PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

"A beta inflated mean regression model with mixed effects for fractional response variables"

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Renzo Fernández Villegas

Asesor: Cristian Bayes

Miembros del jurado:
Dr. Nombre completo jurado 1
Dr. Nombre completo jurado 2
Dr. Nombre completo jurado 3

Lima, Febrero 2017

Dedicatoria

A mis padres por su constante apoyo en todas las etapas de mi vida.

Agradecimentos

Por incluir.

Resumen

En este artículo proponemos un nuevo modelo de regresión con efectos mixtos para variables acotadas fraccionarias. Este modelo nos permite incorporar covariables directamente al valor esperado, de manera que podemos cuantficar exactamente la influencia de estas covariables en la media de la variable de interés en vez de en la media condicional. La estimación se llevó a cabo desde una perspectiva bayesiana y debido a la complejidad de la distribución aumentada a posteriori usamos un algoritmo de Monte Carlo Hamiltoniano, el muestreador No-U-Turn, que se encuentra implementado en el software Stan. Se realizó un estudio de simulación que compara, en términos de sesgo y RMSE, el modelo propuesto con otros modelos tradicionales longitudinales para variables acotadas, resultando que el primero tiene un mejor desempeño. Finalmente, aplicamos nuestro modelo de regresión Beta Inflacionada con efectos mixtos a datos reales los cuales consistían en información de la utilización de las líneas de crédito en el sistema financiero peruano.
Palabras-clave: proporciones, variables fraccionarias, distribución Beta Inflacionada, inferencia bayesiana, métodos MCMC, Monte Carlo Hamiltoniano, modelos mixtos, RStan.

Abstract

In this article we propose a new mixed effects regression model for fractional bounded response variables. Our model allows us to incorporate covariates directly to the expected value, so we can quantify exactly the influence of these covariates in the mean of the variable of interest rather than to the conditional mean. Estimation is carried out from a bayesian perspective and due to the complexity of the augmented posterior distribution we use a Hamiltonian Monte Carlo algorithm, the No-U-Turn sampler, implemented using Stan software. A simulation study for comparison, in terms of bias and RMSE, was performed showing that our model has a better performance than other traditional longitudinal models for bounded variables. Finally, we applied our Beta Inflated mixed-effects regression model to real data which consists of utilization of credit lines in the peruvian financial system.

Keywords: proportions, fractional variables, Beta Inflated distribution, bayesian inference, MCMC methods, Hamiltonian Monte Carlo, mixed models, RStan.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminary Considerations

Multiple researches from different fields are interested in studying and measuring the influence of a set of covariates upon a proportion. There exists a great number of regression models for proportions in the literature, for example: Ferrari y Cribari-Neto (2004) propose a beta regression model based on a reparameterization of the beta distribution; Bayes et al. (2012) propose a beta rectangular regression model which is a mixture of a beta and a uniform distribution, and is robust to the presence of outliers; and Figueroa-Zúñiga et al. (2013) extends the beta regression of Ferrari y Cribari-Neto (2004) to mixed models from a bayesian perspective.

Although these models are suitable and useful for regression analysis on proportions, they assume that these proportions are restricted to the open unit interval. However, these variables could take the 0 or 1 values in real data, yielding to so called fractional bounded variables. Example of fractional bounded variables are: percentage of family income used for recreation purposes; percentage of credit limit used by a credit card client of a bank; or percentage of units of a brand new product sold by a company. The following types of models can be found in the literature which allow to explain a fractional bounded variable:

- Models based on transformations: The response variable is first transformed so any fractional bounded variable is taken from a closed unit interval to an open unit interval, and then any regression model for bounded variables such as the beta or beta rectangular models can be applied. Smithson y Verkuilen (2006) propose, for instance, to transform a fractional bounded variable in the following way:

$$Y^* = \frac{Y(N-1)+0.5}{N}$$

  where $Y \in [0,1]$ is the fractional response and N the sample size. The main disadvantage of these models is that extremely biased estimations can be obtained, as shown in Bayes y Valdivieso (2016).

- Two-part models: First, a multinomial regression model is applied to a categorical variable that clasifies the fractional response in 0, 1 or any value on the open unit interval $(0,1)$. Then, conditional to the fact that the response lies on the $(0,1)$ interval, any regression model for bounded variables, such as the beta regression, is applied. In Ramalho y da Silva (2009) a fractional response model is proposed, where the probability that the response variable takes the value 0 is estimated by fitting a binary model, and then another model is fitted when the variable value lies in open unit interval. Ospina y Ferrari (2010) proposes different models for intervals $(0, 1]$ (one-inflated beta, denoted by BEOI), $[0, 1)$ (zero-inflated beta, denoted by BEZI) and $[0, 1]$ (zero-and-one-inflated beta, denoted by BEINF). For the BEOI and BEZI models, a mixture of a Bernoulli and a beta distribution is proposed; for BEINF a mixture of a Multinomial and beta distribution is proposed. It is worth noting that the expected value of these mixtures are composed by different parameters, so effects on the mean are difficult to interpret under these models.

- One-part models: In this type of models, the mean response γ is directly modelled with the set of covariates. In Papke y Wooldridge (1996), γ is directly related with the vector of covariates through the equation $g(\gamma) = x^\top\beta$, where $g(.)$ is a proper link function. Estimation of this model is based on a quasi-likelihood methodology which maximizes a Bernoulli log-likelihood function, leading to a quasi-maximum likelihood estimator (QMLE). An important property of QMLE is the fact that it is consistent and asymptotically normal regardless of the distribution of the response variable conditional to covariates. On the other hand, Bayes y Valdivieso (2016) propose a beta inflated mean regression model which, based on a convenient reparameterization, allows to model directly the mean of the fractional bounded variable of interest; estimation of this model is carried out from a classical perspective. In this article we will try to extend this model to a mixed effect model since it is shown that in the transversal setting this model outperforms Papke y Wooldridge (1996) model in terms of root of mean squared error (RMSE), bias and information criteria.

Mixed-effects regressions are widely used to model data that consists of multiple measures for each subject over time (longitudinal data), or measures of subjects divided in well-defined groups (clustered data). Similarly to fixed effects models, two-part and one-part models including mixed effects can be found in the literature. We review below these models:

- Two-part models: In Wang y Luo (2016) the two-part regression model is extended to include mixed effects by using a one-augmented beta rectangular distribution that can easily be generalized to a zero-one-augmented beta rectangular distribution. Estimations in this model are carried out from a bayesian perspective. Furthermore, Galvis et al. (2014) propose a zero-and-one augmented beta random effects model, with parameters $p_0 = P(Y = 0)$ and $p_1 = P(Y = 1)$ and the beta parameters. A con of this model is that parameters p_0 and p_1 must meet condition $0 < p_0 + p_1 < 1$, which makes more difficult the estimation in the model. Additionally, in Liu y Kong (2015) a Zero-One Inflated Beta regression model with mixed-effects is proposed with parameters $p = P(Y = 0)$, $q = P(Y = 1 \mid Y \neq 0)$, $\mu = E(Y \mid Y \in (0, 1))$ (conditional mean) and ϕ (precision parameter).

- One-part models: In Papke y Wooldridge (2008) the model proposed in Papke y Wooldridge (1996) is extended to panel data by linking the mean of each observation $E(Y_{ij})$ to a set of covariates x_{ij} and z_{ij} associated to fixed and random effects, respectively. Specifically, the cumulative density function of the normal distribution is chosen as a link function. Estimation is carried out from a classical perspective by using generalized estimating equation (GEE).

The mixed effects regression models do not only provide flexibility, but also allows to identify within-subject and between-subject effects of covariates on the dependent variable, which is certainly useful for applications. In a credit card portfolio, for instance, is common to take decisions based on the percentage of utilization of a credit limit (from here denoted by $UTI\%$) by each client. When $UTI\% = 0\%$ then the client may not need the credit card, and another product can be offered; when $UTI\% = 100\%$ the client may need a greater credit limit; and when $UTI\% \in (0\%, 100\%)$ the bank can take decisions based on ranges. A bank is interested in measuring the effect of a set of covariates, both within-subject (over time) and between-subject (compared to others), on the expected value of $UTI\%$, so that a profile can be identified for expected values close to 0% or 100%.

1.2 Objectives

The main objective of this article is to estimate and apply a new beta inflated mean regression model with mixed effects to simulated and real data from a bayesian perspective, and compare results against models proposed in the literature. More specifically:

- Investigate about mixed effects models for fractional bounded data in the literature.
- Study the properties of a new beta inflated mean regression model with mixed effects.
- Implement the beta inflated mean regression model with mixed effects.
- Conduct a simulation study where the proposed model is compared with other models.
- Apply the proposed model to real data and compare results with other models.

1.3 Work organization

This article is organized in the following way: Chapter 2 describes the beta inflated distribution and points out the importance of the proposed reparameterization that allows to model directly the mean of a fractional response variable; Chapter 3 describes the structure of the beta inflated mean mixed regression, presents the augmented likelihood function, indicates the chosen prior distributions for all parameters which allows to construct the augmented posterior distribution, and describes information criteria indicators for model selection; Chapter 4 shows results obtained from a simulation study; Chapter 5 shows results from application of the proposed model to a real data; final comments are presented in Chapter 6.

Chapter 2

The beta inflated distribution

In this chapter we present the beta inflated distribution, its probability density function, its properties and an alternative reparametrization which allow us to model directly the mean of the independent variable of interest.

2.1 Probability density function

A random variable Y follows the beta inflated distribution if its probability density function is given by:

$$f_Y(y \mid \delta_0, \delta_1, \mu, \phi) = \begin{cases} \delta_0, & y = 0 \\ (1 - \delta_0 - \delta_1) b(y \mid \mu, \phi), & y \in (0, 1) \\ \delta_1, & y = 1 \end{cases} \tag{2.1}$$

where $P(Y = 0) = \delta_0$, $P(Y = 1) = \delta_1$, $E(Y \mid Y \in (0, 1)) = \mu$ and $\phi > 0$ is interpreted as a precision parameter. The function $b(. \mid \mu, \phi)$ is the probability density function of the beta distribution with a convenient parametrization such that μ is the mean and $\frac{\mu(1-\mu)}{1+\phi}$ is the variance of the distribution. The mean and variance of the beta inflated distribution are:

$$E(Y) = \delta_1 + (1 - \delta_0 - \delta_1)\mu$$
$$V(Y) = \delta_1(1 - \delta_1) + (1 - \delta_0 - \delta_1)\left(\tfrac{\mu(1-\mu)}{1+\phi} + (\delta_0 + \delta_1)\mu^2 - 2\mu\delta_1\right)$$

2.2 Alternative parametrization

In the context of a regression model, this parametrization does not allow to measure the effects of the covariates directly on the mean of the dependent variable, since the expected value of Y, to be denoted hereafter by γ, satisfies $\gamma = \delta_1 + (1 - \delta_0 - \delta_1)\mu$ and the parameters δ_0, δ_1 and μ are commonly associated to different effects. Furthermore, it should be noted that γ is restricted to the open interval $(\delta_1, 1 - \delta_0)$. An alternative to solve these problems is described in Bayes y Valdivieso (2016) where a reparametrization of (5.1) is proposed as follows:

$$\gamma = \delta_1 + (1 - \delta_0 - \delta_1)\mu, \quad \alpha_0 = \frac{\delta_0}{1-\gamma} \quad \text{y} \quad \alpha_1 = \frac{\delta_1}{\gamma}$$

where $\gamma \in]0, 1[$, $\alpha_0 \in]0, 1[$ and $\alpha_1 \in]0, 1[$. The mean, variance and probability density function are rewritten under this parametrization as follows:

$$E(Y) = \gamma, \qquad V(Y) = \frac{(1+\alpha_1\phi)}{1+\phi}\gamma + \left(\frac{(1-\alpha_1)^2\phi)}{(1-\alpha_0(1-\gamma)-\alpha_1\gamma)(1+\phi)} - 1\right)\gamma^2$$

and

$$f_Y(y \mid \alpha_0, \alpha_1, \gamma, \phi) = \begin{cases} \alpha_0(1-\gamma), & y = 0 \\ (1-\alpha_0(1-\gamma)-\alpha_1\gamma) \times b\left(y \,\middle|\, \frac{\gamma(1-\alpha_1)}{1-\alpha_0(1-\gamma)-\alpha_1\gamma}, \phi\right), & y \in (0,1) \\ \alpha_1\gamma, & y = 1 \end{cases} \tag{2.2}$$

This reparameterization not only allows to model $E(Y) = \gamma$ directly, but also breaks the restriction for γ, so that a better scenario for a mean regression analysis is established. If (2.2) is the density function of the random variable Y, then it is said that this variable follows a Beta Inflated mean distribution and is denoted by $Y \sim BetaInf(\alpha_0, \alpha_1, \gamma, \phi)$.

Chapter 3

The beta inflated mean mixed regression model

In this chapter we present the definition of the beta inflated mean mixed regression model, the augmented likelihood function, the augmented posterior distribution, the chosen priors for all parameters and the model comparison criteria.

3.1 Model definition

Let $Y_i = [Y_{i1}, \ldots, Y_{in_i}]^\top$, $i = 1, \ldots, n$ be n independent response vector variables for the subjects in the study, where:

$$Y_{ij} \sim BetaInf(\alpha_{0ij}, \alpha_{1ij}, \gamma_{ij}, \phi) \tag{3.1}$$

We can link the parameters α_{0ij}, α_{1ij} and γ_{ij} to covariates through appropriate link functions as shown next:

$$\begin{aligned} g_1(\alpha_{0ij}) &= \hat{x}_{ij}^\top \omega + \hat{z}_{ij}^\top w_i \\ g_2(\alpha_{1ij}) &= \breve{x}_{ij}^\top \delta + \breve{z}_{ij}^\top d_i \\ g_3(\gamma_{ij}) &= x_{ij}^\top \beta + z_{ij}^\top b_i \end{aligned}$$

where $\omega = [\omega_1, \ldots, \omega_k]^\top$, $\delta = [\delta_1, \ldots, \delta_l]^\top$ y $\beta = [\beta_1, \ldots, \beta_m]^\top$ are vectors of regression coefficients (fixed effects) associated to α_{0ij}, α_{1ij} and $E(Y_{ij}) = \gamma_{ij}$, respectively; $w_i = [w_{i1}, \ldots, w_{ip}]^\top$, $d_i = [d_{i1}, \ldots, d_{ir}]^\top$ and $b_i = [b_{i1}, \ldots, b_{is}]^\top$ are random effects associated to α_{0ij}, α_{1ij} and γ_{ij}, respectively; $\hat{x}_{ij} = [\hat{x}_{ij1}, \ldots, \hat{x}_{ijk}]^\top$, $\hat{z}_{ij} = [\hat{z}_{ij1}, \ldots, \hat{z}_{ijp}]^\top$, $\breve{x}_{ij} = [\breve{x}_{ij1}, \ldots, \breve{x}_{ijl}]^\top$, $\breve{z}_{ij} = [\breve{z}_{ij1}, \ldots, \breve{z}_{ijr}]^\top$, $x_{ij} = [x_{ij1}, \ldots, x_{ijm}]^\top$ and $z_{ij} = [z_{ij1}, \ldots, z_{ijs}]^\top$ are covariate vectors which can be different, overlapped or even identical; $\phi > 0$ is a precision parameter; and $g_1(.)$, $g_2(.)$ and $g_3(.)$ are link functions with continuous second derivatives such that $g_v : (0, 1) \mapsto \mathbb{R}, \quad v = 1, 2, 3$.

Any cumulative distribution function of a continuous variable can be an appropriate inverse link function. Among these we have the probit link function which has as a disadvantage to increase the difficulty in interpretation of the effects over the dependent variable. In order to ease the interpretation, we are going to use the inverse of the logistic cumulative distribution as the link function for parameters α_{0ij}, α_{1ij} and γ_{ij}.

With respect to the random effects vectors w_i, d_i and b_i, it will be assumed that these vectors are independent and identically distributed with multivariate normal distributions:

$$
\begin{aligned}
w_i &\sim N_p(0, \Sigma_w) \\
d_i &\sim N_r(0, \Sigma_d) \\
b_i &\sim N_s(0, \Sigma_b)
\end{aligned}
$$

being Σ_w, Σ_d and Σ_b positive definite matrices.

Defining $w = [w_1, \ldots, w_n]^\top$, $d = [d_1, \ldots, d_n]^\top$, $b = [b_1, \ldots, b_n]^\top$, $\theta = [\omega, \delta, \beta, \Sigma_w, \Sigma_d, \Sigma_d, \phi]$ and $Y = [Y_1, \ldots, Y_n]^\top$, the augmented likelihood function for model (3.1) can be written as follows:

$$
\begin{aligned}
L(\theta, w, d, b \mid Y) =& p(Y, w, d, b \mid \theta) \\
=& \prod_{i=1}^{n} f_{Y_i}(y_i \mid \alpha_{0i}, \alpha_{1i}, \gamma_i, \phi) \times \varphi_p(w_i \mid 0, \Sigma_w) \times \varphi_r(d_i \mid 0, \Sigma_d) \times \varphi_s(b_i \mid 0, \Sigma_s)
\end{aligned} \tag{3.2}
$$

where $\varphi_t(. \mid \mu_t, \Sigma_t)$ denotes the probability density function of a multivariate normal distribution with mean vector μ_t and covariance matrix Σ_t, $\alpha_{0i} = [\alpha_{0i1}, \ldots, \alpha_{0in_i}]^\top$, $\alpha_{1i} = [\alpha_{1i1}, \ldots, \alpha_{1in_i}]^\top$ and $\gamma_i = [\gamma_{i1}, \ldots, \gamma_{in_i}]^\top$ are parameter vectors of size n_i with

$$
\begin{aligned}
\alpha_{0ij} &= 1/\{1 + exp(-(\hat{x}_{ij}^\top \omega + \hat{z}_{ij}^\top w_i))\} \\
\alpha_{1ij} &= 1/\{1 + exp(-(\breve{x}_{ij}^\top \delta + \breve{z}_{ij}^\top d_i))\} \\
\gamma_{ij} &= 1/\{1 + exp(-(x_{ij}^\top \beta + z_{ij}^\top b_i))\}
\end{aligned}
$$

and $f_{Y_i}(y_i \mid \alpha_{0i}, \alpha_{1i}, \gamma_{1i}, \phi)$ is the joint probability density function of the vector $Y_i = [Y_{i1}, \ldots, Y_{in_i}]^\top$, which can be expanded as follows assuming conditional independence to random effects:

$$
f_{Y_i}(y_i \mid \theta, w_i, d_i, b_i) = \prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}, \alpha_{1ij}, \gamma_{ij}, \phi)
$$

where $f_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}, \alpha_{1ij}, \gamma_{1ij}, \phi)$ is the probability density function of beta inflated distribution as defined in (2.2).

3.2 Bayesian Inference

Taking into account the augmented likelihood function as described in (3.2), the augmented posterior distribution of θ, w, d, b, denoted by $p(\theta, w, d, b \mid Y)$, can be written as follows:

$$
p(\theta, w, d, b \mid Y) \propto p(Y \mid \theta, w, d, b) \times p(w, d, b \mid \theta) \times p(\theta)
$$

which can be also expressed as:

$$
p(\theta, w, d, b \mid Y) \propto L(\theta, w, d, b \mid Y) \times p(\theta) \tag{3.3}
$$

where $p(\theta)$ is the prior distribution of θ. In this article we consider that ω, δ, β, Σ_w, Σ_d, Σ_b

and ϕ are independent, so we can set the prior distribution as:

$$p(\theta) = p(\omega)p(\delta)p(\beta)p(\Sigma_w)p(\Sigma_d)p(\Sigma_b)p(\phi) \tag{3.4}$$

For fixed effects vectors we propose multivariate normal distributions such that $\omega \sim N_k(0, A)$, $\delta \sim N_l(0, B)$ and $\beta \sim N_m(0, C)$. For covariance matrices we propose as prior an Inverse Wishart distribution such that $\Sigma_w \sim IW(\psi_w, \Psi_w)$, $\Sigma_d \sim IW(\psi_d, \Psi_d)$ and $\Sigma_b \sim IW(\psi_b, \Psi_b)$. For the precision parameter it is set as prior a gamma distribution such that $\phi \sim Gamma(a, b)$. For all these prior distributions, A, B, C, ψ_w, Ψ_w, ψ_d, Ψ_d, Ψ_b, a and b are specified hyperparameters.

Combining the augmented likelihood function defined in (3.2) with the prior distribution defined in (3.4) the augmented posterior distribution defined in (3.3) can be written as:

$$\begin{aligned} p(\theta, w, d, b \mid Y) \propto \prod_{i=1}^{n} & \left[\prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}, \alpha_{1ij}, \gamma_{1ij}, \phi) \right] \\ & \times \varphi_p(w_i \mid 0, \Sigma_w) \times \varphi_r(d_i \mid 0, \Sigma_d) \times \varphi_s(b_i \mid 0, \Sigma_b) \\ & \times \varphi_k(\omega \mid 0, A) \times \varphi_l(\delta \mid 0, B) \times \varphi_m(\beta \mid 0, C) \\ & \times h(\Sigma_w \mid \psi_w, \Psi_w) \times h(\Sigma_d \mid \psi_d, \Psi_d) \times h(\Sigma_b \mid \psi_b, \Psi_b) \\ & \times q(\phi \mid a, b) \end{aligned} \tag{3.5}$$

where $\varphi_t(. \mid \mu_t, \Sigma_t)$ denotes the probability density function of a multivariate normal distribution with mean vector μ_t and covariance matrix Σ_t, $q(. \mid a, b)$ denotes the probability density function of gamma distribution with mean equal to $\frac{a}{b}$, and $h(. \mid \psi, \Psi)$ denotes the probability density function of the Inverse Wishart distribution with mean $\frac{\Psi}{\psi-p-1}$, where ψ is interpreted as a degree of freedom and Ψ is a $p \times p$ scale matrix.

As seen in (3.5), the augmented posterior distribution $p(\theta, w, d, b \mid Y)$ is a complex expression. To obtain samples from it, we will make use of Hamiltonian MCMC methods as the ones implemented in the **R** package *RStan*. This package implements an adaptive Hamiltonian Monte Carlo algorithm (also known as HMC) using a No-U-Turn sampler (NUTS) for the stepsize parameter in order to generate efficient transitions to the posterior distribution. The No-U-Turn sampler was proposed in Hoffman y Gelman (2014) and its main advantage is to select adaptively an adequate number of steps in each iteration so the posterior distribution is reached more efficiently.

3.2.1 Model comparison criteria

A great number of model comparison information criteria can be found in the literature to assess the fit of different models. Information criteria such as the Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2002), Widely Applicable Information Criterion (WAIC), proposed by Watanabe (2010), and expected Akaike's Information Criteria (EAIC) and Bayesian Information Criteria (EBIC) as detailed in Gelman et al. (2014), are based in the deviance $\mathcal{D}(.)$. Before defining the deviance, let $\nu = [\theta, w, d, b]$ be an array parameter that encapsulates all parameters of the augmented posterior distribution (3.5).

We define the deviance as follows:

$$\mathcal{D}(\nu) = -2log(L(\nu \mid Y)) = -2log(L(\theta, w, d, b \mid Y)),$$

where $L(\theta, w, d, b \mid Y)$ is the augmented likelihood function defined in (3.2). Then, we define the DIC criteria as:

$$DIC = \mathcal{D}(\bar{\nu}) + 2 \times p_D,$$

where $p_D = \bar{\mathcal{D}}(\nu) - \mathcal{D}(\bar{\nu})$ can be interpreted as number of effective parameters. Considering a Montecarlo sample size M taken from the augmented posterior distribution defined in (3.5), the terms $\bar{\mathcal{D}}(\nu)$ and $\mathcal{D}(\bar{\nu})$ are calculated as follows:

$$\bar{\mathcal{D}}(\nu) = \sum_{m=1}^{M} \frac{\mathcal{D}(\nu_m)}{M} \quad \text{and} \quad \bar{\nu} = \sum_{m=1}^{M} \frac{\nu_m}{M}.$$

On the other hand, WAIC is calculated similarly to DIC, only differing on the effective parameter count term. Watanabe (2010) define WAIC as follows:

$$WAIC = \mathcal{D}(\bar{\nu}) + 2 \times p_{WAIC},$$

where $p_{WAIC} = \sum_{i=1}^{n} Var(log(p(Y_i \mid \nu)))$.

Finally, the EAIC and EBIC criteria, detailed in Gelman et al. (2014), are defined as follows:

$$EAIC = \mathcal{D}(\bar{\nu}) + p \quad \text{and} \quad EBIC = \mathcal{D}(\bar{\nu}) + p \times log(n),$$

where p is the number of parameters in the model and n is the sample size. It should be noted that EBIC incorporates a penalty term linked to the sample size, so that simpler models are favored.

Since a lower value of DIC, WAIC, EAIC and EBIC indicates a better fit, the model with the lowest value of these indicators will be considered as the best in this article.

Chapter 4

Simulation study

In this section we present a parameter recovery simulation study where we will make use of the Beta Inflated mean mixed-effects regression model (denoted for short by BInf) introduced in Chapter 3, and compare the results with the Beta Transformed mixed-effects regression (denoted by BTran) applying the transformation proposed in Smithson y Verkuilen (2006) to the response variable, and the model proposed in Papke y Wooldridge (2008) for panel data (denoted by PW).

4.1 Generation of data

In the context of longitudinal data, let us consider 100 subjects and 3 measurements for each subject. For fixed effects we construct a design matrix $x = [x_1, \ldots, x_{300}]^\top$ where each x_k is a vector of 3 elements, with the first element being constant and equal to 1, and the other 2 elements were sampled from a multivariate normal distribution with mean vector equal to 0, variances equal to 0.1 and covariance equal to 0.05. For random effects we construct a design matrix $z = [z_1, \ldots, z_{300}]^\top$ where each z_k is a vector of 2 elements, so that $z_{k1} = x_{k2}$ and $z_{k2} = x_{k3}$.

4.2 Parameter recovery

We will incorporate covariates to the parameters α_0, α_1 and γ in the Beta Inflated mixed-effect regression, and only to the conditional mean parameter μ in the Beta Transformed mixed-effect regression, and assume all remaining parameters are constant. Similarly, the Papke and Wooldridge model incorporates covariates only to the mean of the response variable. We set the fixed effects coefficients as $\omega = [\omega_1, \omega_2, \omega_3]^\top = [-1, 1.5, 0.7]^\top$, $\delta = [\delta_1, \delta_2, \delta_3]^\top = [1, -1.5, 0.8]^\top$, $\beta = [\beta_1, \beta_2, \beta_3]^\top = [0.5, -1.0, 1.0]^\top$ and variance-covariance matrices for random effects coefficient as $\Sigma_w = \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 0.8 \end{smallmatrix}\right]$, $\Sigma_d = \left[\begin{smallmatrix} 1 & -0.7 \\ -0.7 & 0.9 \end{smallmatrix}\right]$ and $\Sigma_b = \left[\begin{smallmatrix} 1 & -0.3 \\ -0.3 & 0.2 \end{smallmatrix}\right]$. The 100 random effects w_i, d_i and b_i were then sampled from a multivariate normal distribution with mean equal to 0 and variance-covariance matrices equal to Σ_w, Σ_d and Σ_b, respectively. The precision parameter ϕ, on the other hand, was set to 50. Finally, we sampled 300 responses $y_{ij} \sim BetaInf(\alpha_{0ij}, \alpha_{1ij}, \gamma_{ij}, \phi)$, where $g(\alpha_{0ij}) = x_{ij}^\top \omega + z_{ij}^\top w_i$, $g(\alpha_{1ij}) = x_{ij}^\top \delta + z_{ij}^\top d_i$, $g(\gamma_{ij}) = x_{ij}^\top \beta + z_{ij}^\top b_i$ and $g(.)$ is the inverse logit function.

Regarding the prior distributions, for fixed effects coefficients we set $\omega \sim N_3(0, 10^4 I_3)$, $\delta \sim N_3(0, 10^4 I_3)$ and $\beta \sim N_3(0, 10^4 I_3)$, for variance-covariance matrix we set $\Sigma_w \sim IW(5, 20I_2)$, $\Sigma_d \sim IW(5, 20I_2)$ and $\Sigma_b \sim IW(5, 20I_2)$, and for the precision parameter we set $\phi \sim Gamma(0.0001, 0.0001)$. For estimation, we discarded the first 1000 iterations and obtained

9000 samples considering a thinning equal to 4, leading to 2250 iterations for each parameter. We repeated this estimation 1000 times with 1000 different design matrices.

Table 4.1 shows the bias and root mean squared error (RMSE) of fixed-effects coefficients as estimated by the Beta Inflated and Beta Transformed mixed-effects regressions. As can be clearly seen, the Beta Inflated estimations outperform the Beta Transformed estimations in terms of bias and RMSE.

Table 4.1: Bias and Root Mean Squared Error comparison between Beta Inflated and Beta Transformed mixed-effects regressions.

Parameter	**True Value**	**Bias (BInf)**	**Bias (BTran)**	**RMSE (BInf)**	**RMSE (BTran)**
β_1	0.5	-0.005180	-0.057026	0.101059	0.097314
β_2	-1.0	-0.004127	-0.176107	0.383588	0.310990
β_3	1.0	-0.091412	0.429609	0.375319	0.498653

Table 4.2 shows a comparison of the bias and root mean squared error (RMSE) of Beta Inflated and Papke and Wooldridge mixed-effects regression models. Although estimations of Papke and Wooldridge model are better than Beta Transformed model in terms of bias and RMSE, Beta Inflated model still outperforms Papke and Wooldridge model. It should be noted that the R package *frmpd*, which was used to estimate Papke and Wooldridge model, only provides probit as the link function. For this reason, the conversion proposed in Amemiya (1981) was used to obtain logit coefficients from probit coefficients. Since this limitation would put Papke and Wooldridge model in disadvantage, we also conducted the simulation study considering a probit link function for all 3 models, and obtained very similar results.

Table 4.2: Bias and Root Mean Squared Error comparison between Beta Inflated and Papke mixed-effects regressions.

Parameter	**True Value**	**Bias (BInf)**	**Bias (PW)**	**RMSE (BInf)**	**RMSE (PW)**
β_1	0.5	-0.005180	-0.025101	0.101059	0.104240
β_2	-1.0	-0.004127	0.020626	0.383588	0.452057
β_3	1.0	-0.091412	-0.035931	0.375319	0.445121

We carried out similar simulations with different values for parameters β and ϕ, and varying the number of cases n. Tables 4.3 and 4.4 show the bias and RMSE comparison between Beta Inflated, Beta Transformed and Papke and Wooldridge mixed-effects regression models for multiple scenarios, respectively. As can be seen, results indicate that the Beta Inflated mean mixed effects regression model have lower bias and RMSE than the other 2 regression models.

Table 4.3: Bias comparison between Beta Inflated, Beta Transformed and Papke and Wooldridge mixed-effects regressions for multiple scenarios.

n	Parameter	True Value	Bias (BInf)	Bias (BTran)	Bias (PW)
50	β_1	0.5	0.00085	-0.04146	-0.01898
	β_2	-1.0	0.00411	-0.16788	-0.00681
	β_3	1.0	-0.10966	0.41706	-0.00557
100	β_1	0.5	0.00137	-0.05071	-0.01802
	β_2	-1.0	-0.00072	-0.16971	0.01473
	β_3	1.0	-0.08728	0.43531	-0.02943
50	β_1	0.25	-0.00824	0.14075	-0.00152
	β_2	-3.0	0.16269	-1.29950	-0.07367
	β_3	2.5	-0.11839	1.07320	0.08280
100	β_1	0.25	-0.00543	0.14763	-0.00093
	β_2	-3.0	0.04727	-1.34402	-0.16671
	β_3	2.5	-0.05272	1.09230	0.14585
500	β_1	0.5	-0.00143	-0.06625	-0.02030
	β_2	-1.0	-0.00131	-0.17446	0.00645
	β_3	1.0	-0.05161	0.45966	-0.02832
500	β_1	0.25	0.01119	0.16802	0.00890
	β_2	-3.0	-0.04395	-1.38013	-0.20656
	β_3	2.5	0.03447	1.12106	0.19269

Table 4.4: RMSE comparison between Beta Inflated, Beta Transformed and Papke and Wooldridge mixed-effects regressions for multiple scenarios.

n	Parameter	True Value	RMSE (BInf)	RMSE (BTran)	RMSE (PW)
50	β_1	0.5	0.14019	0.11579	0.13958
	β_2	-1.0	0.56411	0.42070	0.66963
	β_3	1.0	0.53663	0.55705	0.61565
100	β_1	0.5	0.09859	0.09176	0.10069
	β_2	-1.0	0.38370	0.31131	0.46754
	β_3	1.0	0.37928	0.50600	0.44829
50	β_1	0.25	0.16972	0.17142	0.17341
	β_2	-3.0	0.69073	1.34768	0.81149
	β_3	2.5	0.71207	1.13595	0.82234
100	β_1	0.25	0.10750	0.16043	0.10921
	β_2	-3.0	0.46578	1.36623	0.58826
	β_3	2.5	0.46744	1.11926	0.56864
500	β_1	0.5	0.04096	0.07450	0.04652
	β_2	-1.0	0.16048	0.20589	0.19927
	β_3	1.0	0.16373	0.47193	0.19405
500	β_1	0.25	0.04817	0.17028	0.04918
	β_2	-3.0	0.20121	1.38429	0.31784
	β_3	2.5	0.20562	1.12619	0.31030

Chapter 5

Real data analysis

This chapter presents the results obtained from the application of our Beta Inflated mean regression model with mixed effects (denoted by BInf) to real data, and its statistical comparison with the competitive Zero-One Inflated Beta regression model with mixed effects (denoted by ZOIB) proposed in Liu y Kong (2015), which can be considered as a two-part model.

The main objective in our application is to study the effect of a set of covariates in the utilization of a credit line by clients of a bank. With this information a bank could assign profiles to potential clients such as: *Regular Credit Card User*, *Medium Credit Card User* or *Non Credit Card User* which can be used to elaborate a more personalized offer for them.

5.1 Data

Data consisted of 200 individuals that were reported with at least one credit card in the Financial System of Peru during January 2016 and July 2016. We considered as covariates: a flag that determines whether an individual was reported in Financial System with cash advance (1) or not (0), denoted by *flag_cash_adv*; a flag that determines whether an individual was classified as Low Risk (0) or High Risk (1), denoted by *flag_clas*; and standardized age of individual denoted by *age_ind*. As dependent variable we considered utilization of credit line, which is defined as a ratio with numerator equal to the total amount used by the individual in purchases, cash advance or balance transfer, and denominator equal to the total amount of credit line granted to the individual. We will denote utilization of credit line as *uti_cc*.

These four variables were obtained for 3 visits corresponding to January 2016, April 2016 and July 2016. It is important to note that utilization of credit line can be equal to 0 (if individual did not use the credit line) or 1 (if individual used all the credit line) or lie in open unit interval (if individual partially used the credit line). Table 5.1 shows the dependent variable *uti_cc* frequency.

Table 5.1: Frequencies of utilization of credit line equal to 0, 1 or a value in the interval $(0, 1)$.

Value	**Freq.**	**Freq. %**
0	125	21%
1	97	16%
$(0, 1)$	378	63%

5.2 Model structure

Before defining the model structure for the real data analysis, it should be noted that we are considering the following parameterization of the ZOIB regression model proposed in Liu y Kong (2015):

$$f_Y(y \mid p,q,\mu,\phi) = \begin{cases} p, & y=0 \\ (1-p)(1-q)b(y \mid \mu,\phi), & y \in (0,1) \\ (1-p)q, & y=1 \end{cases} \tag{5.1}$$

where $p = P(Y=0)$, $q = P(Y=1 \mid Y \neq 0)$, $\mu = E(Y \mid Y \in (0,1))$ and ϕ is a precision a parameter.

We incorporate covariates for the parameters γ (mean), α_0 and α_1 in the BInf regression model, and for the parameters μ (conditional mean), p and q in the ZOIB regression model. For fixed effects parameters we considered an intercept and coefficients associated to variables *flag_cash_adv*, *flag_clas* and *age_ind*. It is important to note that variable *flag_cash_adv* was incorporated only to γ and α_1 in the Beta Inflated mean regression, and μ and q in the Zero-One Augmented Beta regression because if a person has variable *flag_cash_adv* equal to 0, then the response variable *uti_cc* must be necessarily equal to 0. Regarding random effects we only included an intercept for γ, α_0 and α_1 in the BInf regression model, and for μ, p and q in the ZOIB regression model. We choose the inverse logit function to link the linear predictors to corresponding parameters. All remaining parameters of each distribution were assumed to be constant for all observations. Equations (5.2) and (5.3) summarize the BInf and ZOIB structure of model, respectively.

$$\begin{aligned} Y_{ij} &\sim BetaInf(\gamma_{ij},\alpha_{0ij},\alpha_{1ij},\phi) \\ g(\alpha_{0ij}) &= \omega_0 + \omega_1 flag\_clas + \omega_2 age\_ind + w_i \\ g(\alpha_{1ij}) &= \delta_0 + \delta_1 flag\_clas + \delta_2 age\_ind + \delta_3 flag\_cash\_adv + d_i \\ g(\gamma_{ij}) &= \beta_0 + \beta_1 flag\_clas + \beta_2 age\_ind + \beta_3 flag\_cash\_adv + b_i \end{aligned} \tag{5.2}$$

$$\begin{aligned} Y_{ij} &\sim ZOIB(\mu_{ij},p_{ij},q_{ij},\phi) \\ g(p_{ij}) &= \omega_0 + \omega_1 flag\_clas + \omega_2 age\_ind + w_i \\ g(q_{ij}) &= \delta_0 + \delta_1 flag\_clas + \delta_2 age\_ind + \delta_3 flag\_cash\_adv + d_i \\ g(\mu_{ij}) &= \beta_0 + \beta_1 flag\_clas + \beta_2 age\_ind + \beta_3 flag\_cash\_adv + b_i \end{aligned} \tag{5.3}$$

Regarding the prior distributions, we considered for fixed effects non-informative Multivariate Normal distributions $\omega \sim N_3(0, 10^4 I_3)$, $\delta \sim N_4(0, 10^4 I_4)$ and $\beta \sim N_4(0, 10^4 I_4)$. For the variance-covariance matrices Σ_w, Σ_d, Σ_b, which in this application reduces to an scalar since we are only considering a random intercept, we considered an Inverse-Gamma (univariate version of Inverse Wishart distribution) with 5 degrees of freedom and a scale parameter equal to 20. For the precision parameter ϕ we considered a non-informative prior $Gamma(0.0001, 0.0001)$.

5.3 Results

Both BInf and ZOIB regression models were implemented and estimated in Stan considering 2250 effective iterations after discarding first 1000 and setting a thinning equal to 4 to avoid autocorrelation. Estimation of fixed parameters and random intercepts under the BInf model and ZOIB model for a sample of 5 subjects are presented in Tables 5.2 and 5.3, respectively. Information criteria comparison is shown in Table 5.4 and distribution of random intercepts (b_i parameters) for a sample of 10 subjects is shown in Figure 5.1. Additionally, ordered posterior distributions of random errors b_i, d_i and w_i of BInf regression model for all subjects are shown in Figures 5.2, 5.3 and 5.4, respectively, considering a range equal to 1.5 times the interquartile range; for ZOIB regression model, ordered posterior distributions are shown in Figures 5.5, 5.6 and 5.7.

As can be seen in Tables 5.2 and 5.3, fixed effects ω and δ have a lower standard deviation in BInf regression model, while fixed effects β have similar standard deviation in both regression models. Standard deviations for random effects w_i and d_i are notoriously lower in BInf regression model. Regarding random effects b_i, standard deviations from BInf regression model are slightly greater than ZOIB regression model. However, for some subjects, such as 7 and 17, although the standard deviations are greater in BInf regression model, the credible interval does not contain value zero, while ZOIB regression model does. Furthermore, information criteria shown in Table 5.4 indicates that BInf regression model fitted better to credit card data than ZOIB regression model.

Regarding the interpretation of fixed and random effects in the Beta Inflated regression, results are very useful from a bank perspective. Fixed effect $\beta_1 = 1.71990$ related to covariate *flag_clas* indicates that if an individual is reported as High Risk in the Financial System, mean credit card utilization would be 5.58397 times greater than an individual reported as Low Risk. Besides, fixed effect $\beta_2 = -0.40044$ related to covariate *age_ind* indicates that the older the individual, the lower the mean credit card utilization. Furthermore, fixed effect $\beta_3 = 1.53859$ indicates that if an individual is reported with cash advance balance greater than 0, mean credit card utilization would be 4.658018 greater than an individual reported with cash advance balance equal to 0.

On the other hand, random intercepts point estimations show which subjects are more likely to increment (positive random intercept) or decrement (negative random intercept) their mean credit card utilization. Interpretation of fixed and random effects estimations from ZOIB regression model are very similar, with the disadvantage that these are only related to conditional mean μ rather than the mean $E(Y)$.

Table 5.2: Summary of estimated posterior distribution of parameters of BInf regression model applied to credit card data.

Parameter	Mean	SD	P 2.5%	P 97.5%
ω_0	-2.07275	0.47584	-3.15136	-1.27737
ω_1	0.50591	0.79172	-1.09056	1.98536
ω_2	0.33207	0.32592	-0.28454	0.96937
δ_0	-2.84109	0.50192	-3.98459	-1.99172
δ_1	3.91431	0.56780	2.90692	5.11508
δ_2	-0.03006	0.26343	-0.53668	0.49176
δ_3	0.42369	0.46342	-0.45968	1.37821
β_0	-1.25280	0.13903	-1.52676	-0.98305
β_1	1.71990	0.27088	1.18295	2.25010
β_2	-0.40044	0.12039	-0.63096	-0.17149
β_3	1.53859	0.17636	1.19109	1.89366
Σ_w	11.09	3.75	5.65	20.09
Σ_d	2.98	1.05	1.46	5.55
Σ_b	1.86	0.30	1.35	2.49
w_4	0.07	1.42	-2.92	2.65
w_7	-2.03	2.37	-7.38	1.81
w_{13}	4.62	1.53	1.53	7.66
w_{17}	1.56	1.33	-0.93	4.27
w_{19}	-1.84	2.45	-7.21	2.21
d_4	0.14	1.79	-3.49	3.56
d_7	-0.42	1.53	-3.53	2.42
d_{13}	1.28	1.29	-1.34	3.84
d_{17}	1.26	1.39	-1.54	3.98
d_{19}	-0.47	1.51	-3.57	2.35
b_4	-1.36	0.71	-2.85	-0.04
b_7	1.27	0.58	0.05	2.35
b_{13}	1.81	0.88	0.10	3.51
b_{17}	-2.18	0.96	-4.04	-0.35
b_{19}	1.22	0.51	0.19	2.21

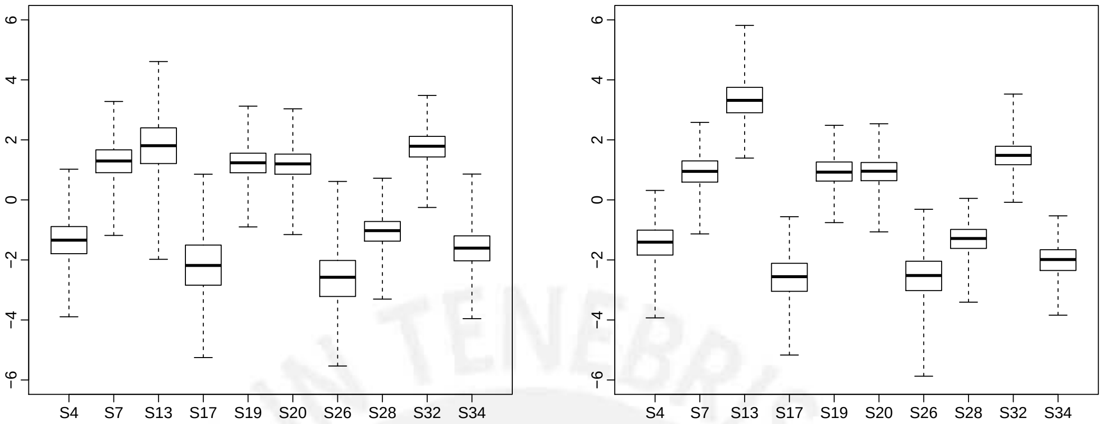
Table 5.3: Summary of estimated posterior distribution of parameters of ZOIB regression model applied to credit card data.

Parameter	**Mean**	**SD**	**P 2.5%**	**P 97.5%**
ω_0	-4.28190	0.83602	-6.21675	-2.86920
ω_1	-2.37523	1.17816	-4.97491	-0.32937
ω_2	0.63549	0.51531	-0.35589	1.69414
δ_0	-4.87325	0.72316	-6.45524	-3.65383
δ_1	5.02683	0.78715	3.69651	6.62932
δ_2	-0.38714	0.3435	-1.07974	0.25293
δ_3	0.76040	0.55560	-0.28424	1.89022
β_0	-0.91590	0.13686	-1.17549	-0.63751
β_1	0.71626	0.23908	0.24165	1.18355
β_2	-0.35553	0.11589	-0.57956	-0.12782
β_3	1.28371	0.17034	0.94471	1.61563
Σ_w	31.59	11.81	15.11	58.87
Σ_d	6.92	2.59	3.04	13.15
Σ_b	1.87	0.27	1.39	2.46
w_4	1.71	1.92	-2.04	5.32
w_7	-3.69	3.99	-12.80	2.45
w_{13}	2.14	1.80	-1.62	5.49
w_{17}	7.04	2.09	3.27	11.46
w_{19}	-3.08	3.95	-12.34	3.07
d_4	-0.40	2.44	-5.29	4.03
d_7	-0.42	2.44	-5.59	3.97
d_{13}	3.66	1.86	0.02	7.30
d_{17}	-1.83	1.98	-6.25	1.70
d_{19}	-0.43	2.30	-5.30	3.79
b_4	-1.44	0.63	-2.78	-0.30
b_7	0.94	0.55	-0.18	2.01
b_{13}	3.34	0.65	2.16	4.70
b_{17}	-2.60	0.67	-3.97	-1.42
b_{19}	0.93	0.49	-0.08	1.88

Table 5.4: Information criteria comparison between BInf and ZOIB regression models applied to credit card data.

Model	**DIC**	**EAIC**	**EBIC**	**WAIC**
BInf	2864.350	2295.993	1878.787	2267.757
ZOIB	3142.884	2611.782	2194.576	2608.742

Figure 5.1: Posterior distribution of random intercepts (b_i parameters) for 10 subjects as modelled by BInf (left panel) and ZOIB (right panel) regression models applied to credit card data.



(a) BInf Random Intercepts related to γ

(b) ZOIB Random Intercepts related to μ

Figure 5.2: Posterior distribution of random intercepts (b_i parameters) for all subjects as modelled in BInf regression model ordered by the median.

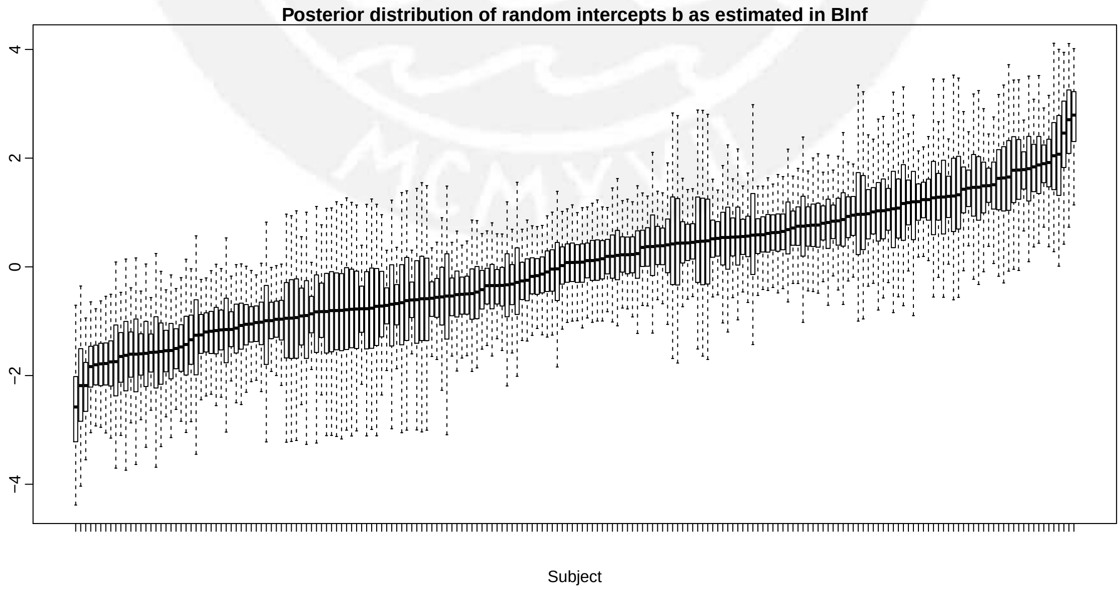


Figure 5.3: Posterior distribution of random intercepts (d_i parameters) for all subjects as modelled in BInf regression model ordered by the median.

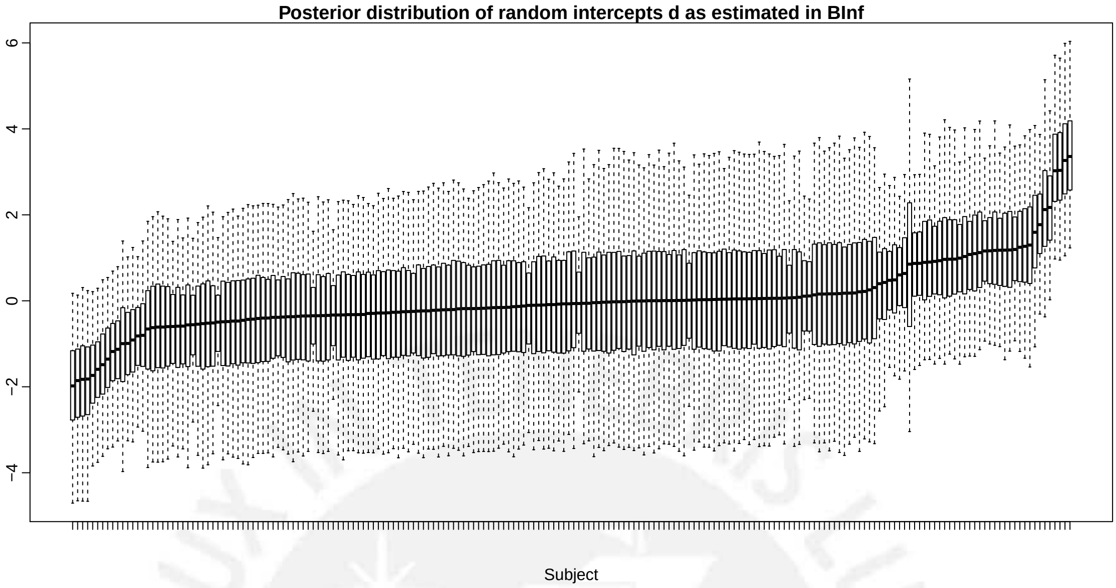


Figure 5.4: Posterior distribution of random intercepts (w_i parameters) for all subjects as modelled in BInf regression model ordered by the median.

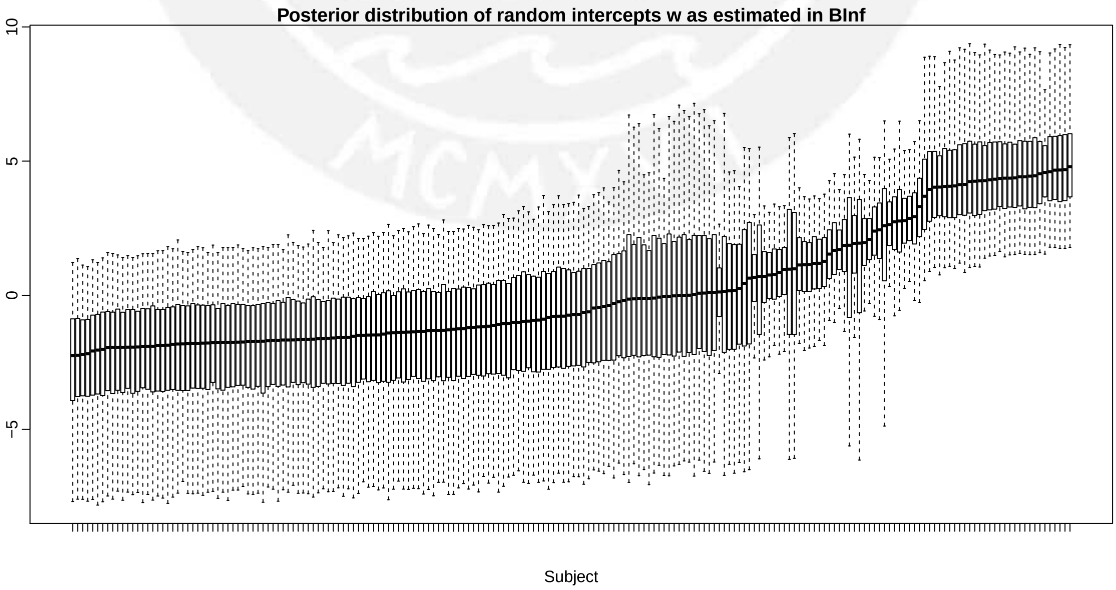


Figure 5.5: Posterior distribution of random intercepts (b_i parameters) for all subjects as modelled in ZOIB regression model ordered by the median.

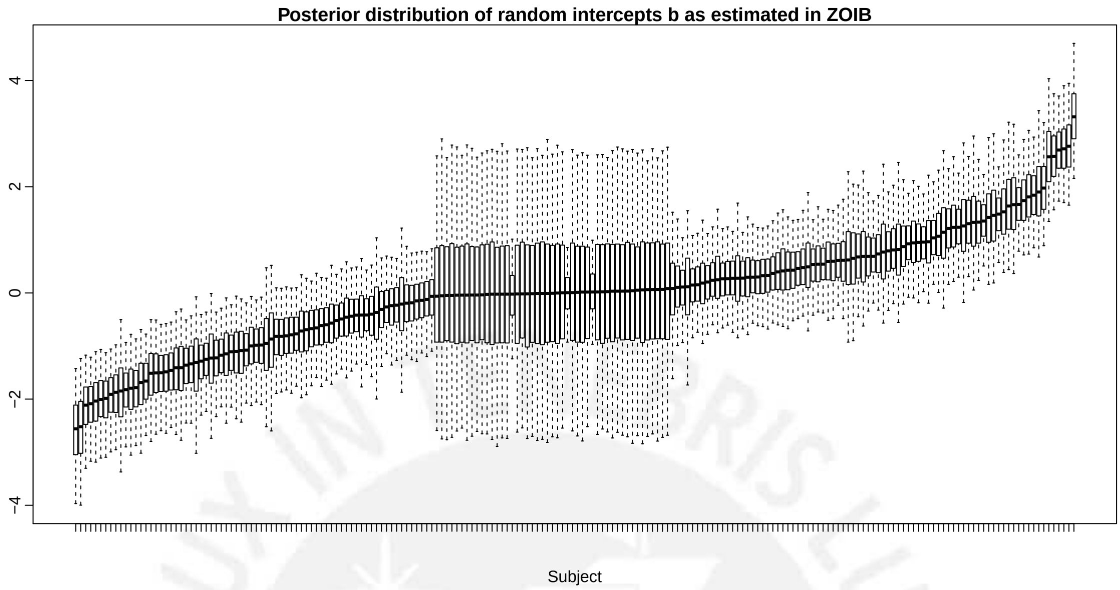


Figure 5.6: Posterior distribution of random intercepts (d_i parameters) for all subjects as modelled in ZOIB regression model ordered by the median.

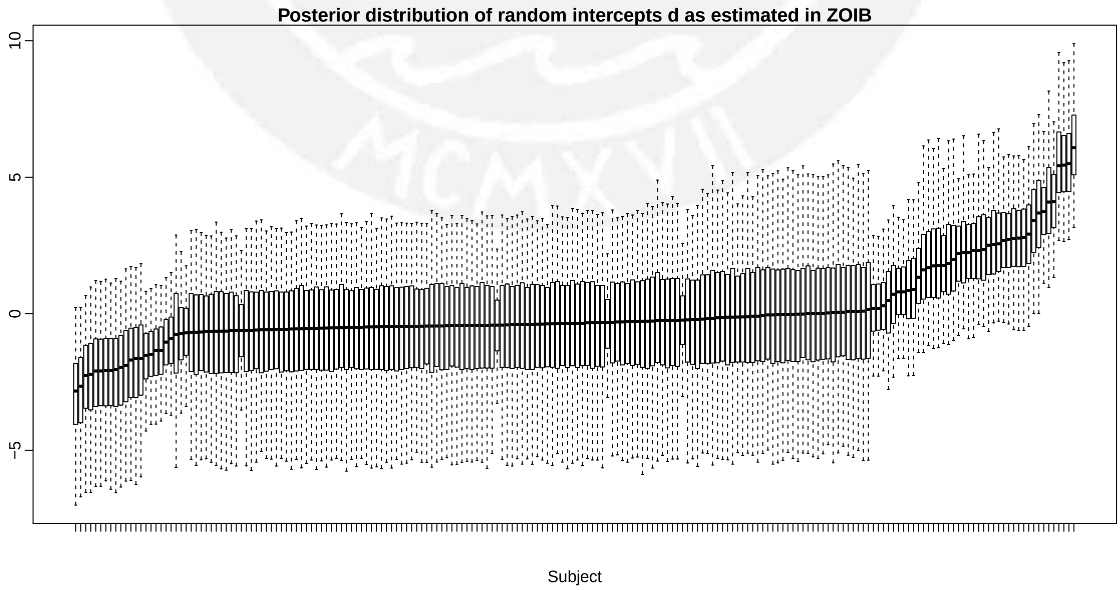
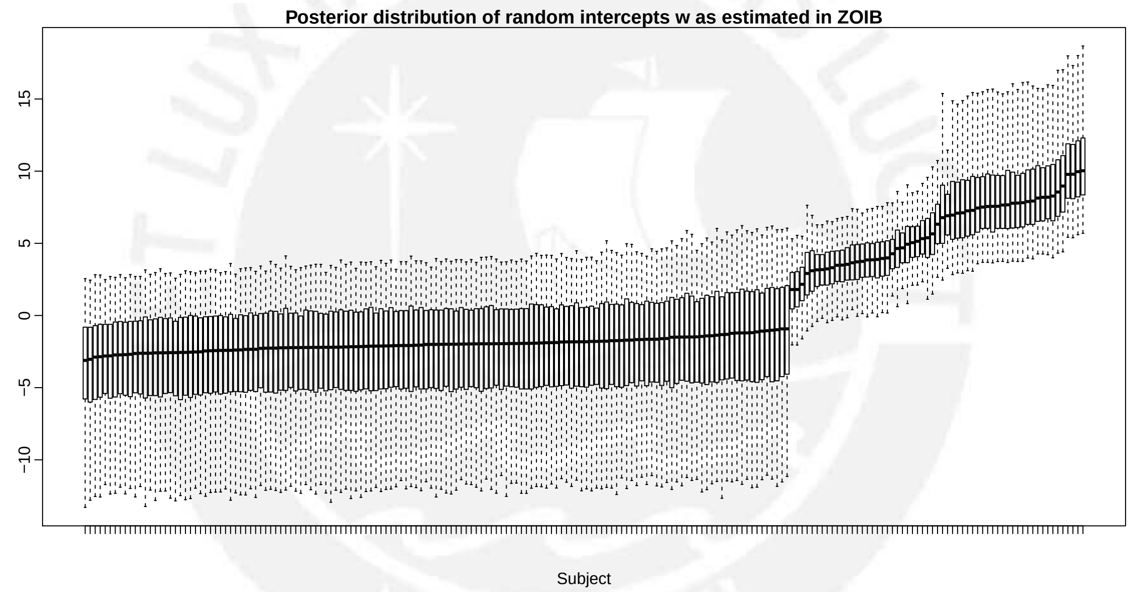


Figure 5.7: Posterior distribution of random intercepts (w_i parameters) for all subjects as modelled in ZOIB regression model ordered by the median.



Chapter 6

Final comments

6.1 Conclusions

In this article we have presented a Beta Inflated mean regression model with mixed effects. The main advantage of this regression is that it is possible to model directly the effects of covariates in the mean of a response variable which can lie in close unit interval $[0, 1]$. A general version of the model was provided in which covariates and fixed and random effects were incorporated to parameters α_0, α_1 and γ, as well as the augmented likelihood function. Estimation was carried out from a bayesian perspective due to the complexity of the model. Non-informative prior distribution for parameters as well as specifiable hyperparameters were also provided.

A simulation study was conducted in order to assess parameter recovery and compare results against Beta Transformed, which is basically a Beta regression with mixed effects considering as dependent variable a transformation of original response as proposed by Smithson y Verkuilen (2006), and model proposed in Papke y Wooldridge (2008) which extends model presented in Papke y Wooldridge (1996) to panel data. Results showed better performance of Beta Inflated than Beta Transformed and Papke models in terms of bias and RMSE.

An application of Beta Inflated mean regression model was conducted to real data. This data consisted of utilization of credit line as response variable and an indicator of cash advance balance greater than zero, an indicator of high-risk individual and standardized individual age as covariates. Performance of Beta Inflated mixed-effects regression (BInf) was compared against Zero-One Beta Inflated mixed-effects regression (ZOIB) as proposed by Liu y Kong (2015) which is a two-part model. Results showed a lower standard deviation for fixed effects ω and δ, and random effects w_i and d_i in the Beta Inflated mean regression. Fixed effects β have similar standard deviation in both models, and random effects b_i are slightly lower in ZOIB regression model. Also, information criteria indicates that BInf regression model outperforms ZOIB regression model.

Stan code used for both simulation study and application to real data can be found in Appendix A.

6.2 Suggestions for future studies

For future studies it would be interesting to consider an asymmetric link function instead of logit in the Beta Inflated mean regression model with mixed effects. Furthermore, considering a multivariate fractional response rather than an univariate fractional response would extend the model revised in this article.

Appendix A

RStan code

A.1 Simulation study code

The following code in Stan implements the Beta Inflated mean regression incorporating fixed and random effects to parameters α_0, α_1 and γ.

```
functions
{
	real betainf_lpdf(real x, real gama, real alfa0, real alfa1, real phi)
	{
		real prob;
		real lprob;

		real mu;
		real a;
		real b;

		mu = (gama*(1-alfa1))/(1-alfa0*(1-gama)-alfa1*gama);
		a = mu*phi;
		b = (1-mu)*phi;

		if (x == 0)
			prob = alfa0*(1-gama);
		else if (x == 1)
			prob = alfa1*gama;
		else
		{
			prob = (1-alfa0*(1-gama)-alfa1*gama);
			prob = prob*exp(beta_lpdf(x|a,b));
		}

		lprob = log(prob);

		return(lprob);
	}
}

data
{
	int N; // the number of observations
	int M; // the number of subjects
	int K1; // the number of columns in the model matrix of fixed effects
```

```
        int K2; // the number of columns in the model matrix of random effects
        real y[N]; // the response
        matrix[N,K1] X; // the model matrix for fixed effects
        matrix[N,K2] Z; // the model matrix for random effects
        int id[N]; // vector with corresponding identification number
        matrix[K2,K2] sigmaw0; // scale covariance matrix for inverse wishart prior (alfa0)
        matrix[K2,K2] sigmad0; // scale covariance matrix for inverse wishart prior (alfa1)
        matrix[K2,K2] sigmab0; // scale covariance matrix for inverse wishart prior (gamma)
}

parameters
{
        vector[K1] omega; // fixed effects (alfa0)
        matrix[M,K2] wi; // random effects (alfa0)
        cov_matrix[K2] sigmaw; // variance-covariance matrix of random effects (alfa0)

        vector[K1] delta; // fixed effects (alfa1)
        matrix[M,K2] di; // random effects (alfa1)
        cov_matrix[K2] sigmad; // variance-covariance matrix of random effects (alfa1)

        vector[K1] beta; // fixed effects (gamma)
        matrix[M,K2] bi; // random effects (gamma)
        cov_matrix[K2] sigmab; // variance-covariance matrix of random effects (gamma)

        real<lower=0> phi; // precision
}

transformed parameters
{
        vector[N] linpred_alfa0;
        vector[N] alfa0_est;

        vector[N] linpred_alfa1;
        vector[N] alfa1_est;

        vector[N] linpred_gama;
        vector[N] gama_est;

        linpred_alfa0 = X*omega; // calculating linear predictor with only fixed effects (alfa0)
        linpred_alfa1 = X*delta; // calculating linear predictor with only fixed effects (alfa1)
        linpred_gama = X*beta; // calculating linear predictor with only fixed effects (gamma)

        for(i in 1:N)
        {
                for(j in 1:K2)
                {
                        // adding random effects to linear predictor (alfa0)
                        linpred_alfa0[i] = linpred_alfa0[i] + Z[i,j]*wi[id[i],j];

                        // adding random effects to linear predictor (alfa1)
                        linpred_alfa1[i] = linpred_alfa1[i] + Z[i,j]*di[id[i],j];
```

```
                // adding random effects to linear predictor (gamma)
                linpred_gama[i] = linpred_gama[i] + Z[i,j]*bi[id[i],j];
            }

            alfa0_est[i] = inv_logit(linpred_alfa0[i]);
            alfa1_est[i] = inv_logit(linpred_alfa1[i]);
            gama_est[i] = inv_logit(linpred_gama[i]);
        }
}

model
{
        // definition of priors

        sigmaw ~ inv_wishart(5,sigmaw0); // covariance matrix (alfa0)
        sigmad ~ inv_wishart(5,sigmad0); // covariance matrix (alfa1)
        sigmab ~ inv_wishart(5,sigmab0); // covariance matrix (gamma)

        for(j in 1:M)
        {
                wi[j,] ~ multi_normal(rep_vector(0,K2),sigmaw); // random effects (alfa0)
                di[j,] ~ multi_normal(rep_vector(0,K2),sigmad); // random effects (alfa1)
                bi[j,] ~ multi_normal(rep_vector(0,K2),sigmab); // random effects (gamma)
        }

        for(i in 1:K1)
        {
                omega[i] ~ normal(0,10000); // fixed effects (alfa0)
                delta[i] ~ normal(0,10000); // fixed effects (alfa1)
                beta[i] ~ normal(0,10000); // fixed effects (gamma)
        }

        phi ~ gamma(50,1); // precision parameter of betainf

        // distribution of dependent variable
        for(j in 1:N)
        {
                y[j] ~ betainf(gama_est[j], alfa0_est[j], alfa1_est[j], phi);
        }
}
```

A.2 Application code

The following code in Stan implements Beta Inflated mean regression incorporating fixed and random effects to parameter γ and considering all remaining parameters as constant for all observations.

```
functions
{
        real betainf_lpdf(real x, real gama, real alfa0, real alfa1, real phi)
        {
                real prob;
                real lprob;

                real mu;
                real a;
                real b;

                mu = (gama*(1-alfa1))/(1-alfa0*(1-gama)-alfa1*gama);
                a = mu*phi;
                b = (1-mu)*phi;

                if (x == 0)
                        prob = alfa0*(1-gama);
                else if (x == 1)
                        prob = alfa1*gama;
                else
                {
                        prob = (1-alfa0*(1-gama)-alfa1*gama);
                        prob = prob*exp(beta_lpdf(x|a,b));
                }

                lprob = log(prob);

                return(lprob);
        }
}

data
{
        int N; // the number of observations
        int M; // the number of subjects
        int K11; // the number of columns in the model matrix of fixed effects (gamma and alfa1)
        int K12; // the number of columns in the model matrix of fixed effects (alfa0)
        int K2; // the number of columns in the model matrix of random effects
        real y[N]; // the response
        matrix[N,K11] X1; // the model matrix for fixed effects (gamma and alfa1)
        matrix[N,K12] X2; // the model matrix for fixed effects (alfa0)
        matrix[N,K2] Z; // the model matrix for random effects
        int id[N]; // vector with corresponding identification number
        matrix[K2,K2] sigmaw0; // scale covariance matrix for inverse wishart prior
        matrix[K2,K2] sigmad0; // scale covariance matrix for inverse wishart prior
        matrix[K2,K2] sigmab0; // scale covariance matrix for inverse wishart prior
}
```

```
parameters
{
	vector[K12] omega; // fixed effects
	matrix[M,K2] wi; // random effects
	cov_matrix[K2] sigmaw; // variance-covariance matrix of random effects

	vector[K11] delta; // fixed effects
	matrix[M,K2] di; // random effects
	cov_matrix[K2] sigmad; // variance-covariance matrix of random effects

	vector[K11] beta; // fixed effects
	matrix[M,K2] bi; // random effects
	cov_matrix[K2] sigmab; // variance-covariance matrix of random effects

	real<lower=0> phi; // precision
}

transformed parameters
{
	vector[N] linpred_omega;
	vector[N] alfa0_est;

	vector[N] linpred_delta;
	vector[N] alfa1_est;

	vector[N] linpred_beta;
	vector[N] gama_est;

	linpred_omega = X2*omega; // calculating linear predictor with only fixed effects
	linpred_delta = X1*delta; // calculating linear predictor with only fixed effects
	linpred_beta = X1*beta; // calculating linear predictor with only fixed effects

	for(i in 1:N)
	{
		for(j in 1:K2)
		{
			// adding random errors to linear predictor

			linpred_omega[i] = linpred_omega[i] + Z[i,j]*wi[id[i],j];
			linpred_delta[i] = linpred_delta[i] + Z[i,j]*di[id[i],j];
			linpred_beta[i] = linpred_beta[i] + Z[i,j]*bi[id[i],j];
		}

		alfa0_est[i] = inv_logit(linpred_omega[i]);
		alfa1_est[i] = inv_logit(linpred_delta[i]);
		gama_est[i] = inv_logit(linpred_beta[i]);
	}
}

model
{
```

```
        // definition of priors

        sigmaw ~ inv_wishart(5,sigmaw0); // covariance matrix
        sigmad ~ inv_wishart(5,sigmad0); // covariance matrix
        sigmab ~ inv_wishart(5,sigmab0); // covariance matrix

        for(j in 1:M)
        {
                wi[j,] ~ multi_normal(rep_vector(0,K2),sigmaw); // random effects
                di[j,] ~ multi_normal(rep_vector(0,K2),sigmad); // random effects
                bi[j,] ~ multi_normal(rep_vector(0,K2),sigmab); // random effects
        }

        for(i in 1:K11)
        {
                beta[i] ~ normal(0,10000); // fixed effects
                delta[i] ~ normal(0,10000); // fixed effects
        }

        for(i in 1:K12)
        {
                omega[i] ~ normal(0,10000); // fixed effects
        }

        phi ~ gamma(0.0001,0.0001); // precision parameter of betainf

        // distribution of dependent variable
        for(j in 1:N)
        {
                y[j] ~ betainf(gama_est[j], alfa0_est[j], alfa1_est[j], phi);
        }
}
```

Bibliography

Amemiya, T. (1981). Qualitative response models: A survey, *Journal of economic literature* **19**(4): 1483–1536.

Bayes, C. L., Bazán, J. L., García, C. et al. (2012). A new robust regression model for proportions, *Bayesian Analysis* **7**(4): 841–866.

Bayes, C. L. y Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables, *Journal of Applied Statistics* **43**: 1814–1830.

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

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. y Ferrari, S. L. (2013). Mixed beta regression: a bayesian perspective, *Computational Statistics & Data Analysis* **61**: 137–147.

Galvis, D. M., Bandyopadhyay, D. y Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data, *Statistics in medicine* **33**(21): 3759–3771.

Gelman, A., Hwang, J. y Vehtari, A. (2014). Understanding predictive information criteria for bayesian models, *Statistics and Computing* **24**(6): 997–1016.

Hoffman, M. D. y Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo., *Journal of Machine Learning Research* **15**(1): 1593–1623.

Liu, F. y Kong, Y. (2015). zoib: an r package for bayesian inference for beta regression and zero/one inflated beta regression, *RJ* **7**: 34–51.

Ospina, R. y Ferrari, S. L. (2010). Inflated beta distributions, *Statistical Papers* **51**(1): 111–126.

Papke, L. E. y Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates, *Journal of Applied Econometrics* **11**(6): 619–632.
URL: *https://dx.doi.org/10.1002/(SICI)1099-1255(199611)11:6¡619::AID-JAE418¿3.0.CO;2-1*

Papke, L. y Wooldridge, J. (2008). Panel data methods for fractional response variables with an application to test pass rates, *Journal of Econometrics* **145**(1-2): 121–133.

Ramalho, J. J. y da Silva, J. V. (2009). A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms, *Quantitative Finance* **9**(5): 621–636.

Smithson, M. y Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables., *Psychological methods* **11**(1): 54.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4): 583–639.

Wang, J. y Luo, S. (2016). Augmented beta rectangular regression models: A bayesian perspective, *Biometrical Journal* **58**(1): 206–221.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(Dec): 3571–3594.