodología para la solución del problema de estimación de los parámetros del modelo de regresión mixto para datos proporcionales con presencia de ceros y/o unos, en particular cuando hay intercepto aleatorio, es por ello que se quiere realizar el siguiente trabajo de investigación.

2.2. Objetivos

2.2.1. Objetivo general

Proponer una metodología para estimación de los parámetros de un modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos.

2.2.2. Objetivos específicos

1. Flexibilizar el modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos a través del uso de diferentes distribuciones para datos proporcionales.
2. Determinar entre las diferentes modificaciones de la cuadratura de Gauss-Hermite adaptativa la que mejor desempeño tiene para estimar los parámetros del modelo de regresión mixto para datos proporcionales entre cero y/o uno.
3. Crear un paquete en el software estadístico R que permita la estimación de los parámetros del modelo de regresión mixto para datos proporcionales inflados con ceros y/o unos.

2.3. Marco Teórico

A continuación se presentan algunos conceptos importantes para el desarrollo de este proyecto relacionados con los modelos de regresión mixto para datos proporcionales y la cuadratura de Gauss-Hermite.

2.3.1. Distribuciones para datos proporcionales.

A continuación se presentan las distribuciones de probabilidad para datos proporcionales beta y simplex.

Beta original

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

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

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

$$E(y) = \frac{p}{p+q} \tag{1}$$

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

Beta parametrización Ferrari y Cribari-Neto

Ferrari and Cribari-Neto (2004) propusieron otra parametrización para la distribución beta en función de los parámetos μ y ϕ donde μ corresponde a la media de la distribución. Si $y \sim Be(\mu, \phi)$ la función de densidad de probabilidad de la distribución está dada por:

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

donde $0 < \mu < 1$ y $\phi > 0$, el parámetro ϕ es interpretado como un parámetro de precisión. Cuando $\mu = 1/2$ la distribución es simétrica y si $\mu \neq 1/2$ es asimétrica, además cuando $\mu = 1/2$ y $\phi = 2$ se convierte en la distribución uniforme y para valores más grandes de ϕ la varianza de μ es más pequeña. El valor esperado y la varianza de y están dados por:

$$E(y) = \mu \tag{3}$$

$$Var(y) = \frac{\mu(1-\mu)}{1+\phi} \tag{4}$$

Beta parametrización Rigby y Stasinopoulos

Rigby and Stasinopoulos (2008) propusieron la función de densidad de probabilidad de la distribución beta con parámetros μ y σ dada por:

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

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 $0 < \sigma < 1$ este último puede ser interpretado como un parámetro de dispersión de y. La media y la varianza de y están dadas por:

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

$$Var(y) = \sigma^2\mu(1-\mu) \tag{6}$$

Distribución Simplex Barndorff-Nielsen and Jørgensen (1991)

la distribución simplex que es introducida por Barndorff-Nielsen and Jørgensen (1991) es un caso particular de los modelos de dispersión propuestos por Jørgensen (1997), cuya función de densidad de probabilidad depende de los parámetros μ y Σ^2 denotado por $S(\mu,\sigma^2)$ y está dado por:

$$f(y;\mu,\sigma^2) = \left\{2\pi\sigma^2[y(1-y)]^3\right\}^{-1/2} exp\left\{-\frac{(y-\mu)^2}{2\sigma^2 y(1-y)\mu^2(1-\mu)^2}\right\} \quad ; \quad 0 < y < 1$$

donde $0 < y < 1$, $0 < \mu < 1$ y $\sigma^2 > 0$. además el valor esperado es dado por:

$$E(y) = \mu \tag{7}$$

$$Var(y) = \mu^3(1-\mu)^3 \tag{8}$$

2.3.2. Distribución para datos proporcionales inflada con ceros y unos

La distribución para datos proporcionales en cero y/o uno se compone de la mezcla de dos distribuciones, una parte dada por una distribución degenerada en cero o uno y la otra por la densidad de una distribución para datos proporcionales, como las presentadas anteriormente, se dice que $y \sim ZOIP(\mu,\sigma,p_0,p_1)$ si sigue la función de desnsidad de probabilidad 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 $p_0 \geq 0$ representa la probabilidad que $y = 0$, $p_1 \geq 0$ representa la probabilidad de que $y = 1$, $0 \leq p_0 + p_1 \leq 1$ y $f(y; \mu, \phi)$ representa algunas de las funciones de densidad de probabilidad para datos proporcionales, como las descritas en la sección anterior. La media y varianza de y, están dados por

$$E(y) = p_1 + (1 - p_0 - p_1)E^*(y)$$

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

donde $E^*(y)$ es el valor esperado de una distribución para datos proporcionales como las descritas en (1), (3), (5) y (7). Además la $Var^*(y)$ es la varianza de una distribución para datos proporcionales como se muestra en (2), (4), (6) y (8).

2.3.3. Modelo de regresión mixto para datos proporcionales inflados en cero y/o uno

Una escritura jerárquica de dos niveles considerada para un modelo con variable respuesta dada por la distribución para datos proporcionales inflados con ceros y/o unos, vista en la sección anterior. Denotando a y_{ij} como la $j-$ésima medida del $i-$ésimo grupo, además si asumisos interceptos aleatorios γ_{i1} y γ_{i2}, los cuales son independientes y cada uno sigue una distribución normal con media cero y desviación estándar λ_1 y λ_2, respectivamente. Asumimos también que los interceptos aleatorios γ_{i1} y γ_{i2} son independientes entre sí. Una escritura matemática para el modelo es el siguiente:

$$\begin{aligned} y_{ij}|\gamma_{i1}, \gamma_{i2} &\overset{\text{ind}}{\sim} ZOIP(\mu_{ij}, \sigma_{ij}, p_0, p_1), \\ \gamma_{i1} &\overset{\text{i.i.d}}{\sim} N(0, \lambda_1^2), \\ \gamma_{i2} &\overset{\text{i.i.d}}{\sim} N(0, \lambda_2^2), \end{aligned}$$

Los parámetros μ, σ, p_0 y p_1 son modelados linealmente en función de un conjunto de covariables respectivamente, por:

$$\begin{aligned} g_1(\mu_{ij}) &= \mathbf{x}_{ij1}^\top \boldsymbol{\beta}_1 + \gamma_{i1}, \\ g_2(\sigma_{ij}) &= \mathbf{x}_{ij2}^\top \boldsymbol{\beta}_2 + \gamma_{i2}, \\ g_3(p_0) &= \mathbf{x}_{ij3}^\top \boldsymbol{\beta}_3, \end{aligned}$$

$$g_4(p_1) = \mathbf{x}_{ij4}^{\top}\boldsymbol{\beta}_4$$

donde $\mathbf{x}_{ij1}$, $\mathbf{x}_{ij2}$, $\mathbf{x}_{ij3}$ y $\mathbf{x}_{ij4}$, son vectores de covariables conocidos de dimensión k_1, k_2, k_3 y k_4 respectivamente. $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ y $\boldsymbol{\beta}_4$ son vectores de parámetros desconocidos fijos asociados a las covariables y γ_{i1}, γ_{i2} son los interceptos aleatorios asociados al $i-$ésimo grupo. Además las funciones $g_1(\cdot)$, $g_2(\cdot)$, $g_3(\cdot)$ y $g_4(\cdot)$ son funciones de enlace conocidas y apropiadas para mapear de los reales a los valores admisibles del parámetro (log, logit, probit, clog-log). ver más en (Usuga and Giampaoli, 2013).

2.3.4. Función de verosimilitud

La estimación de los parámetros del modelo de regresión con intercepto aleatorio para datos proporcionales inflados con ceros y/o unos, por medio de máxima verosimilitud, es necesario hallar la función de verosimilitud.

Considere $\boldsymbol{\theta} = (\boldsymbol{\beta_1}^{\top}, \boldsymbol{\beta_2}^{\top}, \boldsymbol{\beta_3}^{\top}, \boldsymbol{\beta_4}^{\top}, \lambda_1, \lambda_2)^{\top}$ un vector de parámetros a ser estimado en el espacio:

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

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

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

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

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

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

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

donde $f(y_{ij}|\gamma_{i1}, \gamma_{i2}; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4})$ es la función de densidad de probabilidad condicional de y_{ij} que distribuye $ZOIP(\mu, \sigma, p_0, p_1)$. γ_{i1}, γ_{i2} y $f(\gamma_{i1}; \lambda_1)$ y $f(\gamma_{i2}; \lambda_2)$ son funciones de densidades de probabilidad normales de γ_{i1} y γ_{i2}, respectivamente. Encontrar el máximo de la función $log(L(\boldsymbol{\theta}))$ no tiene una solución cerrada analíticamente, por lo que es necesario utilizar técnicas computacionales para la solución de esta, tales como la cuadratura de Gauss-Hermite.

2.3.5. Cuadratura de Gauss-Hermite

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

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

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

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

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

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

Cuadratura de Gauss-Hermite adaptativa

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

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

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

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

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

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

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

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

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

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

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

Se realiza una explicación más a detalle en Hernández et al. (2013).

Cuadratura de Gauss-Hermite adaptativa con Pruning

La cuadratura de Gauss-Hermite adaptativa con pruning consiste en eliminar puntos de la cuadratura, tales que el peso a_i asociado al punto es menor que un valor de referencia dado por θ, que está dado por: (Ver más en (Hernández et al., 2013).)

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

Reducción secuencial

El método de reducción secuencial tiene como objetivo aproximar la función de verosimilitud de los modelos de regresión mixto. El método consiste en combinar métodos de aproximación sobre funciones de almacenamiento, a marginales aproximadas de la distribución con densidades proporcionales a $g(\cdot|y,\phi)$, y así poder aproximar la verosimilitud $L(\phi) = \int_{\Re^n} g(u|y,\phi)du$. Donde u es el vector $u = (u_1, \ldots, u_n)$ de tamaño n, que corresponde a la cantidad de efectos aleatorios dentro del modelo. (Ogden, 2015).

Se tomara un enfoque iterativo que consiste en primero integrar solamente u_1 para encontrar la densidad marginal no normalizada de $u_2, \ldots, u_n$, es decir se escribirá la densidad marginal posterior de $u_2, \ldots, u_n$ como un producto sobre el subconjunto de vértices (puntos) máximos asociados al conjunto total de vértices $\wp$ que a su vez se asocian a $u_1, \ldots, u_n$. Una vez hecho esto, el proceso puede repetirse n veces hasta encontrar a la verosimilitud. Se escribirá a $\wp_i$ para la dependencia posterior de los vértices de $u_i, \ldots, u_n$, por lo que la dependencia posterior de los vértices será $\wp_1 = \wp$. Se escribirá $M_i = M(\wp_i)$ para el conjunto de vértices máximos de $\wp_i$.

La factorización de $g(\cdot|y,\phi)$ sobre el conjunto de vértices máximos de $\wp_1$ es dado por:

$$g(u|y,\phi) = \prod_{C \in M_1} g_C^1(u_C),$$

Un algoritmo que describe como encontrar la función de verosimilitud por el método de reducción secuencial es el siguiente:

1. Escoger el orden de la selección de integración de los u_i. Debido a que el orden de estimación puede afectar la aproximación de la verosimilitud.
2. Factorizar $g(u|y,\phi)$ sobre el conjunto de vértices máximos M_1 de la dependencia posterior de los vértices, como $g(u|y,\phi) = \prod_{C \in M_1} g_C^1(u_C)$.
3. Una vez $u_1, \ldots, u_{i-1}$ ha sido integrado utilizando algún un método de aproximación, P.E la cuadratura de Gauss-Hermite. Tendremos entonces la factorización $\tilde{g}(u_i, \ldots, u_n|y,\theta) = \prod_{C \in M_i} g_C^i(u_C)$, de la posterior no normalizada para $u_i, \ldots, u_n$. Escribir

   $$g_{N_i} = \prod_{C \in M_i : C \subset N_i} g_C^i(u_C).$$

   Integrando sobre u_i utilizando alguna cuadratura (p.e cuadratura de Gauss-Hermite adaptativa), y una representación del almacenamiento aproximada es $\tilde{g}_{N_i \setminus i}(\cdot)$ de la función resultante $g_{N_i \setminus i}(\cdot)$. La expresión '$C \subset N_i$' significa el conjunto de vecinos de vértices 1 en $\wp$

4. Escribir
$$\tilde{g}(u_{i+1},\ldots,u_n|y,\phi) = g_{N_i \setminus i}(u_{N_i \setminus i}) \prod_{C \in M_i : C \nsubseteq N_i} g_C^i(u_C),$$
Definiendo una factorización (aproximada) de la densidad posterior de $u_{i+1},\ldots,u_n$ sobre los máximos valores de los vértices M_{i+1} de la nueva dependencia posterior de $\wp_{i+1}$.

5. Repita los pasos (3) y (4) para $i = 1,\ldots,n-1$, entonces la integral $\tilde{g}(u_n|y,\phi)$ sobre u_n es la función de verosimilitud aproximada.

Se puede ver más a detalle en Ogden (2015).

3. Estado del arte

En muchas ocasiones un investigador se enfrenta a casos de modelación donde la variable de interés está dada por un resultado de una proporción, un porcentaje o una tasa, que por lo general se encuentra en el intervalo (0,1). La distribución más utilizado en la literatura para caracterizar este tipo de variables es la distribución beta que (Ferrari and Cribari-Neto, 2004) utilizan para crear el modelo de regresión beta, sin embargo la utilizan bajo una reparametrización en términos de la media y un parámetro de precisión, Rigby and Stasinopoulos (2005) proponen los modelos gamlss y dentro de ellos desarrollan la distribución beta en términos de la media y la desviación estándar; dando así tres diferentes parametrizaciones de la distribución beta (original y parametrizaciones de Ferrari y Cribari-Neto, y Rigby y Stasinopoulos). Sin embargo, no solo existen distribuciones para datos proporcionales basados en la distribución beta, la distribución simplex es un caso especial de los modelos de dispersión propuestos por Jørgensen (1997), el cual fue derivado por Barndorff-Nielsen and Jørgensen (1991) y el modelo de regresión simple fue propuesto por Qiu et al. (2008) y esta implementado en el software R con el paquete *simplexreg*. Otras distribuciones que se pueden usar son la distribución Logitsep (Hossain, 2015) que se compone de cuatro parámetros (media, dispersión, escala y forma) y Beta rectangular (Hahn, 2008) el cual es una mezcla de dos distribuciones y puede ser parámetrizado en terminos de la media.

Por esto se han desarrollado los distintos modelos de regresión para datos proporcionales, el modelo más estudiado ha sido el modelo de regresión beta, donde (Paolino, 2001) estima mediante máxima verosimilitud modelos de variables dependientes de una distribución beta, Kieschnick and McCullough (2003) realiza un análisis de regresión para proporciones, porcentajes y fracciones entre cero y uno, además Ferrari and Cribari-Neto (2004) proponen el modelo de regresión beta para tasas y proporciones, más adelante Zeileis et al. (2010) implementan un paquete

de dicha regresión en el software R, también Ferrari and Pinheiro (2010) y Simas et al. (2010) realizan mejoras de los estimadores de los parámetros del modelo de regresión beta *betareg*. Además, otros autores como Smithson and Verkuilen (2006), Venezuela (2008) realizan diferentes estudios sobre los modelos de regresión para datos proporcionales principalmente para variables respuestas provenientes de la distribución beta. Por otro lado se han realizado diversas aplicaciones sobre datos proporcionales por ejemplo Cook et al. (2008) realizan una aplicación en finanzas de los modelos de regresión para datos proporcionales, con los modelos descritos.

Los modelos mixtos han sido de mucho interés en la última época, por su capacidad de estimar el efecto de una variable sobre el modelo a través de la estimación de la varianza de la distribución de la variable, estos fueron introducidos de una manera general por Laird and Ware (1982). Rigby and Stasinopoulos (2005) en los modelos aditivos generalizados para localización, escala y forma (gamlss) implementan el modelo de regresión beta con intercepto aleatorio normal. Verkuilen and Smithson (2012) y Bonat et al. (2012) proponen modelos de regresión beta con efectos aleatorios normales, estimando los parámetros por medio de máxima verosimilitud marginal y metodologías bayesianas.(Figueroa-Zuñiga et al., 2012) extienden el modelo propuesto por Ferrari and 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 de la media como del parámetro de precisión, además optan también por estimar los parámetros bajo una perspectiva bayesiana. Usuga and Giampaoli (2013) desarrollan el modelo de regresión beta mixto para datos proporcionales longitudinales, estimando los parámetros por medio de máxima verosimilitud y la cuadratura de Gauss-Hermite. Otros autores como Song et al. (2004) implementan un modelo de regresión mixto para una variable respuesta bajo la distribución simplex y Bonat et al. (2015) tambien realiza un análisis de verosimilitud del modelo beta mixto.

La metodología de estimación de parámetros propuesta por Usuga and Giampaoli (2013) bajo la cuadratura de Gauss-Hermite fue implementada anteriormente por Fahrmeir and Tutz (2001) sobre los modelos lineales generalizados. Diversos estudios para la estimación de parámetros sobre modelos estadísticos mediante esta técnica han sido implementados, por ejemplo el trabajo realizado por Liu and Pierce (1994) y Smithson and 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, la cuadratura de Gauss-Hermite adaptativa y algunas mejoras sobre esta como la cuadratura de Gauss-Hermite adaptativa con pruning (Hernández et al., 2013), y la metodología reducción secuencial propuesta por Ogden (2015).

Los modelos que se han nombrado anteriormente no tienen en cuenta que los datos proporcionales pueden llegar a tomar valores de ceros y/o unos, es por eso que al-

gunos autores como Ospina and Ferrari (2010) presentan una distribución de 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 continua para la parte de los valores continuos entre cero y uno. Esta distribución dio pie para que así mismo Ospina and Ferrari (2012) propusieran una clase general de modelos de regresión beta inflados en cero o uno. Además Kosmidis et al. (2015) han estudiado dichos modelos inflados recientemente, pero con una distribución distinta a la presentada por Ospina and Ferrari (2012). Los anteriores autores han realizado todos sus estudios sobre los modelos de regresión considerando solo efectos fijos para datos proporcionales inflados con ceros y/o unos, por eso otros autores como Galvis and Lachos (2014) incluyen efectos aleatorios dentro del modelo de regresión inflado con ceros y/o unos, teniendo en cuenta además diversas distribuciones tales como beta, simplex y beta-rectangular, la estimación de los parámetros del modelo es realizada por medio de MCMC.

En el trabajo se implementara el modelo de regresión mixto para datos proporcionales inflados en cero y/o uno, en particular cuando hay intercepto aleatorio bajo las distintas distribuciones para datos proporcionales, tales como la distribución beta y Simplex; con una estimación de los parámetros por medio de máxima verosimilitud y las distintas variaciones de la cuadratura de Gauss-Hermite adaptativa mencionadas en la sección 2.3.5.

4. Metodología propuesta

La metodología que se llevara a cabo para la elaboración del trabajo es la siguiente:

Inicialmente se definirán los conceptos fundamentales del modelo de regresión mixto para datos proporcionales inflados en cero y/o uno, que es en realidad la combinación de dos distribuciones una discreta y una continua, donde la parte continua podrá ser caracterizada por alguna distribución de probabilidad para datos proporcionales. Luego se derivara analíticamente la función de verosimilitud para el modelo de regresión planteado.

Posteriormente se implementara las distintas variaciones de la cuadratura de Gauss-Hermite para encontrar los parámetros del modelo vía máxima verosimilitud y se realizaran aplicaciones y estudios de simulación para evaluar el desempeño del proceso de estimación de los parámetros.

Se creara un paquete en el software estadístico R, que permita estimar un modelo de regresión mixto para datos proporcionales inflados en cero y/o uno mediante máxima verosimilitud.

5. Cronograma

Actividad	**Meses**									
	1	2	3	4	5	6	7	8	9	10
Revisión bibliográfica	X	X	X	X	X	X	X			
Desarrollo analítico		X	X	X	X	X				
Simulaciones y comparaciones		X	X	X	X	X	X			
Aplicaciones			X	X	X	X	X			
Paquete en R				X	X	X	X	X		
Redacción del articulo y documento de trabajo de grado escrito				X	X	X	X	X		
Correcciones									X	X

6. Compromisos

- Presentación de los resultados parciales del proyecto en un evento de carácter nacional.
- Sometimiento de un articulo con los resultados relevantes de la metodología propuesta a una revista de circulación nacional o internacional.
- Creación de un paquete en R con la metodología propuesta.

Referencias

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

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

W. Bonat, P. Ribeiro, and 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.

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

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

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

S. Ferrari and 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, and S. Ferrari. Mixed beta regression: a bayesian perspective. 2012.

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

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

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

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

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

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

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

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

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

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

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

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

J. Pinheiro and 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, and M. Tan. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, 35(4):577–596, 2008.

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

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

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

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

M. Smithson and 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, and M. Tan. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, 46(5):540–533, 2004.

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

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

J. Verkuilen and 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, and B. Grün. Beta regression in r. *Journal of statistical software*, 34(2):1–24, 2010.