Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 2, 304–322
https://doi.org/10.1214/18-BJPS423

Bayesian modeling and prior sensitivity analysis for zero–one augmented beta regression models with an application to psychometric data

Danilo Covaes Nogarotto[a], **Caio Lucidius Naberezny Azevedo**[a] **and Jorge Luis Bazán**[b]

[a]*Campinas State University*
[b]*University of São Paulo*

Abstract. The interest on the analysis of the zero–one augmented beta regression (ZOABR) model has been increasing over the last few years. In this work, we developed a Bayesian inference for the ZOABR model, providing some contributions, namely: we explored the use of Jeffreys-rule and independence Jeffreys prior for some of the parameters, performing a sensitivity study of prior choice, comparing the Bayesian estimates with the maximum likelihood ones and measuring the accuracy of the estimates under several scenarios of interest. The results indicate, in a general way, that: the Bayesian approach, under the Jeffreys-rule prior, was as accurate as the ML one. Also, different from other approaches, we use the predictive distribution of the response to implement Bayesian residuals. To further illustrate the advantages of our approach, we conduct an analysis of a real psychometric data set including a Bayesian residual analysis, where it is shown that misleading inference can be obtained when the data is transformed. That is, when the zeros and ones are transformed to suitable values and the usual beta regression model is considered, instead of the ZOABR model. Finally, future developments are discussed.

1 Introduction

In many practical situations, we find the problem of analyzing variables that take values in the $(0, 1)$ interval, as percentages, proportions, rates or fractions. To analyze bounded response variables, the main developed model was the beta regression (BR) model, see Ferrari and Cribari-Neto (2004). It is currently a fairly consolidated model including some extensions as the mixed beta model and beta-mixture model. Also, residual analysis and model comparison are well developed. The literature is extensive on these models, among which we can cite the works of Ferrari and Cribari-Neto (2004), Paolino (2001), Smithson and Verkuilen (2006), Cribari-Neto and Zeiles (2010), which use the maximum likelihood method for parameter estimation, and under a Bayesian approach, the works of Buckley (2003), Branscum, Johnson and Thurmond (2007), Figueroa-Zuñiga, Arellano-Valle and Ferrari (2013), Cepeda-Cuervo et al. (2016).

In addition, when it is possible to observe zeros and/or ones with positive probability, we have the so called augmented data sets. A correspondent augmented statistical model is then commonly proposed for this case, see, for example, Galvis, Bandyopadhyay and Lachos (2014). In this work, we prefer using the term "augmented" instead of "inflated", since the zero and one values do not belong to the original support, that is the interval $(0, 1)$. Also, unless the opposite is stated, the term "augmented" will refer to the presence of discrete values, indicating an augmented observation (the values 0 and 1). Correspondent zero–one (or zero and one) augmented beta regression (ZOABR) models have been proposed in the literature in

Key words and phrases. Augmented beta regression, Bayesian inference, Jeffreys prior, MCMC methods, residual analysis.

Received July 2017; accepted November 2018.

Ospina and Ferrari (2012), Wieczorek and Hawala (2011) and Bayes and Valdivieso (2016). Other alternatives include the approaches developed by Papke and Wooldridge (1996), Tobit models (see, for example, Maddala, 1983) and a family of two part models as those introduced in Kieschnick and McCullough (2003), Ramalho and Silva (2009) and Stavrunova and Yerokhin (2012). Also, several mechanisms of statistical testing and detection of model misspecification to the ZOABR model were studied: see, for example, Pereira and Cribari-Neto (2014) and Souza et al. (2016) and extensions to spatial data proposed by Parker, Bandyopadhyay and Slate (2014). Software is available in Swearingen, Castro and Bursac (2012) and Liu and Kong (2015).

Even though in some works the Bayesian paradigm was considered for the ZOABR model, to the best of our knowledge, none of the works in the literature: explored a sensitivity analysis to prior specifications under the Bayesian paradigm, considering Jeffreys-rule and independence Jeffreys priors and performed a comparison with the maximum likelihood estimation under several scenarios of interest.

In this work, we developed Bayesian inference for the ZOABR model through MCMC algorithms, providing new contributions based in a extensive sensitivity study of prior choice. That is, considering the Jeffreys-rule and independence Jeffreys priors, we compared the Bayesian estimates with the maximum likelihood ones and we measured the accuracy of the estimates under several scenarios of interest. These scenarios correspond to the combination of the levels of some factors such as: the number of subjects (sample size), the number of covariates, the degree of variability of the data and the proportion of augmented (zero and one) observations. The results indicated, in a general way, that: the Bayesian estimates, under the Jeffeys-rule prior, are as accurate as the ML estimates, the higher the sample size, the more accurate the estimates, the opposite occurs with the variability, and the Bayesian paradigm is a feasible alternative to the ML approach (in terms of computational time and flexibility). More comments are provided throughout the paper.

To illustrate the Bayesian paradigm under the Jeffreys-rule prior, an application was performed. The data analyzed corresponds to a psychometric study of risk perception of living close to a nuclear plant, see Carlstrom, Woodward and Palmer (2000), when the value of the response provided in the interval [0, 1] is higer, the higher is the risk perceived by the subject. We concluded that the ZOABR model fitted very well to this data set and is more suitable than other competing models, which are the zero augmented beta regression (ZABR) model, the one augmented beta regression (OABR) model, both defined by Ospina and Ferrari (2012), and the beta regression model (BR) model.

The remainder of the paper is organized as follows. In Section 2, we present the ZOABR model and the correspondent augmented likelihood. In Section 3, we discuss about the prior choice, the full conditional distributions and the MCMC algorithm. In Section 4, we present the simulation studies, including a prior sensitivity analysis. In Section 5, we present the analysis of a psychometric data set using the Bayesian procedure, including a residual analysis to evaluate the goodness of fit of the model as well as a comparison with other competing models. Finally, in Section 6, we present some discussion and suggestions for future research.

2 The ZOABR model

The ZOABR model is based on the zero–one augmented beta distribution (ZOABD) with parameters $(\delta, \gamma, \mu, \phi)^t$, which according to Ospina and Ferrari (2010), is defined as:

$$\text{ZOABD}(\delta, \gamma, \mu, \phi) = \begin{cases} \delta\gamma^y(1-\gamma)^{1-y} & \text{if } y \in \{0, 1\}, \\ (1-\delta)h(y|\mu, \phi) & \text{if } y \in (0, 1), \end{cases}$$

where

$$h(y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\mathbb{1}_{(0,1)}(y), \tag{2.1}$$

that is, a beta distribution parameterized by its mean (μ) and its precision parameter (ϕ), as in Ferrari and Cribari-Neto (2004), with $\mathbb{1}(\cdot)$ being the usual indicator function. Furthermore, δ is the probability of the variable assumes a discrete value (zero or one), that is, assumes extreme augmented values, and γ is the conditional probability of the variable assuming the value one, given that the observation belongs to the discrete part.

The ZOABR model, using a Bayesian hierarchical representation, that is, considering the conditional distribution of the response given the parameters and the prior distribution, is defined by considering a set of random variables, let us say $Y_1,\ldots,Y_n$, such that

$$Y_t|\boldsymbol{\theta}\overset{\text{i.i.d}}{\sim}\text{ZOABD}(\boldsymbol{\theta}),\quad \boldsymbol{\theta}=(\delta,\gamma,\boldsymbol{\beta}^t,\phi)^t, t=1,\ldots,n, \tag{2.2}$$

with $\ln(\frac{\mu_t}{1-\mu_t})=\boldsymbol{x}_t^t\boldsymbol{\beta}$, $\boldsymbol{x}_t^t=(1,x_{t1},\ldots,x_{t(p-1)})$ and $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_{p-1})^t$, where the specification of prior distribution of the parameters is detailed in Section 3.1. That is, we focus on the modeling of the mean of the continuous part through a logistic link with a linear predictor. However, other link functions can be considered as the probit, skew normal or skew Student-t. In the ZOABR model, we distinguish two type of parameters. While that $(\boldsymbol{\beta}^t,\phi)^t$ are parameters associated with the data distribution, $(\delta,\gamma)^t$ are parameters associated with the augmentation. Notice also that no regression structure is adopted to ϕ neither to $(\delta,\gamma)^t$.

If $\delta=0$, we have the beta regression model proposed by Ferrari and Cribari-Neto (2004), namely the BR model here. If $\delta\neq 0$ and $\gamma=0$ or $\gamma=1$ we have, respectively, the zero augmented beta regression (ZABR) model and the one augmented beta regression (OABR) model, both defined by Ospina and Ferrari (2012).

In order to facilitate the obtaining of the posterior distributions and the implementation of the MCMC algorithm, let us define the following augmented observable (indicator) variable z_t which assumes the value 0 if $y_t\in(0,1)$ or 1 if $y_t\in\{0,1\}$. Therefore, the joint distribution of $(y_t,z_t)^t$ given $\boldsymbol{\theta}$ can be written as

$$h(y_t,z_t|\boldsymbol{\theta})=h(y_t|\boldsymbol{\beta},\phi)^{1-z_t}\gamma^{y_tz_t}(1-\gamma)^{z_t(1-y_t)}\delta^{z_t}(1-\delta)^{1-z_t}\mathbb{1}_{\{y_t,z_t\}},$$

where $\mathbb{1}_{\{y_t,z_t\}}=\mathbb{1}_{(0,1)}(y_t)\mathbb{1}_{\{0\}}(z_t)+\mathbb{1}_{\{0,1\}}(y_t)\mathbb{1}_{\{1\}}(z_t)$ and $h(y_t|\boldsymbol{\beta},\phi)$ is given by (2.1), with μ replaced by $\mu_t=\frac{e^{x_t^t\beta}}{1+e^{x_t^t\beta}}$.

Finally the likelihood related to the joint distribution of $(\boldsymbol{y}^t,\boldsymbol{z}^t)^t$ ($\boldsymbol{y}=(y_1,\ldots,y_n)^t$ and $\boldsymbol{z}=(z_1,\ldots,z_n)^t$) is given by

$$\begin{aligned}L(\boldsymbol{\theta})&=h(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})=\prod_{t=1}^{n}h(y_t,z_t|\boldsymbol{\theta})\\&\propto\prod_{t=1}^{n}\left\{\frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)}y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1}\right\}^{1-z_t}\\&\quad\times\gamma^{\sum_{t=1}^{n}y_tz_t}(1-\gamma)^{\sum_{t=1}^{n}z_t(1-y_t)}\delta^{\sum_{t=1}^{n}z_t}(1-\delta)^{n-\sum_{t=1}^{n}z_t}\\&=L(\boldsymbol{\beta},\phi)L(\gamma)L(\delta),\end{aligned} \tag{2.3}$$

where

$$L(\boldsymbol{\beta},\phi)=\prod_{t=1}^{n}\left\{\frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)}y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1}\right\}^{1-z_t}, \tag{2.4}$$

$$L(\gamma)=\gamma^{\sum_{t=1}^{n}y_tz_t}(1-\gamma)^{\sum_{t=1}^{n}z_t(1-y_t)}, \tag{2.5}$$

and

$$L(\delta) = \delta^{\sum_{t=1}^{n} z_t}(1-\delta)^{n-\sum_{t=1}^{n} z_t}. \tag{2.6}$$

Therefore, the likelihood is partially separable. In the next section, we present a discussion about the prior choice, posterior distribution and the related MCMC algorithms.

3 Bayesian inference and MCMC algorithms

3.1 Prior and posterior distributions

For the parameters associated to the data distribution, that is, $(\boldsymbol{\beta}^t, \phi)^t$, we explore the use of the following four prior distributions: (1) usual priors considered for the generalized linear models (see Dey, Ghosh and Mallick, 2000), (2) improper (non-informative) priors, (3) the Jeffreys-rule prior and (4) the independence Jeffreys prior, which are showed in Table 1. For the augmentation parameters, that is, $(\delta, \gamma)^t$, only independent beta distributions are considered, in all cases. Priors 1, 2 and 4 are combined with augmentation priors, using the following structure: $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\phi)p(\delta)p(\gamma)$, that is, we assume that the parameters are mutually independent. More specifically, for prior 1, we consider that: $\boldsymbol{\beta} \sim N(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ and $\phi \sim G(\eta, \lambda)$, where $G(\eta, \lambda)$ stands for a gamma distribution with $E(\phi) = \eta\lambda$ and $\mathcal{V}(\phi) = \eta\lambda^2$. Concerning to prior 4, the former prior for $\boldsymbol{\beta}$ is adopted but $p(\phi) \propto \mathbb{1}_{(0,\infty)}(\phi)$ is assumed. Prior 3 is combined with the priors for augmentation parameters, that is: $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}, \phi)p(\delta)p(\gamma)$ and the Fisher Information is necessary. Due to the separability of the likelihood, see Equation (2.3), the Fisher Information corresponds to the following block diagonal matrix:

$$\boldsymbol{K}(\delta, \gamma, \boldsymbol{\beta}, \phi) = \begin{pmatrix} K(\delta, \gamma) & 0 \\ 0 & \boldsymbol{K}(\boldsymbol{\beta}, \phi) \end{pmatrix} = \begin{pmatrix} \boldsymbol{K}_{\delta\delta} & 0 & 0 & 0 \\ 0 & K_{\gamma\gamma} & 0 & 0 \\ 0 & 0 & \boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{K}_{\boldsymbol{\beta}\phi} \\ 0 & 0 & \boldsymbol{K}_{\phi\boldsymbol{\beta}} & K_{\phi\phi} \end{pmatrix},$$

where $K_{\delta\delta} = 1/[\delta(1-\delta)]$, $K_{\gamma\gamma} = \delta/[\gamma(1-\gamma)]$, $\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \phi^2 \boldsymbol{X}^T\{\boldsymbol{\Delta}_{(0,1)}^{-1}\boldsymbol{T}\boldsymbol{W}\boldsymbol{T}\}\boldsymbol{X}$, $\boldsymbol{K}_{\boldsymbol{\beta}\phi} = \boldsymbol{K}_{\phi\boldsymbol{\beta}}^T = \boldsymbol{X}^T\boldsymbol{\Delta}_{(0,1)}^{-1}\boldsymbol{T}\boldsymbol{c}$, $K_{\phi\phi} = \mathrm{tr}(\boldsymbol{\Delta}_{(0,1)}^{-1}\boldsymbol{D})$, and $\boldsymbol{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$, $\boldsymbol{D} = \mathrm{diag}\{d_1, \ldots, d_n\}$, $\boldsymbol{c} = (c_1, \ldots, c_n)^T$, $w_t = \psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)$, $d_t = (1-\mu_t)^2\psi'((1-\mu_t)\phi) + \mu_t^2\psi'(\mu_t\phi) - \psi'(\phi)$, $c_t = \phi[\mu_t w_t + \psi'((1-\mu_t)\phi)]$, $\boldsymbol{T} = \mathrm{diag}\{\mathrm{d}\mu_1/\mathrm{d}\eta_1, \ldots, \mathrm{d}\mu_n/\mathrm{d}\eta_n\} = \mathrm{diag}\{\frac{1}{g'(\mu_1)}, \frac{1}{g'(\mu_2)}, \ldots, \frac{1}{g'(\mu_n)}\}$, $\boldsymbol{\Delta}_{(0,1)} = \mathrm{diag}\{1/(1-\delta), \ldots, 1/(1-\delta)\}$ being a $n \times n$ diagonal matrix and $\psi(\cdot)$ is the digama function, that is: $\psi(\tau) = \frac{d\ln\Gamma(\tau)}{d\lambda} = \frac{\Gamma'(\tau)}{\Gamma(\tau)}$, $\tau > 0$. More details can be found in Ospina and Ferrari (2012).

Table 1 *Summary of the sets of prior distributions*

Prior distribution	$\boldsymbol{\beta}$	ϕ
Usual—1	$N_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$	$G(\eta, \lambda)$
Independence Jeffreys—2	$\propto \sqrt{\lvert\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}\rvert}$	$\propto \sqrt{\lvert\boldsymbol{K}_{\phi\phi}\rvert}$
Jeffreys—3	$\propto \sqrt{\lvert\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}K_{\phi\phi} - \boldsymbol{K}_{\boldsymbol{\beta}\phi}\boldsymbol{K}_{\phi\boldsymbol{\beta}}\rvert}$	
Improper—4	$N_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$	$\propto 1$
	δ	γ
–	$\mathrm{beta}(a_1, b_1)$	$\mathrm{beta}(a_2, b_2)$

Then, the Jeffreys-rule prior and the independence Jeffreys prior for $(\boldsymbol{\beta}^t, \phi)^t$ are given respectively by: $p^J(\boldsymbol{\beta}, \phi) \propto \sqrt{|\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} K_{\phi\phi} - \boldsymbol{K}_{\boldsymbol{\beta}\phi}\boldsymbol{K}_{\phi\boldsymbol{\beta}}|}$, and $p^{IJ}(\boldsymbol{\beta}, \phi) = p^J(\boldsymbol{\beta})p^J(\phi) \propto \sqrt{|\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}|}\sqrt{|K_{\phi\phi}|}$. In all of those sets of prior distributions, we assume additionally that $\gamma \sim \text{beta}(a_1, b_1)$ and $\delta \sim \text{beta}(a_2, b_2)$ are mutually independent. It is possible to prove that, if $a_1 = b_1 = a_2 = b_2 = 1$, we obtain the independence Jeffreys prior for $(\gamma, \delta)^t$.

An important issue when using improper priors is to ensure that the joint posterior distribution (and consequently, the marginal posterior distributions) is proper, see Gelfand and Sahu (1999). Indeed, we did not prove these results, but the numerical results (good parameter recovery, as we shown ahead), suggest that the posterior distributions exist. However, this issue certainly deserves more investigation in a future paper.

The joint posterior distribution, under the set of priors 1, 2 and 4 is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}) &\propto L(\boldsymbol{\theta})p(\boldsymbol{\beta})p(\phi)p(\delta)p(\gamma) \\ &= L(\boldsymbol{\beta}, \phi)L(\gamma)L(\delta)p(\boldsymbol{\beta})p(\phi)p(\delta)p(\gamma) \end{aligned} \tag{3.1}$$

for the set 2 we have $p(\boldsymbol{\beta})p(\phi) = p^J(\boldsymbol{\beta})p^J(\phi)$, whereas, under the set 3, we have $p(\boldsymbol{\beta}, \phi) = p^J(\boldsymbol{\beta}, \phi)$ in the place of $p(\boldsymbol{\beta})p(\phi)$.

Due to the separability of the likelihood and the structure of the sets of prior distributions adopted, we have that the (marginal) posterior distributions of δ and γ can be obtained analytically. Indeed they are given, respectively, by

$$\delta|\boldsymbol{y}, \boldsymbol{z} \sim \text{beta}(\widehat{a}_1, \widehat{b}_1) \quad \text{and} \quad \gamma|\boldsymbol{y}, \boldsymbol{z} \sim \text{beta}(\widehat{a}_2, \widehat{b}_2), \tag{3.2}$$

where: $\widehat{a}_1 = \sum_{t=1}^n z_t + a_1$; $\widehat{b}_1 = n - \sum_{t=1}^n z_t + b_1$; $\widehat{a}_2 = \sum_{t=1}^n y_t z_t + a_2$; $\widehat{b}_2 = \sum_{t=1}^n z_t(1 - y_t) + b_2$, where $\text{beta}(c, d)$ is the notation of the beta distribution with the usual parameterization.

Therefore, it is possible to make exact Bayesian inference for these parameters in the sense that closed expressions are available for the EAP (expectation a posteriori), MAP (maximum a posteriori) and PSD (posterior standard deviation). To obtain CI (credibility intervals) and HPD intervals it is necessary to employ numerical methods, but this can be easily done through the functions *qbeta* and *hpd*, available in the R program, see R Development Core Team (2015). Since no closed expressions for the marginal posterior distributions for $\boldsymbol{\beta}$ and ϕ are available, more complex numerical methods, as MCMC algorithms, should be employed, as we present in the next subsection.

3.2 Full conditional distributions and MCMC algorithms

For $\boldsymbol{\beta}$ and ϕ, is not possible to obtain the respective (marginal) posterior distributions. Therefore, numerical procedures, as MCMC algorithms, should be employed in order to obtain numerical approximations for them, see Gamerman and Lopes (2006). To implement these algorithms, the so called full conditional distributions are necessary.

They, respectively, for $\boldsymbol{\beta}$ and ϕ, are given by

$$\begin{aligned} p(\boldsymbol{\beta}|\phi, \delta, \gamma, \boldsymbol{y}, \boldsymbol{z}) &= p(\boldsymbol{\beta}|\phi, \boldsymbol{y}, \boldsymbol{z}) \propto L(\boldsymbol{\beta}, \phi)p(\boldsymbol{\beta}, \phi) \\ &= \prod_{t=1}^n \left\{ \frac{1}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi - 1}(1 - y_t)^{(1-\mu_t)\phi - 1} \right\}^{1-z_t} p(\boldsymbol{\beta}, \phi) \end{aligned}$$

and

$$\begin{aligned} p(\phi|\boldsymbol{\beta}, \delta, \gamma, \boldsymbol{y}, \boldsymbol{z}) &= p(\phi|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}) \propto L(\boldsymbol{\beta}, \phi)p(\boldsymbol{\beta}, \phi) \\ &= \prod_{t=1}^n \left\{ \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi - 1}(1 - y_t)^{(1-\mu_t)\phi - 1} \right\}^{1-z_t} p(\boldsymbol{\beta}, \phi). \end{aligned}$$

When $p(\boldsymbol{\beta}, \phi) = p(\boldsymbol{\beta})p(\phi)$ the above expressions reduce to $\propto L(\boldsymbol{\beta}, \phi)p(\boldsymbol{\beta})$ and $\propto L(\boldsymbol{\beta}, \phi)p(\phi)$, respectively. In both cases the full conditional distributions are unknown and it is necessary to use some auxiliary algorithm to sample from them. In this work, we consider the Metropolis-Hastings algorithm, see Gamerman and Lopes (2006). The related details can be found in the Supplementary Material (Section 1 of Nogarotto, Azevedo and Bazán, 2020).

Let $(\cdot)$ denote the set of all necessary parameters. Then, the steps of the Gibbs sampling scheme for the ZOABR model can be summarized as follows:

1. Start the algorithm by using suitable initial values.
   Repeat the steps 2–3:
2. Simulate $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|(\cdot)$.
3. Simulate ϕ from $\phi|(\cdot)$.
4. After the convergence, calculate Bayesian estimates of interest for δ and γ through (3.2).

For the other three models, that is, for BR, ZABR and OABR models the MCMC algorithm can be easily obtained, by skipping the parts related to the known parameters ($\delta = 0$, $\gamma = 1, \gamma = 0$) and simplifying the remaining expressions. Further details about these MCMC algorithms can be found in the Supplementary Material (Section 2 of Nogarotto, Azevedo and Bazán, 2020). A final comment is that one could assume a multivariate prior structure leading to a posterior dependency between $(\boldsymbol{\beta}^t, \phi)^t$ and $(\delta, \gamma)^t$. Therefore, in this case, their posterior distributions would be simulated within the MCMC algorithm, instead of the previously presented two steps approach. This could be developed in another work, even though we believe that the results would be quite similar.

3.3 Residual analysis

The residuals used here are based on those presented by Ferrari and Cribari-Neto (2004), Oliveira (2004), Ospina (2008), Ospina and Ferrari (2012) and Paulino, Turkman and Murteira (2003), named, standard residual (r_t), weighted standard residual ($rp_t^{(01)}$) and the deviance residual (rd_t). We considered the predictive distribution of the response (Y_t) to obtain the residuals.

Using the related MCMC valid samples of the predictive distribution of each observation, we calculate the the predictive mean ($\tilde{\mu_t}$) and the predictive variance ($V(\tilde{y_t})$). They correspond, respectively, to the sample mean and sample variance for each observation using the 1,000 simulated values (valid MCMC sample). In addition, the predictive log-likelihood ($l(\cdot)$), which corresponds to the likelihood calculate using the original observation (y_t) and the predictive mean ($\tilde{\mu_t}$) and variance ($V(\tilde{y_t})$), for each individual, was calculated.

The standard residual is defined as:

$$r_t = \frac{y_t - \tilde{\mu_t}}{\sqrt{V(\tilde{y_t})}}, \tag{3.3}$$

which is only appropriate for BR model. To the other three models, the so-called weighted standard residual should be used. It corresponds to the residual (3.3) suitably modified in order to consider the discrete part of the model. Let us define $\tilde{p_{0t}}$ and $\tilde{p_{1t}}$, the proportion of zeros and ones, respectively, based on the aforementioned simulated values of the predictive distribution. Therefore, using these quantities, it is possible to obtain two other residuals

$$r_t^{(0)} = \frac{\hat{\delta}(1-\hat{\gamma}) - \tilde{p_{0t}}}{\sqrt{\hat{\delta}(1-\hat{\gamma})(1-(\hat{\delta}(1-\hat{\gamma})))}} \tag{3.4}$$

and

$$r_t^{(1)} = \frac{\hat{\delta}\hat{\gamma} - \tilde{p_{1t}}}{\sqrt{\hat{\delta}\hat{\gamma}(1-\hat{\delta}\hat{\gamma})}}. \tag{3.5}$$

Therefore, the weighted standard residual, is given by:

$$rp_t^{(01)} = \hat{\delta}(1-\hat{\gamma})r_t^{(0)} + \hat{\delta}\hat{\gamma}r_t^{(1)} + (1-\hat{\delta})r_t. \tag{3.6}$$

Finally, the deviance residual (DR) is defined as:

$$rd_t = \text{sign}(y_t - \tilde{\mu_t})\sqrt{2\{\ln(y_t) - \ln(\tilde{\mu_t})\}}, \tag{3.7}$$

where

$$\ln(y_t) = \begin{cases} \ln(\hat{\delta}\hat{\gamma}) = \ln(\hat{\delta}) + \ln(\hat{\gamma}) & \text{if } y_t = 1, \\ \ln(\hat{\delta}(1-\hat{\gamma})) = \ln(\hat{\delta}) + \ln(1-\hat{\gamma}) & \text{if } y_t = 0, \\ \ln(h(y_t)) & \text{if } y_t \in (0,1) \end{cases}$$

and

$$\ln(\tilde{\mu_t}) = \begin{cases} \ln(\tilde{p_{1t}}) & \text{if } y_t = 1, \\ \ln(\tilde{p_{0t}}) & \text{if } y_t = 0, \\ \ln(h(\tilde{\mu_t})) & \text{if } y_t \in (0,1), \end{cases}$$

and $h(\cdot)$ is given by Equation (2.1). In summary, the residuals defined in Equation (3.3) were considered for the BR model. On the other hand, the residuals defined in (3.4), (3.5) and (3.6) are used in the models ZABR, OABR and ZOABR, respectively. Under a suitable fit of the (respective) model to the data, it is expected that the residuals present a random behavior, with no systematic pattern, homoscedasticity, and negligible autocorrelation. Under a large sample size, we also expect that the residuals follow, approximately, a standard normal distribution, see Ospina and Ferrari (2012).

4 Simulation studies

Several aspects of interest are explored in this section: we perform a sensitivity study concerning the prior choice of the parameters associated with the data distribution, described in Section 3 and we study and compare the frequentist properties of the Bayesian and maximum likelihood (ML) estimates considering different scenarios of interest. They are defined by crossing the levels of four factors of interest (with their respective levels within parentheses): sample size [n], (20, 50 and 200), number of covariates [p] (2 and 5), value of the precision parameter [ϕ] (20, 50, 200 and 1000) (remembering that the higher the value, the smaller the variance of the data) and the probability of an observation being discrete [δ] (0, 0.1, 0.3 and 0.5). Also, the impact of these factors, in the parameter recovery, was also measured. Notice that when $\delta = 0$, we have the BR model. In addition, when $\delta > 0$, we considered three situations, $\gamma = 0$ (ZABR model), $\gamma = 1$ (OABR model) and $\gamma = 0.5$ (ZOABR model). For the BR model, we have a total of $3 \times 2 \times 4 = 24$ situations, whereas for the other three models we have $3 \times 3 \times 2 \times 4 = 72$ situations. The values considered for the regression parameters were: $\beta_0 = -1.5$, $\beta_1 = 1.5$ and $\beta_0 = -3.0$, $\beta_1 = -1.5$, $\beta_2 = 0.0$, $\beta_3 = 1.5$, $\beta_4 = 3.0$ (when p = 2, we considered only $(\beta_0, \beta_1)^t$).

The respective computation codes were made in the R program, see R Development Core Team (2015), and are available, upon requests, from the authors. Based on the results provided by the usual methods for checking the convergence of the MCMC algorithms, using three parallel chains, it can be concluded that the chains mixed very well and the autocorrelations for a thinning of 9 iterations are negligible. In addition, from a burn-in of 1,000 iterations, and a total 10,000 simulated values was enough to obtain valid MCMC samples of size 1,000, for each parameter.

Table 2 *Hyperparameters for the prior distributions* (1) *and* (4)

Prior distribution	$\boldsymbol{\beta}$	ϕ
Usual-1	$N_p(\mathbf{0}, 25I_p)$	$G(500, 100{,}000)$
Usual-2	$N_p(\mathbf{0}, 25I_p)$	$G(500, 1{,}000{,}000)$
Improper	$N_p(\mathbf{0}, 25I_p)$	$\propto 1$

For each one of the 72 (or 24, in the case of BR model) situations NR = 100 replicas were generated, from model (2.2). Each covariate was simulated from a U(0, 1) distribution, in each replica, and their values were centered in their respective sample averages, in order to improve the convergence of the MCMC algorithm. The hyper parameters for the prior distributions (1) and (4) are presented in Table 2, which induce priors with moderate ($\boldsymbol{\beta}$) and large (ϕ) variances. For the parameters $(\gamma, \delta)^t$, we used $a_1 = b_1 = a_2 = b_2 = 1$. Notice that we have two sets of hyperparameters for the prior distribution named "Usual". The prior distribution for ϕ named "Usual-2" is flatter than that named "Usual-1". Therefore, we have five sets of Bayesian estimates and another related to the ML estimates. The proposal distributions, necessary to implement the Metropolis-Hastings algorithm, are presented in the Supplementary Material (Section 1 of Nogarotto, Azevedo and Bazán, 2020). The ML estimates and the respective asymptotic variances were obtained through the package *betareg* from the R program, see Cribari-Neto and Zeiles (2010). For details concerning the ML estimation, the reader is referred to Ferrari and Cribari-Neto (2004) and Ospina and Ferrari (2012). With these six sets of estimates, obtained in each replica, we calculated the usual statistics for measuring the accuracy of the estimates: bias, variance (Var), root mean squared error (RMSE) and absolute value of relative bias (AVRB). Let θ be the parameter of interest and $\widehat{\theta}_r$ be some estimate related to the replica r, and $\overline{\widehat{\theta}} = \sum_{r=1}^{R} \frac{\widehat{\theta}_r}{R}$. The adopted statistics are defined as: $\text{BIAS} = \overline{\widehat{\theta}} - \theta$, $\text{Var} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\theta}_r - \overline{\widehat{\theta}})^2$, $\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\theta - \widehat{\theta}_r)^2}$, $\text{AVRB} = \frac{|\overline{\widehat{\theta}} - \theta|}{|\theta|}$.

Additionally, for $\boldsymbol{\beta}$, the results were summarized over its components, by taking the respective average over them. The smaller is each one of these statistics, the more accurate is the estimate. Since our main interest, in terms of prior sensitivity analysis, lies on $\boldsymbol{\beta}$ and ϕ, we presented the results for them in Section 4.1 whereas for the other parameters $(\gamma, \delta)^t$ in Section 4.2.

4.1 Parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$

We focus on the results for the ZOABR model with five covariates (for more details see Nogarotto, 2013) and additional results are available from the authors upon request. Tables 3 and 4 present the respective AVRB; whereas, Figures 1, 2 and 3 present the respective RMSE. From an inspection of these results, we can conclude that in general, the Bayesian estimates under the Jeffreys-rule prior and the independence Jeffreys prior are the most accurate, especially under small sample sizes and concerning ϕ. This pattern is also observed for the other scenarios and models, which were not presented, indicating that the Bayesian estimates are as accurate as the maximum likelihood ones. Also, the smaller the variability and/or the higher the sample size is the more accurate the estimates are. Notice that the y-axis of Figure 3(b), for ϕ, is shown in logarithm scale, in order to improve its visualization.

For a given sample size and variability, the larger the number of covariates is, the less accurate the estimates are. For each scenario, the parameters were more accurate estimated according to the complexity of the model, that is, the estimates of the parameters of the BR model were more accurate compared with those related to the OABR/ZABR models which, in their turn, presented more accurate estimates than those associated with the ZOABR model.

Table 3 *Absolute values of relative bias (AVRB) associated to* $\boldsymbol{\beta}$ *for different values of* ϕ *and* δ *under* p = 5 *covariates for the ZOABR model*

ϕ	n	Usual-1	Usual-2	Independence Jeffreys	Jeffreys	Improper	ML
				$\delta = 0.1$			
50	20	6.2%	3.1%	2.7%	5.2%	4.3%	5.8%
	50	1.2%	0.6%	0.5%	1.7%	0.4%	0.9%
	200	0.3%	0.2%	0.2%	0.5%	0.2%	0.3%
200	20	1.9%	1.4%	1.2%	2.0%	1.7%	1.8%
	50	0.2%	0.3%	0.3%	0.6%	0.2%	0.2%
	200	0.3%	0.2%	0.2%	0.2%	0.3%	0.3%
				$\delta = 0.3$			
50	20	5.7%	1.5%	1.4%	16.5%	3.9%	5.7%
	50	1.2%	1.3%	1.3%	3.1%	1.3%	1.1%
	200	1.3%	0.9%	0.8%	0.5%	1.1%	1.2%
200	20	0.9%	1.0%	0.8%	5.3%	0.7%	1.2%
	50	0.9%	0.7%	0.7%	0.5%	0.7%	1.0%
	200	0.3%	0.4%	0.4%	0.4%	0.4%	0.4%
				$\delta = 0.5$			
50	20	8.5%	4.1%	4.4%	3.3%	6.2%	10.1%
	50	2.9%	2.0%	2.0%	3.7%	1.9%	2.7%
	200	0.7%	0.2%	0.2%	0.5%	0.3%	0.6%
200	20	4.9%	3.0%	4.5%	8.9%	4.6%	5.9%
	50	1.0%	0.7%	0.6%	0.6%	0.9%	1.0%
	200	0.7%	0.7%	0.7%	0.6%	0.7%	0.8%

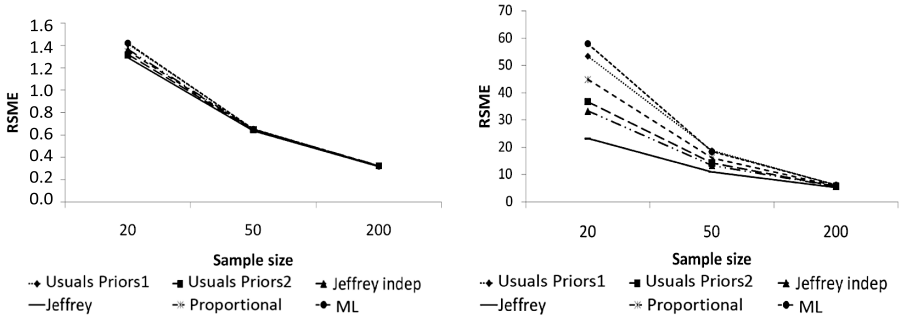
Also, it can be seen that as the sample size and the value of ϕ increase and the value of δ decreases, the smaller is the RMSE of $\boldsymbol{\beta}$. Also, notice that, for $\boldsymbol{\beta}$, the frequentist and Bayesian estimates are very similar (Table 3). However, for ϕ, the Bayesian estimates were more accurate, mainly under Jeffreys-rule prior and smaller sample sizes (Table 4). For this parameter, notice that most of the values of AVRB are larger for ML method, presenting, for example, 4,100% ($\delta = 0.5$, n = 20 and $\phi = 50$). This is expected, especially under small sample sizes, since the estimation of ϕ tends to be less accurate.

In Figure 4, we present the histograms of the marginal posterior distributions for a particular case (n = 50, p = 5, $\phi = 50$ and $\delta = 0.3$). In general, for each parameter, they are similar, regardless the adopted prior. They are even more similar as the sample size and ϕ increase and δ decreases. However, for the regression parameters, the posterior distributions induced by the Jeffreys-rule prior tend to present smaller variances. On the other hand, for ϕ, the smaller the sample size and the higher the value of δ are, the more concentrated around the respective true values the posterior distributions are, induced by the Jeffreys-rule prior, compared with the others. In conclusion, mainly for the precision parameter ϕ, Jeffreys-rule and the independence Jeffreys priors are more appropriate choices.

The spent time to run the algorithms varies according to the adopted prior, as well as the scenario and the estimation method considered. For instance, to estimate the parameters for one replica, considering n = 200, p = 5 and $\phi = 200$, the spent time (in seconds) for each method was: 8 (usual prior 1), 9 (usual prior 2/improper), 53 (independency Jeffreys prior), 332 (Jeffreys prior) and 1 (ML). In general, the spent time for the usual and improper priors were the same, and the largest observed spent time was related to Jeffreys prior and the independence Jeffreys prior (which lead to more complicated calculations, compared with the other priors). As expected, the ML method was the fastest.

Table 4 *Absolute values of relative bias (AVRB) associated to ϕ for different values of ϕ and δ under* $p = 5$ *covariates for the ZOABR model*

ϕ	n	Usual-1	Usual-2	Independence Jeffreys	Jeffreys	Improper	ML
				$\delta = 0.1$			
50	20	74.5%	36.1%	28.9%	19.9%	53.5%	76.9%
	50	26.1%	13.9%	11.4%	1.3%	18.6%	24.7%
	200	5.6%	3.1%	2.6%	0.3%	4.0%	5.2%
200	20	39.4%	25.0%	21.4%	14.2%	43.0%	71.3%
	50	10.1%	3.5%	2.7%	6.1%	8.3%	15.2%
	200	1.6%	0.1%	0.2%	2.3%	1.0%	2.5%
				$\delta = 0.3$			
50	20	130.0%	72.2%	60.2%	30.2%	104.6%	149.8%
	50	25.2%	9.9%	7.6%	9.6%	16.2%	23.5%
	200	8.5%	5.2%	4.4%	0.9%	6.7%	8.1%
200	20	55.4%	44.9%	43.7%	18.2%	79.8%	129.5%
	50	22.2%	14.3%	13.0%	0.3%	20.7%	31.3%
	200	4.5%	2.5%	2.2%	0.5%	3.8%	5.9%
				$\delta = 0.5$			
50	20	237.1%	284.0%	1,668.4%	541.3%	2,579.4%	4,100.0%
	50	48.5%	25.7%	21.0%	6.9%	34.6%	47.8%
	200	7.3%	2.5%	1.8%	3.1%	4.6%	6.7%
200	20	71.6%	73.1%	191.7%	37.2%	450.2%	902.0%
	50	29.3%	18.9%	16.5%	3.6%	30.2%	45.8%
	200	7.1%	4.4%	4.0%	0.2%	6.4%	9.4%

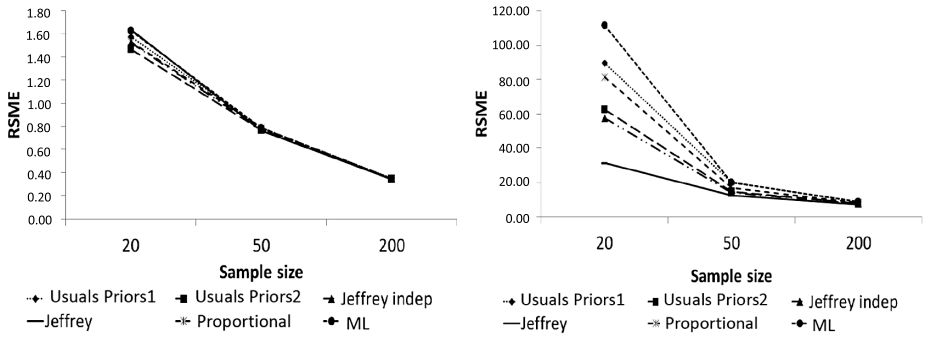


(a) RMSE - $\boldsymbol{\beta}$ (b) RMSE - ϕ

Figure 1 *RMSE of the* (a) $\boldsymbol{\beta}$ *and* (b) ϕ *for the ZOABR model under different sample sizes considering* $\phi = 50$, $p = 5$ *and* $\delta = 0.1$, *for parameter* $\boldsymbol{\beta}$ *and* ϕ *for the ZOABR model.*

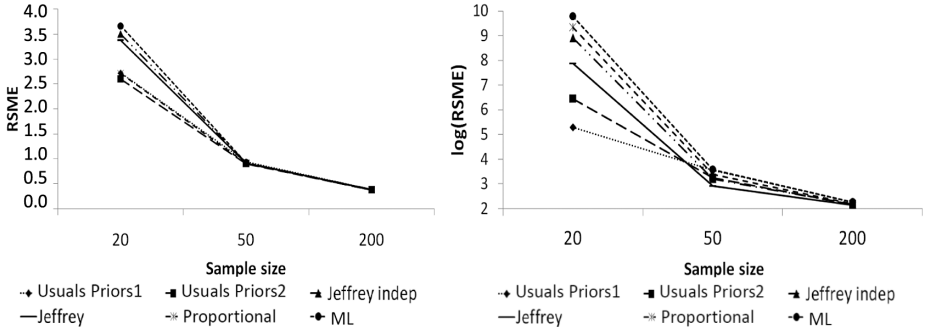
However, we suggest the use of the Bayesian paradigm for fitting the ZOABR model (and the related particular cases) as an alternative to the frequentist inference, since the spent time is not prohibitive (for the cases that we explored) and it is as accurate as the ML method. Of course, for larger data sets, the MCMC algorithms would require more computational time and in such cases, the using of MCMC algorithms would become less attractive.

In the Supplementary Material (Section 2 of Nogarotto, Azevedo and Bazán, 2020) are available the results of the simulation studies related to other scenarios, where small values (very large variability) of ϕ were considered, that is $\phi \in \{0.5, 1, 5\}$. The results indicated, when the Bayesian methods did not present numerical problems, that they present the most accurate results, as before. That is, for a real data analysis, when at least one Bayesian method



(a) RMSE - $\boldsymbol{\beta}$ (b) RMSE - ϕ

Figure 2 *RMSE for* (a) $\boldsymbol{\beta}$ *and* (b) ϕ *for ZOABR model under different sample sizes considering* $\phi = 50$, $\mathrm{p} = 5$ *and* $\delta = 0.3$, *for parameters* $\boldsymbol{\beta}$ *and* ϕ *for the ZOABR model.*



(a) RMSE - $\boldsymbol{\beta}$ (b) log(RMSE) - ϕ

Figure 3 *RMSE for* (a) $\boldsymbol{\beta}$ *and* $\log(RMSE)$ *for* (b) ϕ *for the ZOABR model under different sample sizes considering* $\phi = 50$, $\mathrm{p} = 5$ *and* $\delta = 0.5$, *for parameters* $\boldsymbol{\beta}$ *and* ϕ *for the ZOABR model.*

did not present any problem, it should be used, otherwise, the ML method should be considered.

4.2 Parameters δ and γ

In this section, we describe some results related to δ and γ. In this case, our goal was only to compare the Bayesian (using the priors described in Section 4) and ML estimates, instead of performing a sensitivity analysis.

From Tables 5 and 6 it is possible to see that Bayesian and ML results were similar, except for δ (when $\delta = 0.1$), where the ML estimate are better. However, there is a slight advantage for the Bayesian approach when estimating γ. Also, the higher the value of δ and the sample size, the more accurate the estimates. Since the higher the value of δ is, the more augmented (discrete) observations are observed, it is expected to obtain more accurate estimates. Also, since the likelihood is separable, $(\boldsymbol{\beta}^t, \phi)^t$ from $(\delta, \gamma)^t$, it is expected that the factors related to the continuous part do not affect the estimate of the parameters of the discrete part. Therefore, again, the Bayesian estimates are as good as the frequentist ones. In conclusion, also, in this case, we suggest the use of Bayesian inference to fit the ZOABR model.

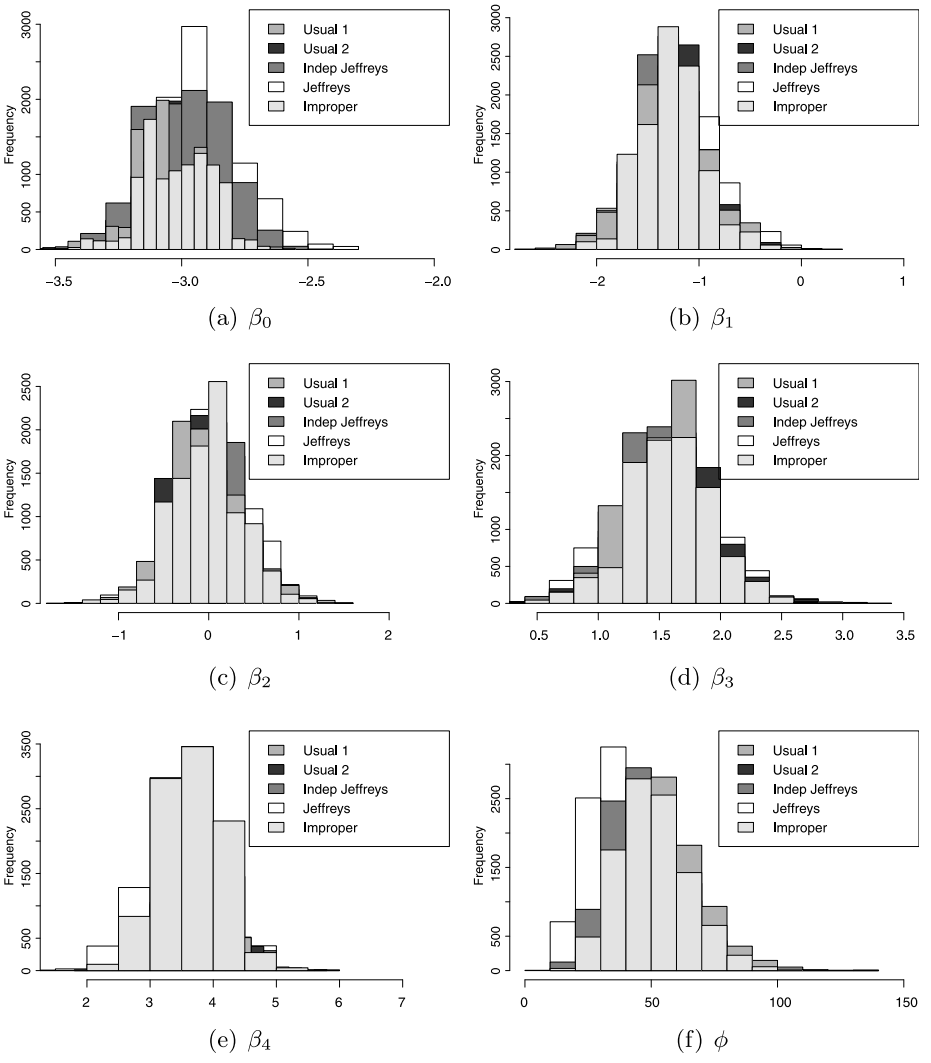


Figure 4 *Histogram of the posteriors distribution for* $\boldsymbol{\beta}$ (a)–(e) *and for* ϕ (f) *the ZOABR model to different priors considering* $n = 50$, $\phi = 50$, $p = 5$ *and* $\delta = 0.3$.

Table 5 *Absolute values of relative bias* (*AVRB*) *associated to* δ *for different values of* ϕ *and* δ *under* $p = 5$ *covariates for the ZOABR model*

		$\delta = 0.1$		$\delta = 0.3$		$\delta = 0.5$	
ϕ	n	Bayesian	ML	Bayesian	ML	Bayesian	ML
50	20	81.8%	50.0%	12.1%	6.7%	2.1%	2.3%
	50	25.6%	10.6%	1.0%	1.6%	1.5%	1.6%
	200	6.4%	2.5%	2.3%	1.7%	0.8%	0.8%
200	20	80.9%	49.0%	7.3%	1.3%	3.7%	4.1%
	50	23.7%	8.6%	2.9%	0.3%	0.9%	0.9%
	200	5.0%	1.0%	1.0%	0.3%	1.7%	1.7%

Table 6 *Absolute values of relative bias (AVRB) associated to γ for different values of ϕ and δ and with* $\mathrm{p} = 5$ *covariates for the ZOABR model*

		$\delta = 0.1$		$\delta = 0.3$		$\delta = 0.5$	
ϕ	n	Bayesian	ML	Bayesian	ML	Bayesian	ML
50	20	0.1%	0.3%	0.9%	1.0%	3.7%	4.5%
	50	7.2%	9.8%	0.1%	0.1%	0.7%	0.8%
	200	1.9%	2.1%	0.3%	0.3%	0.6%	0.6%
200	20	0.4%	0.7%	1.4%	1.3%	0.1%	0.3%
	50	0.3%	0.6%	3.4%	4.2%	1.0%	1.0%
	200	3.7%	4.1%	2.8%	2.9%	0.3%	0.4%

5 Application

In order to illustrate the Bayesian approach here developed, we conducted a novel approach of analysis of psychometric data. Specifically, the analyzed data was obtained from Carlstrom, Woodward and Palmer (2000) and it is available from http://www.stat.ucla.edu/projects/datasets/risk_perception.html. It corresponds to a psychometric study of risk perception. Specifically, we consider the so-called subjective part, where subjects were asked about the risk perceived by them, related to several financial and health activities. Each subject were asked to provide a number in the interval $[0, 100]$ such that the higher the value, the higher the risk perceived, being 0 non-risk at all and 100 maximum risk. In order to use the ZOABR model, the observations were transformed to the interval $[0, 1]$. Also, several covariates were measured and the goal was to analyze the impact of them on the risk perception. They are: age (measured in years), gender (male and female), world view (wvcat), classified as hierarchicalist, individualist, egalitarian or other (unclassifiable) and ethnicity (Caucasian, African-American, Mexican-American or Taiwanese-American).

The data set analyzed correspond to the perception of the subjects about the risk related to living close to a nuclear plant. We have a total of 592 observations, being 3 observations equal to zero and 181, to one. That is, approximately 30.6% of the participants are extremely afraid to live close to a nuclear plant.

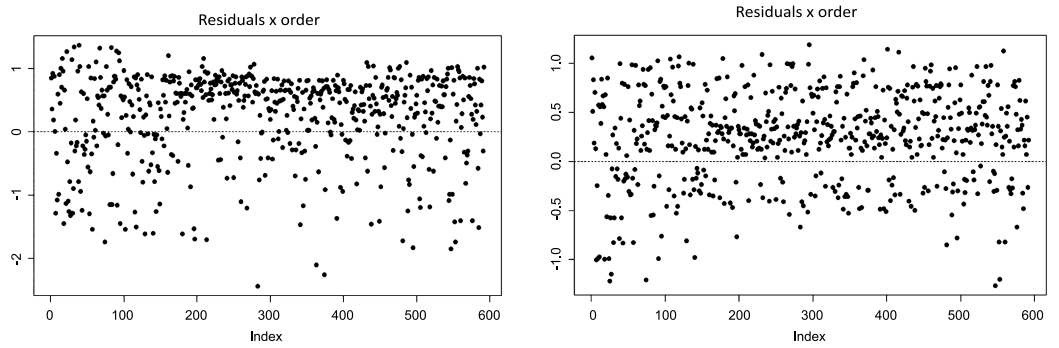
Following the results of simulation study, see Section 4, we choose the Jeffreys-rule prior for $\boldsymbol{\beta}$ and ϕ, and the beta$(1, 1)$ distribution for δ and γ. The other quantities, related to the MCMC algorithms, were exactly the same.

We started fitting a ZOABR model with all covariates (main effects) without interactions. Then, we excluded the non-significant covariates and introduced all first order interactions. We found that no interaction was significant and we present only the results related to the final model, that is:

$$\begin{aligned} Y_{tij} &\overset{\text{ind.}}{\sim} \text{ZOABD}(\delta, \gamma, \mu_{tij}, \phi), \\ \text{logit}(\mu_t) &= \mu + \nu x_t + \beta_2 z_{2t} + \theta_2 w_{2t} + \theta_3 w_{3t}, \end{aligned} \tag{5.1}$$

where $t = 1, \ldots, 512$; $z_{2t} = 1$ if male, 0 otherwise; $w_{2t} = 1$ if African-American, 0 otherwise; $w_{3t} = 1$ if Mexican-American, 0 otherwise, the parameter ν is related to age (x_t) and parameters β_2 and $\boldsymbol{\theta} = (\theta_2, \theta_3)^t$ are related to gender and ethnicity, respectively (notice that the covariate worldview was not significant and, then, it was not considered in the final model, as well as the groups Taiwanese-American and Caucasian were equivalent).

The residuals used here were the standard residuals (r_t), weighted standard residuals $(rp_t^{(01)})$ and the deviance residuals (rd_t) (see Ferrari and Cribari-Neto, 2004, Espinheira, Ferrari and Cribari-Neto, 2008 and Ospina and Ferrari, 2012). Different from these authors, we



(a) Weighted standard residuals (b) Deviance residuals

Figure 5 *Residual plots for the ZOABR model.*

Table 7 *Bayesian estimates for the final model*

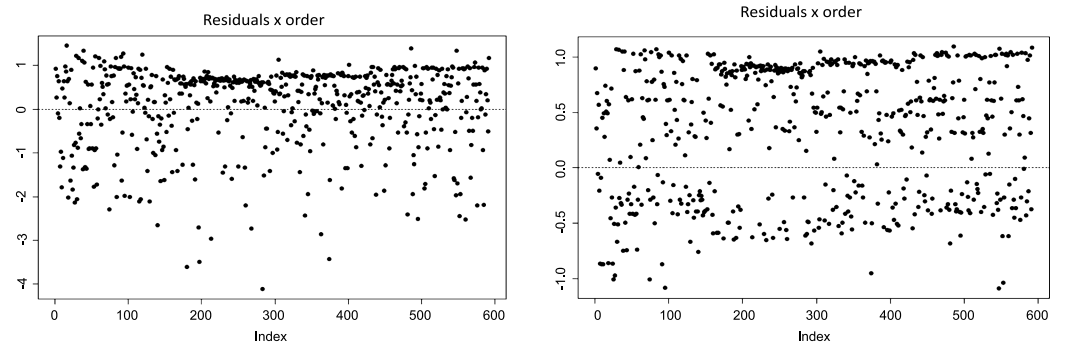
Parameter	EAP	PSD	CI (95%)
μ	1.10	0.12	[0.85; 1.32]
ν	−0.02	<0.01	[−0.02; −0.01]
β_2	−0.34	0.10	[−0.53; −0.14]
θ_2	0.57	0.15	[0.29; 0.86]
θ_3	0.48	0.12	[0.24; 0.69]
ϕ	2.96	0.20	[2.58; 3.37]
δ	0.31	0.02	[0.27; 0.35]
γ	0.98	0.01	[0.95; 0.99]

considered the predictive distribution of the response (Y_t) to obtain the residuals. That is, for each one of the 592 observations we generated, based on the model (5.1) and the MCMC sample of size 1,000 of each parameter, 1,000 values for each observation.

Figure 5 presents the weighted standard and deviance residuals. Some observations present large values of the residuals but, in a general way, they indicate that the model fitted to the data, satisfactorily.

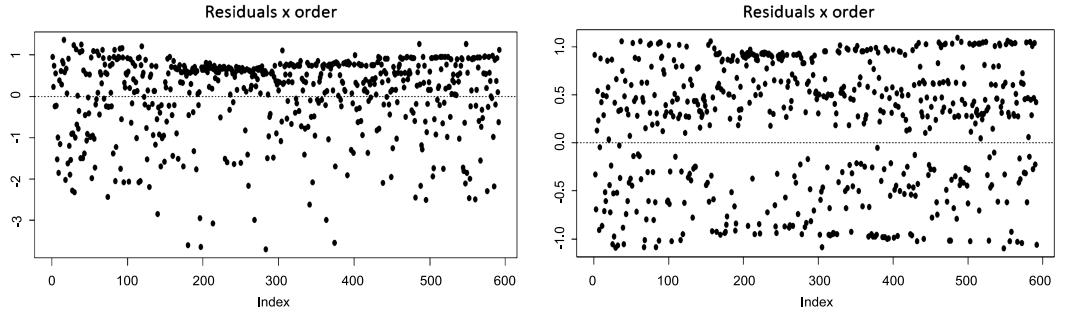
Table 7 presents the expectation a posteriori (EAP), the posterior standard deviation (PSD) and the respective 95% equi-tailed credibility intervals CI(95%) for all parameters. We can see that all regression parameters are significant. Also, the estimate of ϕ indicates that the data presents high variability. Notice that β_2 presents a negative sign (−0.34), indicating that for women, to live close to a nuclear plant is more dangerous than for men. Also, from the estimates of $(\theta_2, \theta_3)^t$ it is possible to conclude that the risk perceived is the same for the African-American and Mexican-American, being higher for this group compared with the Taiwanese-American/Caucasian group. In addition, the older the person is the less dangerous to him/her it is to live close to a nuclear plant, even though the impact, for a difference of one year, is small. The estimates of δ and γ indicate that the probability of a person providing an extreme value for the risk is around 30.4%, regardless their profile. Also, given that a person perceives an extreme risk, it is very likely that it is close to the maximum, due to magnitude of the estimate of γ.

Furthermore, we analyzed the impact of transforming the extreme risks (zero and one) on the estimates, that is, replacing these values by 0.001 and 0.999, respectively, and fitting the BR, OABR and ZABR models. Figures 5–8 present the residuals for the four models. We can notice that more extreme residuals are observed for the BR, OABR and ZABR models, compared with those for the ZOABR model, besides the former residuals present more skewed



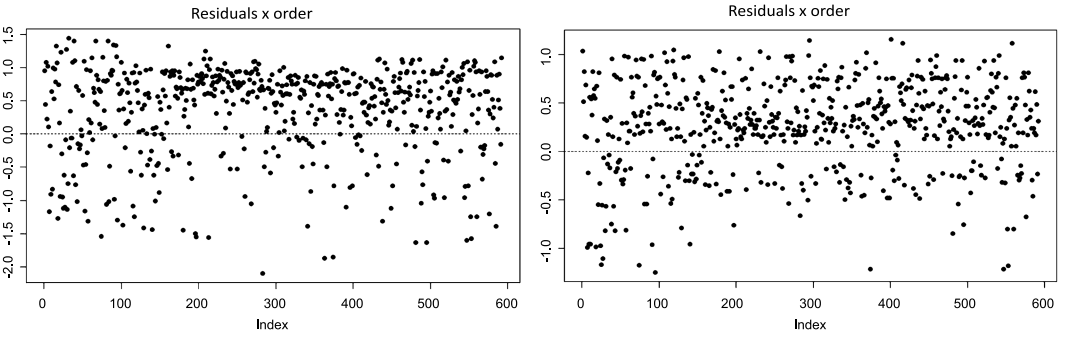
(a) Standard residuals (b) Deviance residuals

Figure 6 *Residual plots for the BR model.*



(a) Weighted standard residuals (b) Deviance residuals

Figure 7 *Residual plots for the ZABR model.*

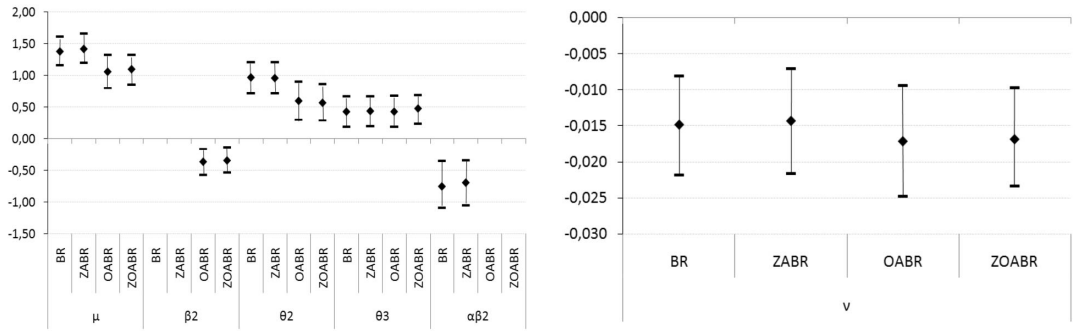


(a) Weighted standard residuals (b) Deviance residuals

Figure 8 *Residual plots for the OABR model.*

distributions than the latter. Therefore, we can conclude, based on the residual analysis, that the ZOABR model is more suitable to analyze the data and a poor fitting can be obtained when the data is transformed.

Another difference among the four models is related to the parameter estimates. In fact, depending on the model, some of the parameters are not significant. This is also an important aspect that illustrates how the use of non-augmented models to the transformed data can lead to mislead inference. Figure 9 presents the EAP and the respective 95% equi-tailed credibility intervals for the four models. For example, according to the OABR and ZOABR model, the parameter $(\alpha\beta)_{22}$, related to the interaction between gender and worldview, was



(a) Parameters associated to the factor covariabels (b) Age

Figure 9 *EAP and* 95% *equi-tailed credibility interval for the parameters.*

not significant, occurring the opposite for the BR and ZABR models. For the parameter β_2, related to the effect of gender, an opposite pattern is observed. Also, the similarity between the results obtained from the OABR and ZOABR model is due to the low presence of zeros and the high presence of ones in the sample.

6 Concluding remarks

In this work we compared Bayesian estimation of the ZOABR model (and the related particular cases), under different prior distributions, with the ML estimation. We found that the Bayesian estimates under the Jeffreys-rule prior as accurate as the others, including the ML ones, mainly for the precision parameter (ϕ). Also, all parameters are properly recovered. In addition, the higher the sample size and/or the smaller the variability is, the more accurate the estimates are, for all methods. For a given sample size and degree of variability (inverse of ϕ), the larger the number of covariates is, the less accurate the estimates are. For a given scenario, the parameters are more accurately estimated according to the complexity of the model. That is, estimates related to the BR model are more accurate than those of OABR/ZABR models which, in their turn, are more accurate than those of the ZOABR model.

We suggest the use of the Bayesian paradigm to fit the ZOABR model (and the respective particular cases) as an alternative to the ML approach, since the spent time is not prohibitive (for the cases that we explored), besides the aforementioned comments. Also, Bayesian influence diagnostics as well as mechanisms for posterior predictive assessment can be developed and implemented straightforwardly.

As future research, we suggest to extend the present simulation study to the ZOABR model where all parameters are modeled (using covariates) or mixed augmented limited regression models as that presented in Galvis, Bandyopadhyay and Lachos (2014) or to models that consider alternative distributions to the beta as in López (2013) and Lemonte and Bazán (2016). Also, the correspondent Jeffreys-rule prior and the independence Jeffreys prior can be explored for those models.

Other auxiliary algorithms as the Hamiltonian Monte Carlo (see Homan and Gelman, 2014), adaptive reject sampling and slice sampling (see Gamerman and Lopes, 2006) could be compared. Another aspect of interest is the use of other link functions for the response mean, as the probit, clolog, cauchit, skew probit, among others. To the best of our knowledge no works concerning this topic, for limited response regression models, are available in the literature. Specially for unbalanced data, similarly to the binary regression, asymmetric/heavy-tailed links could be preferable, see Bazan, Romeo and Rodrigues (2014) and Kim, Chen and

Dey (2008), for example. Finally, other numerical methods to obtain approximation for the marginal posterior distributions, as the INLA algorithm, can be useful, see Rue and Martino (2009).

Acknowledgments

The authors would like to thank CNPq ("Conselho Nacional de Desenvolvimento Científico e Tecnológico") for the finnancial support through a Master's scholarship granted to the first author under the guidance of the second. Also, the second author would like to thank to CNPq, Grant 308339/2015-0, related to a research scholarship. The third author was partially supported by the Brazilian agency FAPESP (Grant 2017/07773-6).

Supplementary Material

Supplement to "Bayesian modeling and prior sensitivity analysis for zero–one augmented beta regression models with an application to psychometric data" (DOI: 10.1214/18-BJPS423SUPP; .pdf). In the Supplementary Material we provide technical details for all MCMC algorithms and also additional results related to the Simulation Studies.

References

Bayes, C. L. and Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables. *Journal of Applied Statistics* **43**, 1814–1830. MR3492148 https://doi.org/10.1080/02664763.2015.1120711

Bazan, J. L., Romeo, J. R. and Rodrigues, J. (2014). Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics* **28**, 467–482. MR3263060 https://doi.org/10.1214/13-BJPS218

Branscum, A. J., Johnson, W. O. and Thurmond, M. C. (2007). Bayesian beta regression: Application to household data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics* **49**, 287–301. MR2405396 https://doi.org/10.1111/j.1467-842X.2007.00481.x

Buckley, J. (2003). Estimation of models with beta-distributed dependent variables: A replication and extension of Paolinos study. *Political Analysis* **11**, 204–205.

Carlstrom, L., Woodward, J. and Palmer, C. (2000). Evaluating the simplified conjoint expected risk model: Comparing the use of objective and subjective information. *Risk Analysis* **20**, 385–392.

Cepeda-Cuervo, E., Jaimes, D., Marín, M. and Rojas, J. (2016). Bayesian beta regression with Bayesian beta reg R-package. *Computational Statistics* **31**, 165–187. MR3481800 https://doi.org/10.1007/s00180-015-0591-9

Cribari-Neto, F. and Zeiles, A. (2010). Beta regression in R. *Journal of Statistical Software* **34**, 1–24.

Dey, D. K., Ghosh, S. K. and Mallick, B. N. (2000). *Generalized Linear Models: A Bayesian Perspective*. MR1893779

Espinheira, P., Ferrari, S. L. P. and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics* **35**, 407–419. MR2420486 https://doi.org/10.1080/02664760701834931

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815. MR2095753 https://doi.org/10.1080/0266476042000214501

Figueroa-Zuñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. P. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis* **61**, 137–147. MR3063006 https://doi.org/10.1016/j.csda.2012.12.002

Galvis, D. M., Bandyopadhyay, D. and Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in Medicine* **33**, 3759–3771. MR3260658 https://doi.org/10.1002/sim.6179

Gamerman, D. and Lopes, H. (2006). *Stochastic Simulation for Bayesian Inference*, 2nd ed. New York: Chapman & Hall/CRC.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94**, 247–253. MR1689229 https://doi.org/10.2307/2669699

Homan, M. D. and Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623. MR3214779

Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions, and fractions. *Statistical Modelling* **3**, 193–213. MR2005473 https://doi.org/10.1191/1471082X03st053oa

Kim, S., Chen, M.-H. and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* **95**, 93–106. MR2409717 https://doi.org/10.1093/biomet/asm079

Lemonte, A. J. and Bazán, J. L. (2016). New class of Johnson SB distributions and its associated regression model for rates and proportions. *Biometrical Journal* **58**, 727–746. MR3527412 https://doi.org/10.1002/bimj.201500030

Liu, F. and Kong, Y. (2015). zoib: An R package for Bayesian inference for beta regression and zero/one inflated beta regression. *The R Journal* **7**, 34–51.

López, F. O. (2013). A Bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression. *Revista Colombiana de Estadística* **36**, 1–21. MR3075189

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Economics*. New York: Cambridge University Press. MR0799154 https://doi.org/10.1017/CBO9780511810176

Nogarotto, D. C. (2013). Bayesian inference EM beta and inflated beta regression model. Master's dissertation (in Portuguese), IMECC-Unicamp.

Nogarotto, D. C., Azevedo, C. L. N. and Bazán, J. L. (2020). Supplement to "Bayesian modeling and prior sensitivity analysis for zero–one augmented beta regression models with an application to psychometric data." https://doi.org/10.1214/18-BJPS423SUPP

Oliveira, M. S. (2004). A beta regression model: Theory and application. Master's dissertation (in Portuguese), IME-USP.

Ospina, R. (2008). Inflated beta regression modelo. Doctoral's thesis (in Portuguese), IME-USP.

Ospina, R. and Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers* **51**, 111–126. MR2556590 https://doi.org/10.1007/s00362-008-0125-4

Ospina, R. and Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* **56**, 1609–1623. MR2892364 https://doi.org/10.1016/j.csda.2011.10.005

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* **9**, 325–346.

Papke, L. and Wooldridge, J. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* **11**, 619–632.

Parker, A. J., Bandyopadhyay, D. and Slate, E. H. (2014). A spatial augmented beta regression model for periodontal proportion data. *Statistical Modelling* **14**, 503–521. MR3284553 https://doi.org/10.1177/1471082X14535515

Paulino, C. D., Turkman, M. A. A. and Murteira, B. (2003). *Bayesian Statistic* (in Portuguese). Lisboa: Fundação Calouste Gulbenkian.

Pereira, T. L. and Cribari-Neto, F. (2014). Detecting model misspecification in inflated beta regressions. *Communications in Statistics Simulation and Computation* **43**, 631–656. MR3200996 https://doi.org/10.1080/03610918.2012.712183

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org.

Ramalho, J. J. S. and 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**, 621–636. MR2548100 https://doi.org/10.1080/14697680802448777

Rue, H. and Martino, S. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71**, 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54–71.

Souza, T. C., Pereira, T. L., Cribari-Neto, F. and Lima, M. C. V. (2016). Testing inference in inflated beta regressions under model misspecification. *Communications in Statistics Simulation and Computation* **45**, 625–642. MR3457110 https://doi.org/10.1080/03610918.2013.867995

Stavrunova, O. and Yerokhin, O. (2012). Two-part fractional regression model for the demand for risky assets. *Applied Economics* **44**, 21–26.

Swearingen, C. J., Castro, M. S. M. and Bursac, Z. (2012). Inflated beta regression: Zero, one, and everything in between. In *SAS Global Forum 2012*. Available at http://support.sas.com/resources/papers/proceedings12/325-2012.pdf.

Wieczorek, J. and Hawala, S. (2011). A Bayesian zero–one inflated beta model for estimating poverty in U.S. counties. In *Proceedings of the American Statistical Association. Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

D. Covaes Nogarotto
School of Technology
Campinas State University
Limeira
Brazil
E-mail: nogarotto.danilo@gmail.com

C. Lucidius Naberezny Azevedo
Department of Statistics
Campinas State University
Campinas
Brazil
E-mail: cnaber@ime.unicamp.br

J. Luis Bazán
Department of Applied Mathematics and Statistics
University of São Paulo
São Carlos
Brazil
E-mail: jlbazan@icmc.usp.br