Likelihood-based inference for zero-or-one augmented rectangular beta regression models

Ana R. S. Santos, [1], Caio L N Azevedo[1]*, Jorge L. Bazan[2],
Juvêncio S. Nobre[3]

[1] Department of Statistics, State University of Campinas, Brazil
[2] Department of Applied Math and Statistics, University of São Paulo, Brazil
[2] Department of Statistics and Applied Math, Federal University of Ceará, Brazil

Abstract

A new zero-and/or-one augmented beta rectangular regression model is introduced in this work, which is based on a new parameterization of the rectangular beta distribution. Maximum likelihood estimation is performed by using a combination of the EM algorithm (for the continuous part) and Fisher scoring algorithm (for discrete part). Also, we develop techniques of model fit assessment, by using the randomized quantile residuals and model selection, considering criteria, such as AIC and BIC. We conducted several simulation studies, considering some situations of practical interest, in order to evaluate the parameter recovery of the proposed model and estimation method, the impact of transforming the observed zeros and ones with the use of non-augmented models and the behavior of the model selection criteria. A psychometric real data set was analyzed to illustrate the performance of the new approach considering the model studied.

keywords: Augmented rectangular beta distribution; diagnostic analysis; frequentist inference; generalized linear models; proportional data

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. Some examples include the fraction of income contributed to a retirement fund, the proportion of weekly hours spent on work-related activities, the fraction of household income spent on food, etc. To analyze bounded response variables, the main developed model was the beta regression model; see Ferrari and Cribari-Neto (2004). It is currently a fairly consolidated model including some extensions as the mixed beta model (see Galvis et al. (2014)) and beta-mixture model (see Ma and Leijon (2011)). Also, residual analysis and model comparison are well developed for them. 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) and Cribari-Neto and Zeileis (2010).

When it is possible to observe zero and/or one values 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 et al. (2014). In this work, we prefer use the term "augmented", as in Galvis et al. (2014), instead of "inflated", as in Ospina and Ferrari

[*] Corresponding author: Caio L N Azevedo, Department of Statistics, State University of Campinas, Mailbox 6065, SP, Brazil. Email: cnaber@ime.unicamp.br

(2012), 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), also named discrete values. Correspondent zero-or-one (or zero and one) augmented beta (ZOAB) regression models have been proposed in the literature as in Ospina and Ferrari (2012) and Bayes and Valdivieso (2016). Pereira (2010) presents several applications of the augmented beta regression models. Furthermore, Pereira (2012) introduces the truncated inflated beta regression model that is considered in situations where the data is coming from a distribution that is a mixture of a beta distribution in the range $(c, 1)$ and a trinomial distribution that takes values zero, one and c. Also, several statistical testing and model misspecification detection techniques to the ZOAB regression model were studied; see for example Cribari-Neto and Pereira (2014) and Souza et al. (2016). Additionally, extensions to spatial data were proposed by Parker et al. (2014).

Bayes et al. (2012) proposed a rectangular beta regression model as a robust alternative to model limited data. The parameters that they use allows to model directly the response mean using a linear predictor and a general link function, such that the specification is similar to the generalized linear models. Wang and Luo (2015) proposed a framework that consists on a multivariate Bayesian augmented rectangular beta regression model for longitudinal outcomes belonging to the closed unit interval $[0, 1]$ and a Cox proportional hazard model for the dependent censoring event. They consider a zero-and/or-one augmented rectangular beta regression model with random effects, including covariables of interest for modeling the mean and the probabilities of occurrence of zero and one. In addition, to detect the presence of outliers and extreme observations, they considered the Kullback Leibler (K-L) divergence and, for model comparison, they considered the usual statistics in the Bayesian context. They also conducted a simulation study to compare the performance of the proposed model.

Wang and Luo (2016) generalize the model proposed by Bayes et al. (2012) and developed, under the Bayesian perspective, an augmented rectangular beta regression model to account for the occurrence of boundary values 0 and 1 for (0,1) data. Moreover, they account for the within-subject correlation in a longitudinal setup by introducing random effects under the generalized linear mixed models framework. However, they only developed the one-augmented rectangular beta random effects model, modeling the mean, the precision parameter and the probabilities of occurrence of one. As in Wang and Luo (2015), they used the Kullback Leibler (K-L) divergence to detect the occurrence of outliers and extreme observations and, for model comparison. They used some commonly employed tools in the Bayesian context. In addition, they conducted a simulation study to compare the performance of the usual one-augmented beta regression model (see Ospina and Ferrari (2012)) and the proposed regression model.

In this paper, we developed, under the frequentist perspective, statistical modeling of data distributed in the closed unit interval [0,1]. The idea is to assume, for the response variable, a mixture structure between the rectangular beta distribution and the Bernoulli distribution, which assigns probabilities to the integers 0 and 1. We developed a zero-and/or-one augmented rectangular beta distribution, using a new parametrization, as well as a correspondent zero-and/or-one augmented rectangular beta regression model. We considered appropriate regression structures for the mean, the dispersion parameter and the probabilities of occurrence of ones and zeros. The maximum likelihood estimates are obtained through a combination of the EM algorithm and Fisher scoring algorithm. The respective standard errors are obtained through the Fisher Information and the Score function. Furthermore, we developed techniques of residual analysis, by using the randomized quantile residuals. Statistics for model comparison were explored. We conducted simulation studies in order to measure: the parameter recovery of the developed model and the estimation method, the impact of transforming the discrete values, in order to use non-augmented regression models, on the parameter estimation, the behavior of the residuals and the performance of the statistics of model comparison.

It is noteworthy that, unlike the works of Wang and Luo (2015) and Wang and Luo (2016), in this paper we present a new parametrization of the rectangular beta distribution, we developed the estimation of the parameters under the frequentist perspective, we include regression structures for more parameters, we develop techniques of residual analysis, by using the randomized quantile residuals and we conduct simulation studies in order to: measure the parameter recovery, to study the behavior of the residuals and the performance of the statistics of model comparison, considering different scenarios, defined by crossing the levels of some factors of interest.

The remainder of the paper is organized as follows. In Section 2 we present the zero-and/or-one augmented rectangular beta distribution. In Section 3, we present the zero-and/or-one augmented rectangular beta (ZOABR) regression model. In Section 4, we discuss maximum likelihood estimation, that we use the EM algorithm to the parameters of the continuous part and, for the discrete part, we use the Fisher scoring algorithm. In Section 5, we develop techniques of residual analysis, by using the randomized quantile residuals and we present some statistical tools for model selection. In Section 6, we present some simulation studies. In Section 7, we present the analysis of a psychometric data set using the developed methodology, including a residual analysis to evaluate the goodness of fit of the model as well as a comparison with other competing model. Finally, in Section 8, we present some discussion and suggestions for future research.

2 Zero-and/or-one augmented rectangular beta distribution

The beta distribution with parameters μ and ϕ, proposed by Ferrari and Cribari-Neto (2004) and denoted by beta(μ, ϕ), has the following density

$$b(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\mathbb{1}_{(0,1)}(y),$$

where $0 < \mu < 1$, $\phi > 0$ and $\Gamma(\cdot)$ is the gamma function. If $Y \sim$ beta(μ, ϕ), then $\mathbb{E}(Y) = \mu$ and $\mathbb{V}\text{ar}(Y) = \dfrac{V(\mu)}{1+\phi}$.

The beta distribution presents a reasonably flexibility, since its density can have different shapes, implied by changes in the values of μ and ϕ. However, as was noted by Hahn (2008) and García et al. (2011), the beta distribution neither considers tail-area events, neither allows more flexibility in the variance specification. This fact can limit its application for modeling proportions. In order to get some additional flexibility, Bayes et al. (2012) provide a regression model which permits varying amounts of dispersion and greater likelihood of more extreme tail-area events; by considering the beta rectangular distribution with parameters μ, ϕ and θ, which was proposed by Hahn (2008), denoted by BR(μ, ϕ, θ), whose density is given by

$$g(y; \mu, \phi, \theta) = \theta\mathbb{1}_{(0,1)}(y) + (1-\theta)b(y|\mu, \phi)\mathbb{1}_{(0,1)}(y),$$

where $0 \leq \theta \leq 1$ is a mixture parameter. If $Y \sim$ BR(μ, ϕ, θ), then $\mathbb{E}(Y) = \dfrac{\theta}{2} + (1-\theta)\mu$ and $\mathbb{V}$ ar$(Y) = \dfrac{V(\mu)}{1+\phi}(1-\theta)[1-\theta(1+\phi)] + \dfrac{\theta}{12}(4-3\theta)$.

For a regression analysis, the mean of the response is typically modeled. However, the mean of the beta rectangular distribution is a function of the parameters μ and θ. According to Bayes et al. (2012) if we consider $\mathbb{E}(Y) = \dfrac{\theta}{2} + (1-\theta)\mu = \gamma$, the parameter space of μ is restricted to $0 < \theta < 1 - |2\gamma - 1| < 1$. In order to obtain a more appropriate regression structure for the mean, Bayes and Bazán (2014) defined

$$\gamma = \frac{\theta}{2} + (1-\theta)\mu \quad \text{and} \quad \alpha = \frac{\frac{\theta}{2}\left(1-\frac{\theta}{2}\right)}{\frac{\theta}{2}\left(1-\frac{\theta}{2}\right) + (1-\theta)^2\mu(1-\mu)},$$

as a new parametrization. In this case, the parameter space of γ and α is the rectangle given by $\{0 \le \gamma \le 1, 0 \le \alpha \le 1\}$.

Under this parametrization, we have

$$\theta = 1 - \sqrt{1-4\alpha\gamma(1-\gamma)} \quad \text{and} \quad \mu = \frac{\gamma - \frac{1}{2} + \frac{1}{2}\sqrt{1-4\alpha\gamma(1-\gamma)}}{\sqrt{1-4\alpha\gamma(1-\gamma)}}$$

and, consequently, the reparameterized density of the rectangular beta distribution, denoted by $BRr(\gamma,\phi,\alpha)$, may be expressed as

$$\begin{aligned} h(y;\gamma,\phi,\alpha) &= \left(1-\sqrt{1-4\alpha\gamma(1-\gamma)}\right)\mathbb{1}_{(0,1)}(y) + \sqrt{1-4\alpha\gamma(1-\gamma)} \times \\ &\times \; b\left(\frac{\gamma - \frac{1}{2} + \frac{1}{2}\sqrt{1-4\alpha\gamma(1-\gamma)}}{\sqrt{1-4\alpha\gamma(1-\gamma)}}, \phi\right)\mathbb{1}_{(0,1)}(y). \end{aligned} \tag{2.1}$$

For proportions observed in the interval [0,1], we assume that the probability of observing the values zero and one are positive. For this type of data, we use a distribution obtained from the mixing between a rectangular beta distribution and a Bernoulli distribution, which assigns positive probabilities to the integers 0 and 1. In this case, the rectangular beta distribution is used to model the continuous component, while the Bernoulli distribution accounts for the discrete component, that is, the probabilities related to zero and one. We also consider the particular cases when only either zero or one can be observed. Then, we say that Y has a zero-and/or-one augmented rectangular beta distribution (ZOABR), with parameters $(\tau,\eta,\gamma,\phi,\alpha)^\top$, if its density is given by

$$f(y;\tau,\eta,\gamma,\phi,\alpha) = \left[\tau(1-\eta)^{1-y}(\eta)^y\right]\mathbb{1}_{\{0,1\}}(y) + (1-\tau)h(y;\gamma,\phi,\alpha)\mathbb{1}_{(0,1)}(y),$$

where $h(y;\gamma,\phi,\alpha)$ is as in (2.1) and $(\tau,\eta) \in (0,1)^2$. In this case we use the notation $Y \sim$ ZOABR $(\tau,\eta,\gamma,\phi,\alpha)$, where τ is the probability of observing 0 or 1. On the other hand, η is the probability of the observation be equal to one given it is augmented (i.e., given that it belongs to the discrete part).

The mean and variance of the ZOABR distribution are given by $\mathbb{E}(Y) = \tau\eta + (1-\tau)\gamma$ and $\mathbb{V}\mathrm{ar}(Y) = \tau V_1 + (1-\tau)V_2 + \tau(1-\tau)(\eta-\gamma)^2$, respectively, where $V_1 = \eta(1-\eta)$, $V_2 = \dfrac{\theta}{3} + (1-\theta)\left[\dfrac{\mu(1-\mu)}{1+\phi} + \mu^2\right] - \gamma^2$. Note that $\mathbb{E}(Y)$ is a weighted average between the first moment of the Bernoulli distribution and the corresponding moment of the reparameterized rectangular beta distribution with weights τ and $1-\tau$, respectively.

Additionally, other (re)parameterizations of the ZOABR distribution can be introduced. For example, following the proposal of Ospina (2008), we can define another version in which the Bernoulli distribution parameter satisfies the relation $\eta = p_1/\tau$ and the mixture parameter is such that $\tau = p_0 + p_1$. This parameterization is useful to define regression models, since the probabilities of occurrence of zeros and ones are directly specified. Thus, the ZOABR density can be written as

$$f(y; p_0, p_1, \gamma, \phi, \alpha) = p_0^{1-y} p_1^y \mathbb{1}_{\{0,1\}}(y) + (1 - p_0 - p_1) h(y|\gamma, \phi, \alpha) \mathbb{1}_{(0,1)}(y) \tag{2.2}$$

where $h(y; \gamma, \phi, \alpha)$ is as (2.1). Note that this parameterization induces a restriction in the parameter space given by $0 < p_0 + p_1 < 1$. Also, when $\alpha = 0$ we obtain the augmented beta distribution proposed by Ospina (2008). In this case, we use the notation $Y \sim \text{ZOABR}(p_0, p_1, \gamma, \phi, \alpha)$. On the other hand, the cumulative distribution function (cdf) of ZOABR, which is useful to define the so-called quantile residuals, can be defined as follows

$$\mathrm{F}(y; \tau, \eta, \gamma, \phi, \alpha) = \tau \mathrm{Ber}(y; \eta) + (1 - \tau) \mathrm{BR}(y; \gamma, \phi, \alpha),$$

where $\mathrm{Ber}(y; \eta)$ is the cdf of a Bernoulli with parameter η and $\mathrm{BR}(y; \gamma, \phi, \alpha)$ is the cdf of the rectangular beta given in (2.1).

3 Zero-and/or-one augmented rectangular beta regression model

Let $Y_t \overset{ind.}{\sim} \text{ZOABR}(p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha)$, $t = 1, \ldots, n$. The ZOABR regression model is defined by (2.2) and the systematic components

$$\begin{aligned} g_1(\gamma_t) &= \sum_{i=1}^{p} x_{ti} \beta_i = \mathbf{x}_t^\top \boldsymbol{\beta}, \quad g_2(\phi_t) = \sum_{j=1}^{k} -w_{tj} \delta_j = -\mathbf{w}_t^\top \boldsymbol{\delta} \\ H(p_{0t}, p_{1t}) &= (h_0(p_{0t}, p_{1t}), h_1(p_{0t}, p_{1t})) = (\zeta_{0t}, \zeta_{1t}) = (\mathbf{v}_t^\top \boldsymbol{\rho}, \mathbf{z}_t^\top \boldsymbol{\psi}), \end{aligned} \tag{3.1}$$

where $\gamma_t = \mathbb{E}(Y_t | Y_t \in (0,1)), p_{0t} = \mathbb{P}(Y_t = 0), p_{1t} = \mathbb{P}(Y_t = 1)$ and $1 - p_{0t} - p_{1t} = \mathbb{P}(Y_t \in (0,1))$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top, \boldsymbol{\delta} = (p_1, \ldots, \delta_k)^\top, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_{k_0})^\top, \boldsymbol{\psi} = (\psi_1, \ldots, \psi_{k_1})^\top$ are vectors of unknown regression parameters such that $\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\delta} \in \mathbb{R}^k, \boldsymbol{\rho} \in \mathbb{R}^{k_0}$ and $\boldsymbol{\psi} \in \mathbb{R}^{k_1}$. Here, $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^\top, \mathbf{w}_t = (w_{t1}, \ldots, w_{tk})^\top, \mathbf{v}_t = (v_{t1}, \ldots, v_{k_0})^\top$ and $\mathbf{z}_t = (z_{t1}, \ldots, z_{k_1})^\top$ are vectors with p, k, k_0 and k_1 covariates, respectively.

Following to Bayes et al. (2012) we use the negative sign in $g_2(\phi_t)$, as indicated by Smithson and Verkuilen (2006), to make the interpretation of the coefficients $\boldsymbol{\delta}$ easier. Since ϕ_t is a precision parameter, a positive-signed δ_j indicates smaller variance, which is potentially confusing. It seems more natural to model the dispersion rather than the precision parameter, and the negative sign enables us to do so. Also the estimates of the dispersion parameter, or the related regression parameters, are more accurate than those associated with the precision parameter, see Cribari-Neto and Souza (2012).

Assume that the link functions $g_1 : (0,1) \to \mathbb{R}$ and $g_2 : \mathbb{R}^+ \to \mathbb{R}$ are strictly monotonic and twice differentiable. Also H is a bijective transformation of the set $\mathbb{C} = \{(p_{0t}, p_{1t}) : 0 < p_{0t} < 1, 0 < p_{1t} < 1 - p_{0t}\}$ to $\mathbb{R}^2$, doubly differentiable. The conditions imposed on H ensures that the partial derivatives of $p_{0t} = h_0^*(\zeta_{0t}, \zeta_{1t})$ and $p_{1t} = h_1^*(\zeta_{0t}, \zeta_{1t})$ are continuous in $\mathbb{R}^2$ and p_{0t}, p_{1t} can be written in terms of ζ_{0t} and ζ_{1t}, uniquely. According to Ospina (2008), we can consider H such that

$$\begin{aligned} H(p_{0t}, p_{1t}) &= (h_0(p_{0t}, p_{1t}), h_1(p_{0t}, p_{1t})) \\ &= \left(h\left(\frac{p_{0t}}{1 - p_{0t} - p_{1t}} \right), h\left(\frac{p_{1t}}{1 - p_{0t} - p_{1t}} \right) \right), \end{aligned}$$

where $h : \mathbb{R}^+ \to \mathbb{R}$ is strictly monotonic and twice differentiable. Notice that h_0 and h_1 are functions of $\mathbb{R}^2$ in $\mathbb{R}$. We consider the logit link for g_1, that is, $g_1(\gamma_t) = \log(\gamma_t/1-\gamma_t)$ and the log link for g_2, this is, $g_2(\phi_t) = \log(\phi_t)$. Following Ospina (2008), we chose h as the log link, that is, $h_0(p_{0t}, p_{1t}) = \log(p_{0t}/(1-p_{0t}-p_{1t})) = \zeta_{0t}$ and $h_1(p_{0t}, p_{1t}) = \log(p_{1t}/(1-p_{0t}-p_{1t})) = \zeta_{1t}$.

In order to structure in a clearer way the mixture between the rectangular reparameterized beta distribution and Bernoulli, let us define the following variable

$$z_t^* = \begin{cases} 0 \text{ if } y_t \in (0,1), \\ \\ 1 \text{ if } y_t \in \{0,1\}. \end{cases}$$

When $Z_t^* = 0$, we have that $Y_t \sim \mathrm{BRr}(\gamma_t, \phi_t, \alpha)$ and when $Z_t^* = 1$, we have that $Y_t \sim$ Bernoulli$(p_{1t}/(p_{0t}+p_{1t}))$. The joint distribution of $(Y_t, Z_t^*)^\top$ is given by:

$$\begin{aligned} f(y_t, z_t^*; p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha) &= f_1(y_t; \gamma_t, \phi_t, \alpha)^{1-z_t^*} f_2(y_t; \eta_t)^{z_t^*} \times \\ &\times (p_{0t}+p_{1t})^{z_t^*}(1-p_{0t}-p_{1t})^{1-z_t^*} \mathbb{1}_{\{y_t, z_t^*\}} \end{aligned} \quad (3.2)$$

where $\mathbb{1}_{\{y_t,z_t^*\}} = \mathbb{1}_{(0,1)}(y_t)\mathbb{1}_{\{0\}}(z_t^*) + \mathbb{1}_{\{0,1\}}(y_t)\mathbb{1}_{\{1\}}(z_t^*)$, where $f_1(y_t; \gamma_t, \phi_t, \alpha)$ is the density of the rectangular beta distribution as defined in (2.1) and $f_2(y_t; \eta_t)$ is the probability function of a Bernoulli with parameter $\eta_t = p_{1t}/(p_{0t}+p_{1t})$. Let us denote $f(y_t, z_t^*; p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha) \equiv f(y_t, z_t^*)$, for short.

4 Maximum likelihood estimation

The likelihood for the ZOABR regression model is given by

$$L(\boldsymbol{\Upsilon}) = \prod_{t=1}^{n} f(y_t, z_t^*; p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha) = L_1(\boldsymbol{\rho}, \boldsymbol{\psi}) L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha),$$

where $f(y_t, z_t^*; p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha)$ is the joint distribution of $(y_t, z_t^*)^\top$ defined in (3.2) and

$$\begin{aligned} L_1(\boldsymbol{\rho}, \boldsymbol{\psi}) &= \prod_{t=1}^{n} (p_{0t}^{1-y_t} p_{1t}^{y_t})^{z_t^*} (1-p_{0t}-p_{1t})^{1-z_t^*} \\ L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha) &= \prod_{t=1}^{n} h(y_t; \gamma_t, \phi_t, \alpha)^{1-z_t^*}, \end{aligned}$$

where p_{0t}, p_{1t}, γ_t and ϕ_t are defined by (3.1) as functions of $\boldsymbol{\rho}, \boldsymbol{\psi}, \boldsymbol{\beta}$ and $\boldsymbol{\delta}$, respectively.

In order to obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\Upsilon}$, let us define an unobserved variable u_t, so that

$$U_t = \begin{cases} 0 \text{ if } Y_t \sim \text{beta}(\mu_t, \phi_t), \text{with probability } 1-\theta_t, \\ \\ 1 \text{ if } Y_t \sim \text{Unif}(0,1), \text{with probability } \theta_t. \end{cases} .$$

Therefore $U_t \overset{ind.}{\sim}$ Bernoulli(θ_t). The joint distribution of $(y_t, u_t, z_t^*)^\top$ can be written as

$$\begin{aligned} f(y_t, u_t, z_t^* | \boldsymbol{\Upsilon}) &= (p_{0t}^{1-y_t} p_{1t}^{y_t})^{z_t^*} (1-p_{0t}-p_{1t})^{1-z_t^*} \times \\ &\times [\theta_t^{u_t}(1-\theta_t)^{1-u_t} b(y_t; \mu_t, \phi_t)^{1-u_t}]^{1-z_t^*} \mathbb{1}_{\{y_t, u_t, z_t^*\}}, \end{aligned} \quad (4.1)$$

where $\mathbb{1}_{\{y_t,u_t,z_t^*\}} = \mathbb{1}_{(0,1)}(y_t)\mathbb{1}_{\{0,1\}}(u_t)\mathbb{1}_{\{0\}}(z_t^*)+\mathbb{1}_{\{0,1\}}(y_t)\mathbb{1}_{\{1\}}(z_t^*)$, $\mu_t = \dfrac{\gamma_t - \frac{1}{2} + \frac{1}{2}\sqrt{1-4\alpha\gamma_t(1-\gamma_t)}}{\sqrt{1-4\alpha\gamma_t(1-\gamma_t)}}$.

The complete likelihood associated to $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{z}^{*\top})^\top$ is given by

$$L_c(\boldsymbol{\Upsilon};\mathbf{y}_c) = \prod_{t=1}^{n} f(y_t, u_t, z_t^*) = L_1(\boldsymbol{\rho},\boldsymbol{\psi})L_{2c}(\boldsymbol{\beta},\boldsymbol{\delta},\alpha),$$

where $f(y_t, u_t, z_t^*)$ is defined in (4.1) where

$$L_{2c}(\boldsymbol{\beta},\boldsymbol{\delta},\alpha,u_t) \;\;=\;\; \prod_{t=1}^{n}[\theta_t^{u_t}(1-\theta_t)^{1-u_t}b(y_t;\mu_t,\phi_t)^{1-u_t}]^{1-z_t^*},$$

and the respective log-likelihood given by

$$\ell_c(\boldsymbol{\Upsilon};\mathbf{y}_c) = \sum_{t=1}^{n} f(y_t, u_t, z_t^*; p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha) = \ell_1(\boldsymbol{\rho},\boldsymbol{\psi}) + \ell_2(\boldsymbol{\beta},\boldsymbol{\delta},\alpha),$$

where

$$\begin{aligned}
\ell_1(\boldsymbol{\rho},\boldsymbol{\psi}) &= \sum_{t=1}^{n}\{z_t^*[(1-y_t)\log(p_{0t}) + y_t\log(p_{1t})] \\
&+ (1-z_t^*)\log(1-p_{0t}-p_{1t})\}, \\
\ell_{2c}(\boldsymbol{\beta},\boldsymbol{\delta},\alpha,u_t) &= \sum_{t=1}^{n}\{(1-z_t^*)\{u_t\log\theta_t + (1-u_t)\log(1-\theta_t) + (1-u_t)\times \\
&\times [\log(\Gamma(\phi_t)) - \log(\Gamma(\mu_t\phi_t)) - \log(\Gamma((1-\mu_t)\phi_t)) \\
&+ (\mu_t\phi_t - 1)\log y_t + ((1-\mu_t)\phi_t - 1)\log(1-y_t)]\}\}.
\end{aligned}$$

Note that the likelihood $L_c(\boldsymbol{\Upsilon};\mathbf{y}_c)$ can be factored into two terms, one depending only on the parameters $(\boldsymbol{\rho}^\top,\boldsymbol{\psi}^\top)^\top$ and the other depending only on the parameters $(\boldsymbol{\beta}^\top,\boldsymbol{\delta}^\top,\alpha)^\top$. That is, the likelihood and, consequently, the log-likelihood, are separable. Therefore, it is possible to estimate the parameters $(\boldsymbol{\rho}^\top,\boldsymbol{\psi}^\top)^\top$ (related to the discrete part) separated from the parameters $(\boldsymbol{\beta}^\top,\boldsymbol{\delta}^\top,\alpha)^\top$ (related to the continuous part). While for the former it is easy to maximize the likelihood directly through the Fisher Scoring algorithm, for example, for the latter we use the EM-algorithm, considering the augmented variables $u_t, t = 1,\ldots,n$ as non-observable. In the following two subsections we present the estimation for each set of parameters.

It is noteworthy that the term "augmented", concerning the variable u_t, refers to the usual definition of the augmented variables, which is different from the meaning related to the model definition (see Section 1).

4.1 Estimation of the parameters related to the discrete part - Fisher Scoring algorithm

The score function (see Appendix A), is given by

$$\mathbf{U}(\boldsymbol{\varphi}) = (\mathbf{U}_{\boldsymbol{\rho}}(\boldsymbol{\rho},\boldsymbol{\psi})^\top, \mathbf{U}_{\boldsymbol{\psi}}(\boldsymbol{\rho},\boldsymbol{\psi})^\top)^\top, \tag{4.2}$$

where

$$\begin{aligned}
\mathbf{U}_{\boldsymbol{\rho}}(\boldsymbol{\rho},\boldsymbol{\psi}) &= \mathbf{V}^\top\mathbf{T}_0[\boldsymbol{\Delta}_0\mathbf{Z}^*(\mathbf{1}-\mathbf{y})-\boldsymbol{\Delta}_{(\mathbf{0},\mathbf{1})}(\mathbf{1}-\mathbf{z}^*)]+\mathbf{V}^\top\mathbf{T}_{\mathbf{10}}[\boldsymbol{\Delta}_{\mathbf{1}}\mathbf{Z}^*\mathbf{y}-\boldsymbol{\Delta}_{(\mathbf{0},\mathbf{1})}(\mathbf{1}-\mathbf{z}^*)],\\
\mathbf{U}_{\boldsymbol{\psi}}(\boldsymbol{\rho},\boldsymbol{\psi}) &= \mathbf{Z}^\top\mathbf{T}_{01}[\boldsymbol{\Delta}_0\mathbf{Z}^*(\mathbf{1}-\mathbf{y})-\boldsymbol{\Delta}_{(\mathbf{0},\mathbf{1})}(\mathbf{1}-\mathbf{z}^*)]\\
&+ \mathbf{Z}^\top\mathbf{T}_1[\boldsymbol{\Delta}_1\mathbf{Z}^*\mathbf{y}-\Delta_{(0,1)}(\mathbf{1}-\mathbf{z}^*)],
\end{aligned}$$

$\mathbf{y} = (y_1,\ldots,y_n)^\top, \mathbf{z}^* = (z_1^*,\ldots,z_n^*)^\top, \mathbf{1}$ =(1,. . .,1)$^\top$ are $n-$dimensional vectors and $\boldsymbol{\Delta}_0 = \mathrm{diag}\,\{1/\delta_{01},\ldots,1/\delta_{0n}\}\,, \boldsymbol{\Delta}_1 = \mathrm{diag}\,\{1/\delta_{11},\ldots,1/\delta_{1n}\}\,, \boldsymbol{\Delta}_{(0,1)} = \mathrm{diag}\{1/(1-\delta_{01}-\delta_{11}),\ldots,1/(1-\delta_{0n} - \delta_{1n})\}, \mathbf{T}_0 = \mathrm{diag}\{\partial\delta_{01}/\partial\zeta_{01},\ldots,\partial\delta_{0n}/\partial\zeta_{0n}\}$, $\mathbf{T}_1 = \mathrm{diag}\,\{\partial\delta_{11}/\partial\zeta_{11},\ldots,\partial\delta_{1n}/\partial\zeta_{1n}\}\,, \mathbf{T}_{01} = \mathrm{diag}\,\{\partial\delta_{01}/\partial\zeta_{11},\ldots,\partial\delta_{0n}/\partial\zeta_{1n}\}\,$, $\mathbf{T}_{10} = \mathrm{diag}\{\partial\delta_{11}/\partial\zeta_{01},\ldots,\partial\delta_{1n}/\partial\zeta_{0n}\}$. Also, $\mathbf{V}_{n\times k_0}$ and $\mathbf{Z}_{n\times k_1}$ are covariate matrices, where $\mathbf{v}_t$ and $\mathbf{z}_t$ are the correspondents $t-$th rows. We can see that the likelihood equations have no analytical solutions and some nonlinear optimization algorithm should be employed. e.g., NewtonRaphson, Fisher scoring, quasi-Newton algorithms such as BFGS, among others. In this work, we adopted the Fisher scoring algorithm.

The expected Fisher information matrix (see Appendix B), is given by

$$\mathbf{K}(\boldsymbol{\varphi}) = \begin{bmatrix} \mathbf{K}_{\boldsymbol{\rho}\boldsymbol{\rho}} & \mathbf{K}_{\boldsymbol{\rho}\boldsymbol{\psi}} \\ \\ \mathbf{K}_{\boldsymbol{\psi}\boldsymbol{\rho}} & \mathbf{K}_{\boldsymbol{\psi}\boldsymbol{\psi}} \end{bmatrix}, \tag{4.3}$$

where

$$\begin{aligned}
\mathbf{K}_{\boldsymbol{\rho}\boldsymbol{\rho}} &= \mathbf{V}^\top[\mathbf{T}_0^2\boldsymbol{\Delta}_0+\boldsymbol{\Delta}_{(0,1)}(\mathbf{T}_0+\mathbf{T}_{10})^2+\boldsymbol{\Delta}_1\mathbf{T}_{10}^2]\mathbf{V}, \mathbf{K}_{\boldsymbol{\psi}\boldsymbol{\rho}} = \mathbf{K}_{\rho,\boldsymbol{\psi}}^\top\\
\mathbf{K}_{\boldsymbol{\psi}\boldsymbol{\psi}} &= \mathbf{Z}^\top[\mathbf{T}_1^2\boldsymbol{\Delta}_1+\boldsymbol{\Delta}_{(0,1)}(\mathbf{T}_1+\mathbf{T}_{01})^2+\boldsymbol{\Delta}_0\mathbf{T}_{01}^2]\mathbf{Z},\\
\mathbf{K}_{\boldsymbol{\rho}\boldsymbol{\psi}} &= K_{\boldsymbol{\psi}\boldsymbol{\rho}}^\top = \mathbf{Z}^\top[\mathbf{T}_0\boldsymbol{\Delta}_0\mathbf{T}_{01}+(\mathbf{T}_0+\mathbf{T}_{10})\boldsymbol{\Delta}_{(0,1)}(\mathbf{T}_1+\mathbf{T}_{01})+\mathbf{T}_1\boldsymbol{\Delta}_1\mathbf{T}_{10}]\mathbf{V}.
\end{aligned}$$

Then the Fisher scoring algorithm is given by

$$\boldsymbol{\varphi}^{(m+1)} = \boldsymbol{\varphi}^{(m)} + \mathbf{K}^{-1}(\boldsymbol{\varphi}^{(m)})\mathbf{U}(\boldsymbol{\varphi}^{(m)}),$$

where $\boldsymbol{\varphi}^{(m)}$ is the estimate of $\boldsymbol{\varphi}$ at the $m-th$ iteration, $\mathbf{K}(\boldsymbol{\varphi})$ is defined in (4.3), $\mathbf{U}(\boldsymbol{\varphi})$ is defined in (4.2) and $m = 0,1,2,\ldots,$ until $||\boldsymbol{\varphi}^{(m+1)} - \boldsymbol{\varphi}^{(m)}|| < \epsilon, \epsilon > 0$.

4.2 Estimation of the parameters related to the continuous part - EM algorithm

Let $\mathbf{y} = (y_1,\ldots,y_n)^\top$ be the vector of observable responses, $\mathbf{z}^* = (z_1^*,\ldots,z_n^*)^\top$ the vector of observable variables, $\mathbf{u} = (u_1,\ldots,u_n)^\top$ the vector of unobservable variables and $\widehat{\boldsymbol{\vartheta}}^{(m)} = (\widehat{\boldsymbol{\beta}}^{(m)},\widehat{\boldsymbol{\delta}}^{(m)},\widehat{\alpha}^{(m)})^\top$ the estimates of $\boldsymbol{\vartheta}$ in the $m-$th iteration. The EM algorithm proceeds in the following two steps:

E-Step: Calculates the expectation of the log-likelihood concerning the unobservable variables, conditioned on the observed variable and current parameters estimates, that is $Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)}) = \mathbb{E}[\ell_{2c}(\boldsymbol{\vartheta}|\mathbf{y},\mathbf{u},\mathbf{z}^*)|\mathbf{y},\mathbf{z}^*,\widehat{\boldsymbol{\vartheta}}]$.

M-Step: Maximize $Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)})$ concerning to $\boldsymbol{\vartheta}$, obtaining $\widehat{\boldsymbol{\vartheta}}^{(m+1)}$.

where

$$\begin{aligned} Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)}) &= \sum_{t=1}^{n}\{(1-z_t^*)\{\widehat{u}_t^{(m)}\log\theta_t + (1-\widehat{u}_t^{(m)})\log(1-\theta_t) + (1-\widehat{u}_t^{(m)}) \times \\ &\times \ [\log(\Gamma(\phi_t)) - \log(\Gamma(\mu_t\phi_t)) - \log(\Gamma((1-\mu_t)\phi_t)) \\ &+ \ (\mu_t\phi_t - 1)\log y_t + ((1-\mu_t)\phi_t - 1)\log(1-y_t)]\}\} \\ &= \sum_{t=1}^{n} Q_t(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)}), \end{aligned}$$

$$\begin{aligned} \widehat{u}_t &= \mathbb{E}[U_t|y_t, z_t^*, \widehat{\boldsymbol{\vartheta}}] = \mathbb{P}(U_t = 1|y_t, z_t^*, \widehat{\boldsymbol{\vartheta}}) \\ &= \left(\frac{\theta_t}{\theta_t + (1-\theta_t)b(y_t;\mu_t,\phi_t)}\right)^{1-z_t^*}, \end{aligned} \tag{4.4}$$

$\theta_t = 1 - \sqrt{1-4\alpha\gamma_t(1-\gamma_t)}, \mu_t = \dfrac{\gamma_t - \frac{1}{2} + \frac{1}{2}\sqrt{1-4\alpha\gamma_t(1-\gamma_t)}}{\sqrt{1-4\alpha\gamma_t(1-\gamma_t)}}$ and $b(y_t;\mu_t,\phi_t)$.

The optmization of $Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)})$ is not analytically feasible. Then, the estimators must be obtained numerically. We use the optimization function `optim` of the `R` through the L-BFGS-B method (Byrd et al. (1995)) to maximize the function $Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)})$. The EM algorithm is implemented as follows:

E-Step: Given $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$, calculate $u_t^{(m)}$ for $t = 1, \ldots, n$ using (4.4).

M-Step: Update $\widehat{\boldsymbol{\vartheta}}^{(m+1)}$ maximizing $Q(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}^{(m)})$ on $\boldsymbol{\vartheta}$ through the `optim` function, using the L-BFGS-B method.

This approach also applies to the augmented regression model proposed by Ospina (2008), by setting $u_t = 0$ and $\alpha = 0$, which implies that $\theta_t = 0$ in Equation (4.1).

4.3 Standard errors

The asymptotic standard errors of $\widehat{\boldsymbol{\varphi}} = (\widehat{\boldsymbol{\rho}}^\top, \widehat{\boldsymbol{\psi}}^\top)^\top$ can obtained through the inverse of the Fisher information, defined in (4.3), through $\mathrm{EP}(\widehat{\boldsymbol{\varphi}}) = [\mathrm{diag}(\mathbf{K}_{\varphi}^{-1/2}(\widehat{\boldsymbol{\varphi}}))]$.

Since the estimates of the parameters of the continuous part, $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top\alpha)^\top$, are obtained by the EM algorithm, we will approximate the respective asymptotic variance-covariance matrix using the inverse of the empirical information matrix. Louis' missing information principle (Louis, 1982) relates the score function of the incomplete data log-likelihood with the complete data log-likelihood through the conditional expectation $\bigtriangledown_o(\boldsymbol{\vartheta}) = \mathbb{E}_{\boldsymbol{\vartheta}}[\bigtriangledown_c(\boldsymbol{\vartheta};\mathbf{Y}_c|\mathbf{Y}_{obs})]$, where $\bigtriangledown_o(\boldsymbol{\vartheta}) = \partial\ell_o(\boldsymbol{\vartheta};\mathbf{Y}_{obs})/\partial\boldsymbol{\vartheta}$ and $\bigtriangledown_c(\boldsymbol{\vartheta}) = \partial\ell_c(\boldsymbol{\vartheta};\mathbf{Y}_c)/\partial\boldsymbol{\vartheta}$ are the score functions for the incomplete and complete data, respectively. As defined in Meilijson (1989), the empirical information matrix can be computed as

$$\mathbf{I}_e(\boldsymbol{\vartheta}|\mathbf{y}) = \sum_{t=1}^{n}\mathbf{s}(y_t|\boldsymbol{\vartheta})\mathbf{s}(y_t|\boldsymbol{\vartheta})^\top - \frac{1}{n}\mathbf{S}(\mathbf{y}|\boldsymbol{\vartheta})\mathbf{S}(\mathbf{y}|\boldsymbol{\vartheta})^\top, \tag{4.5}$$

where $\mathbf{S}(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{t=1}^{n}\mathbf{s}(y_t|\boldsymbol{\vartheta})$ and $\mathbf{s}(y_t|\boldsymbol{\vartheta}), t = 1, 2, \ldots, n$ is given by

$$\mathbf{s}(y_t|\boldsymbol{\vartheta}) = \mathbb{E}\left[\frac{\partial\ell_{2c}(\boldsymbol{\vartheta}; y_t, z_t^*, u_t)}{\partial\boldsymbol{\vartheta}}|y_t, z_t^*, \boldsymbol{\vartheta}\right].$$

Replacing $\boldsymbol{\vartheta}$ by maximum likelihood estimator $\widehat{\boldsymbol{\vartheta}}$ and considering $\bigtriangledown_o(\widehat{\boldsymbol{\vartheta}}) = \mathbf{0}$, Equation (4.5) takes the simple form $\mathbf{I}_e(\widehat{\boldsymbol{\vartheta}}|\mathbf{y}) = \sum_{t=1}^{n} \mathbf{s}(y_t|\boldsymbol{\vartheta})\mathbf{s}(y_t|\boldsymbol{\vartheta})^\top$.

The empirical score function for the $t-$th observation is decomposed into $\mathbf{s}(y_t|\boldsymbol{\vartheta}) = \big(s_{\boldsymbol{\beta}}(y_t|\boldsymbol{\vartheta}), s_{\boldsymbol{\delta}}(y_t|\boldsymbol{\vartheta}), s_\alpha(y_t|\boldsymbol{\vartheta})\big)^\top$.

Thus, the observed empirical information matrix can be calculated using equations (C.1)–(C.3) in Appendix C. Finally, the variance-covariance matrix is given by $\mathbf{I}_e^{-1}(\widehat{\boldsymbol{\vartheta}}|\mathbf{y})$.

4.4 Hypothesis testing

Under suitable regularity conditions, we have that the maximum likelihood estimators $\widehat{\boldsymbol{\varphi}} = (\widehat{\boldsymbol{\rho}}^\top, \widehat{\boldsymbol{\psi}}^\top)^\top, \widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\delta}}^\top, \widehat{\alpha})^\top$, $\mathbf{K}(\widehat{\boldsymbol{\varphi}})$ and $\mathbf{I}_e(\widehat{\boldsymbol{\vartheta}}|\mathbf{y})$ are consistent for $\boldsymbol{\varphi}$, $\boldsymbol{\vartheta}$, $\mathbf{K}(\boldsymbol{\varphi})$ and $\mathbf{I}_e(\boldsymbol{\vartheta}|\mathbf{y})$, respectively. Assuming that $\mathbf{I}(\boldsymbol{\varphi}) = \lim_{n\longrightarrow\infty}\left\{n^{-1}\mathbf{K}(\boldsymbol{\varphi})\right\}$ and $\mathbf{I}(\boldsymbol{\vartheta}) = \lim_{n\longrightarrow\infty}\left\{n^{-1}\mathbf{I}_e(\boldsymbol{\vartheta}|\mathbf{y})\right\}$ exists and are nonsingular, we have that $\sqrt{n}(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{\mathcal{D}} \mathcal{N}_{k_0+k_1}(\mathbf{0}, \mathbf{I^{-1}}(\boldsymbol{\varphi}))$ and $\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}) \xrightarrow{\mathcal{D}} \mathcal{N}_{p+q+1}(\mathbf{0}, \mathbf{I^{-1}}(\boldsymbol{\vartheta}))$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

For testing the significance of the $i-$th regression parameter, for example, related to γ, we can use the Wald's statistic, $\widehat{\beta_i}/\text{se}(\widehat{\beta_i})$, where $\text{se}(\widehat{\beta_i})$ is the asymptotic standard error of the MLE of β_i obtained from the inverse of observed empirical information matrix evaluated at the maximum likelihood estimates. The limiting null distribution of the test statistic is standard normal. Hypothesis tests on the $\boldsymbol{\rho}$ and $\boldsymbol{\psi}$ can be performed in a similar fashion.

5 Model fit assessment and model comparison

The residual analysis is an important tool for model fit assessment. It is possible, through the residual analysis, checking the presence of outliers, as well as the departing from model assumptions. In this work we use the randomized quantile residual (RQR). Also, we present and compare some statistics for model comparison.

5.1 Randomized quantile residuals

We adapted the randomized quantile residual (Dunn and Smyth (1996)) for our model, which is a randomized version of Cox and Snell (1968) residual, and it is given by $r_t^q = \Phi^{-1}(W_t), \;\; t = 1, \ldots, n$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, W_t is a uniform random variable on the interval $(a_t, b_t]$, with $a_t = \lim_{y\uparrow y_t} \mathrm{F}(y; \widehat{\tau}_t, \widehat{\eta}_t, \widehat{\gamma}_t, \widehat{\phi}_t, \widehat{\alpha})$ and $b_t = \mathrm{F}(y_t; \widehat{\tau}_t, \widehat{\eta}_t, \widehat{\gamma}_t, \widehat{\phi}_t, \widehat{\alpha})$ where $\widehat{\eta}_t = \frac{\widehat{p}_{1t}}{\widehat{\tau}_t}$ and $\widehat{\tau}_t = \widehat{p}_{0t} + \widehat{p}_{1t}$. Here, $\mathrm{F}(y; \widehat{\tau}_t, \widehat{\eta}_t, \widehat{\gamma}_t, \widehat{\phi}_t, \widehat{\alpha})$ is the cdf of ZOABR distribution. For example, for zero-and-one augmented rectangular beta regression model, W_t is a uniform random variable on $(0, \widehat{\tau}(1 - \widehat{\eta})]$ if $y_t = 0$, is a uniform random variable on $(1 - \widehat{\tau}\widehat{\eta}, 1]$ if $y_t = 1$ and $W_t = \mathrm{F}(Y_t; \widehat{\tau}, \widehat{\eta}, \widehat{\gamma}, \widehat{\phi}, \widehat{\alpha})$ if $y_t \in (0, 1)$. Since the variable W_t is no longer continuous, we need to simulate several set of values of r_t^q and take, for example, the respective medians, for each observation. However, since the maximum likelihood estimators are consistent, these medians are expected to follow, approximately, a standard normal distribution. In practice, it is important to simulate at least four sets of RQR.

A plot of RQR against the index of the observations (t) or against the predicted values should present a random pattern. A systematic behavior may suggest a misspecification of the model. Also, a quantile-quantile plot, based on the standard normal distribution, with simulated envelopes, is a helpful diagnostic tool, see Atkinson (1985). The simulated envelopes were constructed simulating standard Normal distribution values.

5.2 Model selection

The Akaike information criterion (AIC) (Akaike (1974)) and the Bayesian information criterion (BIC) (Schwarz (1978)) have been successfully used as statistic tools for model comparison. The AIC is based on the likelihood, penalized by the number parameters, while the BIC includes the sample size. The smaller values of AIC and BIC, the better is the model fit. They are, respectively, defined as $\text{AIC} = -2\log[L(y;\widehat{\boldsymbol{\Upsilon}})] + 2k$ and $\text{BIC} = -2\log[L(y;\widehat{\boldsymbol{\Upsilon}})] + k\log(n)$, where n denotes the sample size and k denotes the number parameters.

6 Simulation studies

In this section, we present simulation studies related to the parameter recovery (Study 1) and another comparing the impact of transforming the values zero and one, in order to be able to consider the rectangular beta regression models, instead using our ZOABR model (Study 2). Also, we perform one study to analyze the behavior of the RQR (Study 3) and another to study the performance of the statistics of model comparison (Study 4).

In Study 1, several relevant scenarios were considered, which correspond to the combination of the levels of some factors of interest. The factors (with the respective levels within parenthesis) are: sample size (n) (50, 100, 500), regression models (rectangular beta, zero augmented rectangular beta, one augmented rectangular beta, zero and one augmented rectangular beta), modeled parameters (mean, mean and dispersion parameter, mean, dispersion and the parameters that model the probabilities of occurrences of zeros and/or ones). Therefore, 33 scenarios were considered, since for the BRr model the probabilities of occurrence of zeros and ones are null. For each scenario, we simulated the response and estimated the parameters under the same model. More details will be provided in Subsection 6.1.

In Study 2, two factors were fixed: the sample size (n) (50, 100, 500) and the percentages of zeros and ones in the sample ($p_0 = p_1 = 1\%$, $p_0 = 5\%$ and $p_1 = 3\%$, $p_0 = 10\%$ and $p_1 = 8\%$, $p_0 = p_1 = 20\%$). The data sets were simulated from the ZOABR, modeling only the mean. In addition, we fitted the regression models ZOABR and BRr, in both modeling only the mean.

Study 3 was subdivided in two studies. In the first one we considered four scenarios, in which we used the ZOABR regression model to simulate the data according to the regression structure defined for each situation and we fitted the ZOABR regression model. The regression structure for this study is defined by $g_1(\gamma_t) = \beta_0 + \beta_1 x_t$ and $g_2(\phi_t) = \delta_0 + \delta_1 w_t$, where $t = 1, \ldots, n$. In Table 1, we present the four scenarios considered, where in scenario C3, $F(\cdot)$ is the cumulative distribution function of the t-Student distribution with $\nu = 4$ degrees of freedom. In the second scenario, we simulated the data of ZOABR regression model, modeling the mean and dispersion parameter, and we fitted the ZOABR and ZOAB regression models, in both modeling the mean and the dispersion parameter.

Study 4 also was subdivided in two studies. In the first one, we simulated the data of ZOABR regression model modeling the mean, dispersion parameter and the parameters that model the probabilities of occurrences of zeros and ones, considering two sample sizes (100 and 500). Then we fitted the ZOABR and ZOAB regression models, modeling all parameters (unless α). In the second study, we simulated the data of ZOAB regression model modeling all parameters, considering two sample sizes (100 and 500), then we fitted ZOABR and ZOAB regression models modeling the mean, dispersion parameter and the parameters that model the probabilities of occurrences of zeros and ones.

The results of Study 1 will be presented only for the ZOABR regression model. In addition, the results of Study 2 will be presented only for the following percentages of zeros and ones

Table 1: Scenarios considered in the Study 3.

Scenarios	Model	
	Simulated	Fitted
C1	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$ $g_2(\phi_t) = -\log(\phi_t)$	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$ $g_2(\phi_t) = -\log(\phi_t)$
C2	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$ $g_2(\phi_t) = -\log(\phi_t)$	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$
C3	$g_1(\gamma_t) = F^{-1}(\gamma_t)$ $g_2(\phi_t) = -\log(\phi_t)$	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$ $g_2(\phi_t) = -\log(\phi_t)$
C4	$g_1(\gamma_t) = \log[-\log(1-\gamma_t)]$ $g_2(\phi_t) = -\log(\phi_t)$	$g_1(\gamma_t) = \log(\gamma_t/(1-\gamma_t))$ $g_2(\phi_t) = -\log(\phi_t)$

in the sample, $p_0 = 5\%$ and $p_1 = 3\%$, $p_0 = 10\%$ and $p_1 = 8\%$. The results related to the other scenarios, for Studies 1 and 2, are presented in the Supplementary Material. The results regarding the residual analysis (first scenario) and model selection will not be presented, for the sake of simplicity. However, they are presented in the Supplementary Material. The results indicated that the RQR perform well in detecting the departing from some model assumptions. Regarding the model selection study, the results indicate that the true underlying model was chosen in at least 99% of the 100 generated replicas.

6.1 Study 1

The true parameters were fixed as: $\rho_0 = -1.8, \rho_1 = 1.5, \psi_0 = -1.8, \psi_1 = 1.5, \beta_0 = -1.5, \beta_1 = 1.5, \delta_0 = -3.0, \delta_1 = -1.8$ and $\alpha = 0.5$. We generated 100 replicas from $Y_t \sim \text{ZOABR}(p_{0t}, p_{1t}, \gamma_t, \phi_t, \alpha)$, considering $\log(p_{0t}/(1-p_{0t}-p_{1t})) = \rho_0 + \rho_1 v_t, \log(p_{1t}/(1-p_{0t}-p_{1t})) = \psi_0 + \psi_1 z_t, \log(\gamma_t/(1-\gamma_t)) = \beta_0 + \beta_1 x_t, \log(\phi_t) = -\delta_0 - \delta_1 w_t$ and $t = 1, \ldots, n$, where x_t, w_t, v_t and z_t were generated independently from a Uniform distribution on $(0, 1)$. We fixed three sample sizes, namely $n = 50, 100, 500$.

Using the estimates obtained in each replica, we calculated the usual statistics to measure the accuracy of the estimates: mean, variance (Var), bias, root mean squared error (RMSE) and absolute value of relative bias (AVRB). Let v be the parameter of interest and let $\hat{v}_r$ be some estimate related to the replica r. The formulas of the adopted statistics are: Mean=$\sum_{r=1}^{R} \frac{\hat{v}_r}{R} = \hat{\hat{v}}_R$, Var=$\sum_{r=1}^{R} \frac{(\widehat{v}_r - \widehat{\widehat{v}}_R)^2}{R-1}$, Bias=$\widehat{\widehat{v}}_R - v$, RMSE=$\sqrt{\frac{1}{R}\sum_{r=1}^{R}(v - \hat{v}_r)^2}$, AVRB=$\frac{|\widehat{\widehat{v}}_R - v|}{|v|}$. The smaller the value of each one of these statistics is, the more accurate the estimate is, except for the mean.

The results of Study 1 are shown in Table 2. Note that as the sample size increases, the better is the accuracy of the estimates, as expected. We can observe that the Bias, Variance and AVRB for β_0, β_1 and α tend to approach to zero as the sample increases (n), indicating that the maximum likelihood estimators are consistent. The same occurs with the other parameters, even though the respective Bias and AVRB are higher compared with the other parameters. In a general way, we can say that the parameters were properly recovered.

Table 2: Mean, Variance, Bias, RMSE and AVRB for the parameters of ZOABR model, under different sample size - Study 1 - $\rho_0 = -1.8, \rho_1 = 1.5, \psi_0 = -1.8, \psi_1 = 1.5, \beta_0 = -1.5, \beta_1 = 1.5, \delta_0 = -3.0, \delta_1 = -1.8$ and $\alpha = 0.5$.

Parameter	n	Mean	Variance	Bias	RMSE	AVRB
ρ_0	50	-1.988	.731	-0.188	.875	.104
	100	-2.041	.522	-0.241	.761	.134
	500	-1.839	.058	-0.039	.245	.021
ρ_1	50	1.699	2.092	.199	1.460	.133
	100	1.805	1.031	.305	1.060	.203
	500	1.559	.136	.059	.373	.039
ψ_0	50	-1.795	.875	.005	.936	.003
	100	-1.869	.293	-0.069	.545	.038
	500	-1.794	.077	.006	.277	.003
ψ_1	50	1.332	2.530	-0.168	1.599	.112
	100	1.621	.740	.121	.869	.081
	500	1.503	.183	.003	.428	.002
β_0	50	-1.537	.072	-0.037	.271	.025
	100	-1.536	.029	-0.036	.174	.024
	500	-1.497	.007	.003	.083	.002
β_1	50	1.541	.125	.041	.356	.027
	100	1.532	.042	.032	.208	.021
	500	1.492	.009	-0.008	.097	.005
δ_0	50	-3.079	1.351	-0.079	1.165	.026
	100	-2.970	.487	.030	.699	.010
	500	-3.024	.082	-0.024	.287	.008
δ_1	50	1.966	4.920	-0.166	2.224	.092
	100	-1.888	1.366	-0.088	1.172	.049
	500	-1.794	.214	.006	.463	.003
α	50	.432	.039	-0.068	.209	.137
	100	.475	.011	-0.025	.110	.050
	500	.495	.003	-0.005	.055	.010

6.2 Study 2

We fixed the values of the parameters as: $\beta_0 = -1.5, \beta_1 = 1.5, \phi = 50$ and $\alpha = 0.5$. Then, we generated 100 replicas from $Y_t \sim$ZOABR$(p_0, p_1, \gamma_t, \phi, \alpha)$ considering $\log(\gamma_t/(1-\gamma_t)) = \beta_0 + \beta_1 x_t$ and $t = 1, \ldots, n$, where x_t was generated from a Uniform distribution on $(0, 1)$. We fixed three sample sizes $n = 50, 100, 500$ and the following percentages of zeros and ones: $(a)\, p_0 = 5\%, p_1 = 3\%$ and $(b) p_0 = 10\%, p_1 = 8\%$. Then, we fitted the ZOABR and the BRr regression models, modeling only the mean in the two models. In the case of the BR model (non-augmented), the observations equal to zero were replaced by 0.001, whereas those equal to one were replaced by 0.999.

For the ZOABR model we present only the results related to continuous part $(\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top, \boldsymbol{\alpha}^\top)^\top$ (once we are comparing them with those obtained by the BRr regression model and these are the unique parameters presented in the two models).

The results are shown in Tables 3 and 4. We can notice that the higher is the percentage of zeros and ones, the less accurate are the estimates associated to the BRr regression model. This behavior is expected, since the higher those quantities are, the greater is the impact of the transformation applied in the data, on the parameter estimation, see Galvis et al. (2014) and Nogarotto et al. (2015). Table 3 shows the results for $p_0 = 5\%$ and $p_1 = 3\%$. Considering the sample sizes equal to 50, 100 and 500, the number of observations equal to zeros and ones in the sample are approximately 4, 8 and 40, respectively. The variance associated with ϕ is much higher for the BRr regression model, under sample sizes equal to 50 and 100. Moreover, as the sample size increases, the estimates tend to be more accurate for the ZOABR regression model.

Table 4 presents the results under $p_0 = 10\%$ and $p_1 = 8\%$. Differently from the previous situation, as the sample size increases, the Bias and AVRB increase for the BRr regression model. In addition, the estimation of ϕ is substantially affected by the data transformation, since the AVRB were very high compared to those related to the ZOABR regression model, although the variance associated with ϕ is higher for the ZOABR regression model. However, the ZOABR regression model performs better according to all other statistics.

Table 3: Mean, Variance, Bias, RMSE and AVRB for the parameters of ZOABR and BRr models, under different sample size - (a) $p_0 = 5\%, p_1 = 3\%$ - Study 2 - $\beta_0 = -1.5, \beta_1 = 1.5, \phi = 50$ and $\alpha = 0.5$.

Parameter	n	Model	Mean	Variance	Bias	RMSE	AVRB
β_0	50	BRr	-1.181	.251	.319	.594	.212
		ZOABR	-1.520	.034	-0.020	.186	.013
	100	BRr	-1.184	.109	.316	.456	.210
		ZOABR	-1.499	.015	.001	.122	.001
	500	BRr	-1.307	.003	.193	.201	.129
		ZOABR	-1.507	.003	-0.007	.054	.005
β_1	50	BRr	1.169	.385	-0.331	.703	.220
		ZOABR	1.525	.044	.025	.212	.017
	100	BRr	1.146	.198	-0.354	.568	.236
		ZOABR	1.497	.022	-0.003	.150	.002
	500	BRr	1.313	.004	-0.187	.198	.125
		ZOABR	1.508	.004	.008	.066	.006
ϕ	50	BRr	33.817	1.166.488	-16.183	37.794	.324
		ZOABR	58.762	594.277	8.762	25.905	.175
	100	BRr	46.093	795.931	-3.907	28.482	.078
		ZOABR	55.880	218.808	5.880	15.918	.118
	500	BRr	55.865	28.470	5.865	7.929	.117
		ZOABR	50.434	22.165	.434	4.728	.009
α	50	BRr	.448	.073	-0.052	.274	.104
		ZOABR	.479	.020	-0.021	.143	.042
	100	BRr	.528	.052	.028	.230	.055
		ZOABR	.491	.007	-0.009	.082	.017
	500	BRr	.635	.001	.135	.139	.271
		ZOABR	.495	.001	-0.005	.036	.011

6.3 Study 3

We will present the results for the second scenario. The interest is to compare the behavior of the residuals, according to the fit of the ZOAB and ZOABR regression models to the simulated data set. We simulate the data from $Y_t \sim$ZOABR$(p_0, p_1, \gamma_t, \phi_t, \alpha)$, considering $\log(\gamma_t/(1-\gamma_t)) = \beta_0 + \beta_1 x_t, \log(\phi_t) = -\delta_0 - \delta_1 w_t$; $t = 1, \ldots, n$, where x_t and w_t were generated independently from a Uniform distribution on $(0, 1)$. Then we fitted the ZOABR and ZOAB regression models, in both modeling the mean and the dispersion parameter.

We observed in the Figures 1(a) and 1(b) that the residuals are not randomly dispersed, and in addition, there is a large concentration of points around zero thus indicating a poor fit of the

Table 4: Mean, Variance, Bias, RMSE and AVRB for the parameters of ZOABR and BRr models using different sample size - (b) $p_0 = 10\%, p_1 = 8\%$ - Study 2 - $\beta_0 = -1.5, \beta_1 = 1.5, \phi = 50$ and $\alpha = 0.5$.

Parameter	n	Model	Mean	Variance	Bias	RMSE	AVRB
β_0	50	BRr	-0.629	.370	.871	1.063	.581
		ZOABR	-1.509	.037	-0.009	.194	.006
	100	BRr	-0.484	.190	1.016	1.106	.678
		ZOABR	-1.506	.019	-0.006	.137	.004
	500	BRr	-0.442	.071	1.058	1.091	.705
		ZOABR	-1.508	.004	-0.008	.065	.005
β_1	50	BRr	.596	.582	-0.904	1.183	.603
		ZOABR	1.507	.059	.007	.243	.005
	100	BRr	.415	.232	-1.085	1.187	.724
		ZOABR	1.501	.029	.001	.172	.001
	500	BRr	.394	.084	-1.106	1.143	.737
		ZOABR	1.508	.005	.008	.074	.005
ϕ	50	BRr	4.394	149.581	-45.606	47.217	.912
		ZOABR	59.027	727.490	9.027	28.443	.181
	100	BRr	2.726	90.759	-47.274	48.225	.945
		ZOABR	53.977	141.181	3.977	12.530	.080
	500	BRr	.932	.061	-49.068	49.068	.981
		ZOABR	50.116	28.682	.116	5.357	.002
α	50	BRr	.386	.120	-0.114	.365	.227
		ZOABR	.487	.019	-0.013	.139	.026
	100	BRr	.464	.114	-0.036	.340	.072
		ZOABR	.493	.009	-0.007	.094	.013
	500	BRr	.426	.097	-0.074	.319	.149
		ZOABR	.491	.002	-0.009	.048	.018

ZOAB regression model to the data. However, in Figures 2(a) and 2(b) it is possible to note a random pattern of the residuals, indicating the non-existence of a trend in observations, which suggests a good fit of the ZOABR regression model to the data. The histogram (see Figure 1(c)) indicates a possible asymmetry to the left of the residuals, whereas the histogram (see Figure 2(c)) suggests a symmetry of the residuals. In the Figure 1(d), we note that there is initially a increasing trend, then there seems to be a decreasing trend and finally a increasing trend again. In addition, there appear to be two concavities, one facing up and the other down, causing the residuals to leave the confidence bands of the simulated envelopes. However, in the Figure 2(d), there is no strong evidence to depart from the assumption that the ZOABR regression model is adequate for the data, since most of the residuals remain within the confidence bands of the simulated envelopes. In addition, there is no evidence of trend within the bands of confidence.

7 Application

The analyzed data set was obtained from Carlstrom et al. (2000) which is available from `http://www.stat.ucla.edu/projects/datasets/risk_perception.html`. It corresponds to a psychometric study of risk perception. The part that we are interested in this study corresponds to 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 is, the higher the risk perceived

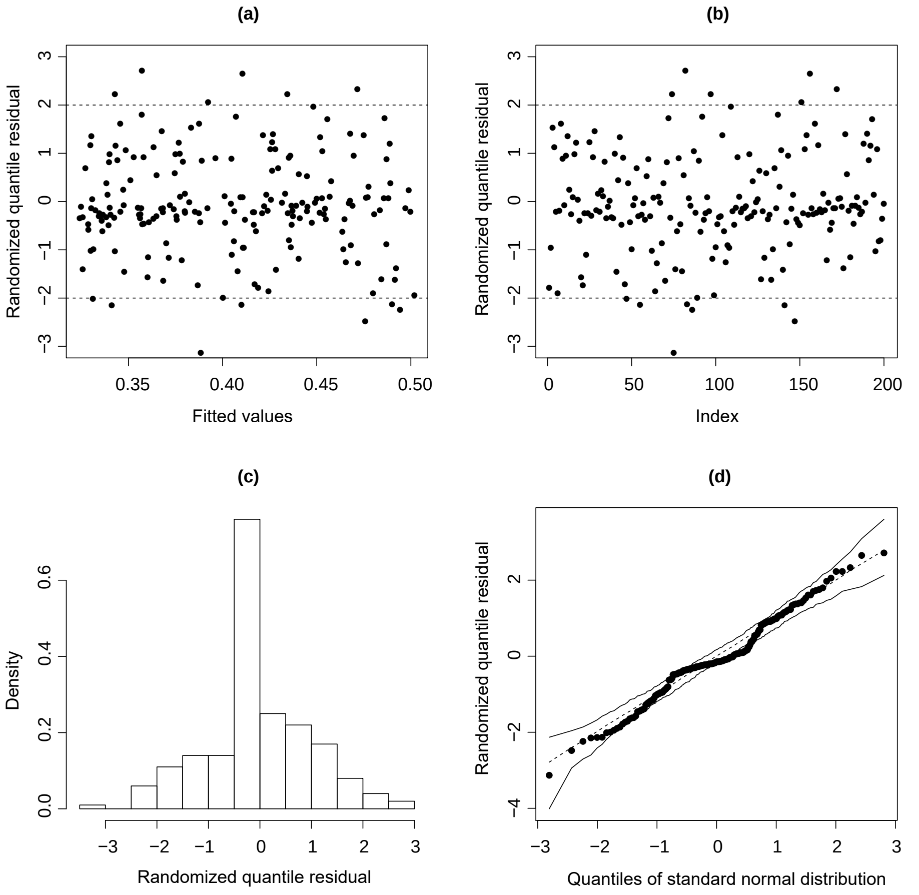

Figure 1: Residual plots fitting the ZOAB model.

is, being 0 non-risk and 100 the maximum risk. In order to use the ZOABR model, the observations were transformed to the interval [0,1]. Also, several covariables were measured and we aim to study their impact on the risk perception. The covariables 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).

We analyzed the perception of the subjects about the risk related to a screening for genes that may predispose subjects to heart disease. We have a total of 588 observations, being 86 equal to zero and 21 to one. That is, approximately 14.63% of the participants provide a null risk perception whereas 3.57% provide a maximum risk.

We started fitting the ZOABR and ZOAB regression models including all covariables that were apparently significant through a descriptive analysis (the related results will not be presented for the sake of simplicity) without interactions. Assuming $Y_{tijk} \overset{ind.}{\sim}$ ZOABR$(p_{0tijk}, p_{1tijk}, \gamma_{tijk}, \phi_{tijk}, \alpha)$ for the ZOABR model and $Y_{tijk} \overset{ind.}{\sim}$ ZOAB$(p_{0tijk}, p_{1tijk}, \gamma_{tijk}, \phi_{tijk})$ for the ZOAB model, the initial structure for the linear predictors, for both models, is:

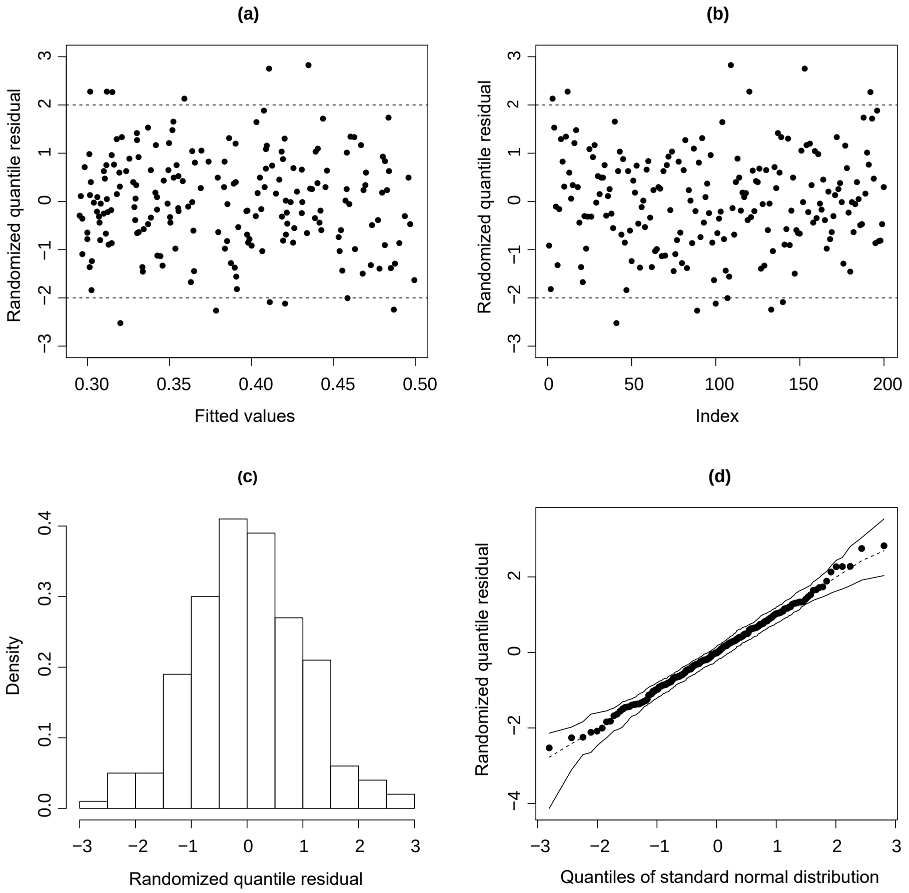

Figure 2: Residual plots fitting the ZOABR model.

$$
\begin{aligned}
\log(\gamma_{tijk}/(1-\gamma_{tijk})) &= \mu_1 + (\beta_1)_i + (\beta_2)_j + (\beta_3)_k, \\
\log(\phi_{tijk}) &= -\mu_2 - (\delta_1)_i - (\delta_2)_j - (\delta_3)_k, \\
\log(p_{0tijk}/(1-p_{0tijk}-p_{1tijk})) &= \mu_3 + \rho_1 x_{tijk} + (\rho_2)_k, \\
\log(p_{1tijk}/(1-p_{0tijk}-p_{1tijk})) &= \mu_4 + \psi_1 x_{tijk} + (\psi_2)_k, \qquad (7.1)
\end{aligned}
$$

where $t = 1, \ldots, n$; i=1 (female), 2 (male); j=1 (caucasian), 2 (African-American), 3 (Mexican-American), 4 (Taiwanese-American); k=1 (unclassifiable), 2 (individualist), 3 (hierarchicalist), 4 (egalitarian) and $(\beta_1)_0 = (\beta_2)_1 = (\beta_3)_0 = 0$, $(\delta_1)_0 = (\delta_2)_1 = (\delta_3)_0 = 0$, $(\rho_2)_0 = 0$, $(\psi_2)_0 = 0$. The parameters $(\beta_1, \delta_1), (\beta_2, \delta_2)$ and $(\beta_3, \delta_3, \rho_2, \psi_2)$ are related to gender, ethnicity and wvcat, respectively and x_{tijk} is the age subject t, from gender i, ethnicity j and world view k.

Considering the regression structure defined in (7.1), we fitted the ZOAB model. Then we excluded all non significant covariables, combining the equivalent levels within each significant covariable (according to the non-significance of the respective parameters). The selected regression structure is:

$$\begin{array}{rcl}\log(\gamma_{tijk}/(1-\gamma_{tijk})) & = & \mu_1 + (\beta_2)_j, \\ \log(\phi_{tijk}) & = & -\mu_2 - (\delta_2)_j, \\ \log(p_{0tijk}/(1-p_{0tijk}-p_{1tijk})) & = & \rho_0 + \rho_1 x_{tijk}, \\ \log(p_{1tijk}/(1-p_{0tijk}-p_{1tijk})) & = & \psi_0 + \psi_1 x_{tijk},\end{array}$$

where $(\beta_2)_1 = (\beta_2)_3 = (\beta_2)_4 = 0$, $(\delta_2)_1 = (\delta_2)_3 = (\delta_2)_4 = 0$. Therefore, for the mean and the dispersion parameters, we have only the effect of ethnicity, being only the group African-American different from the others. For the p_{0tijk} and p_{1tijk}, only "age" was significant.

Also, considering the regression structure defined in (7.1), we fitted the ZOABR model, following the same steps as those for the ZOAB model. Thus, the final regression structure was:

$$\begin{array}{rcl}\log(\gamma_{tijk}/(1-\gamma_{tijk})) & = & \mu_1 + (\beta_2)_j, \\ \log(\phi_{tijk}) & = & -\mu_2 - (\delta_2)_j, \\ \log(p_{0tijk}/(1-p_{0tijk}-p_{1tijk})) & = & \rho_0 + \rho_1 x_{tijk}, \\ \log(p_{1tijk}/(1-p_{0tijk}-p_{1tijk})) & = & \psi_0 + \psi_1 x_{tijk},\end{array}$$

where $(\beta_2)_1 = (\beta_2)_3 = 0$, $(\delta_2)_1 = (\delta_2)_3 = 0$, $j = 1, 2, 3, 4$. Therefore for the mean and for the dispersion parameters, only the covariable ethnicity was significant, being the African-American and Taiwanese-American levels different from the others and from each other. For the p_{0tijk} and p_{1tijk} only the covariable age was significant. We can see that the final regression structures for the two models were different, pointing out that different inferences can be drawn from these two models.

For the ZOAB model we obtained AIC= 500.24, BIC = 535.25, whereas for the ZOABR model we obtained AIC = 487.41 and BIC = 535.55, which indicate that the ZOABR model presents the best fit.

In Figure 3, we present the histogram of the predicted distribution for the ZOAB and ZOABR models. In the histograms, the bar with the dot above represents the zeros and ones. From Figure 3(a) we can notice that the ZOABR model performs better on the right tail when compared to the ZOAB model.

In Figures 4 and 5, we present residual analysis for the ZOAB and ZOABR models, respectively. Concerning the ZOAB regression model, we observed in Figure 4(a) a different variability along the fitted values, indicating that only a part of the variability of the data was captured by the model. From Figure 4(d), we can notice that, although most of the residuals are within the confidence bands, some are outside or close to the limits, especially in the tails of the distribution. This behavior is, probably, due to the observations in the tails, that were not well accommodated by the ZOAB model. Concerning the ZOABR model, from Figure 5(a), we can notice a behavior similar to that for the residuals related to the ZOAB regression model. However, from Figure 5(d) we can see that all residuals are well within the confidence bands and those observations that highlighted in the ZOAB model, were well accommodated by the ZOABR model.

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 beta and rectangular beta models. Figure 6 presents the simulated envelopes for the two models. We can notice that many of the residuals are out of the confidence bands and they present a systematic behavior. Then, we can conclude that the non-augmented models are not suitable to analyze this data, even transforming the zero/one observations.

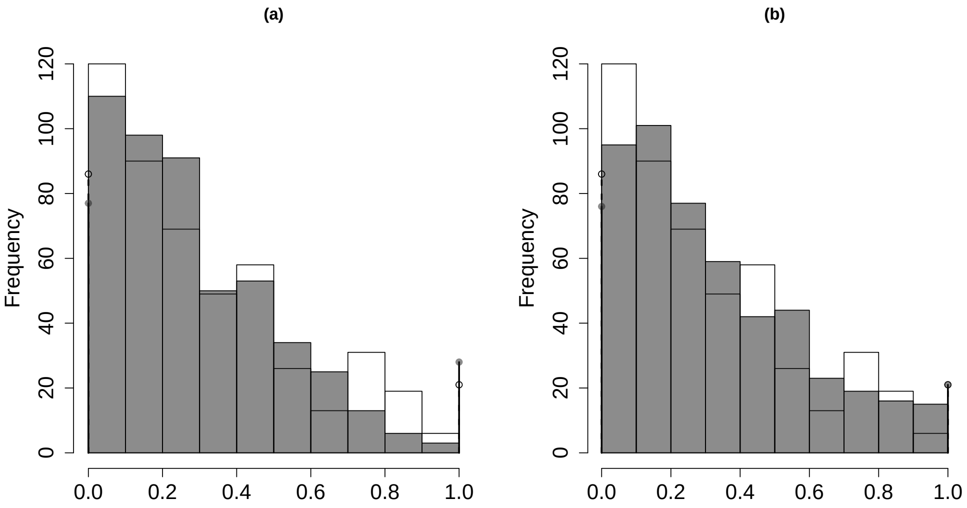

Figure 3: Histogram of the predicted distribution for the regression model: (a) ZOAB (b) ZOABR.

Also, depending on the model, some of the parameters are not significant (for the sake of simplicity, we did not present the results for the non-augmented models). This is also an important aspect that illustrates how the use of non-augmented models, in the transformed data, can lead to misleading inference.

Finally, Tables 5 and 6 presents the estimates, standard errors (SE), 95% equi-tailed confidence intervals (CI), test statistic and p-value of the ZOABR and ZOABR final models, respectively. All parameters are significant at a significance level of 5%. In Table 6, we can see that the estimate of α indicates that its true value is around 0.5, which, in its turn, suggest heavy tails for the conditional distribution of the response. According to the ZOABR model and the estimates of $((\beta_2)_2, (\beta_2)_4)^\top$, it is possible to conclude that the risk perceived is higher for the African-American and Taiwanese-American groups compared with the Caucasian and Mexican-American ones. Also, from the estimates of $((\delta_2)_2, (\delta_2)_4)^\top$, it is possible to conclude that the dispersion of the risk perceived is smaller for the African-American and Taiwanese-American groups compared with the Caucasian and Mexican-American groups. Finally, from the estimates of $(\rho_0, \rho_1)^\top$ and $(\psi_0, \psi_1)^\top$ we obtain that the proportion of the subjects that provide a non-risk perception is 0.1463 (14.64%) whereas 0.0357 (3.57%) provides a maximum risk.

Table 5: Parameter estimates, standard errors, 95% confidence intervals, test statistic and p-value for the ZOAB final model.

Parameter	Estimate	SE	CI(95%)	Statistic	p-value
μ_1	-0.703	.052	[-0.805; -0.601]	-13.519	< 0.001
$(\beta_2)_2$	.397	.123	[.156; .638]	3.228	0.001
μ_2	-1.119	.065	[-1.246; -0.992]	-17.215	< 0.001
$(\delta_2)_2$	.309	.140	[.035; .583]	2.207	0.027
ρ_0	-2.418	.265	[-2.937; -1.899]	-9.125	< 0.001
ρ_1	.024	.008	[.008; .040]	3	0.003
ψ_0	-4.548	.505	[-5.538; -3.558]	-9.006	< 0.001
ψ_1	.044	.012	[.020; .067]	3.667	< 0.001

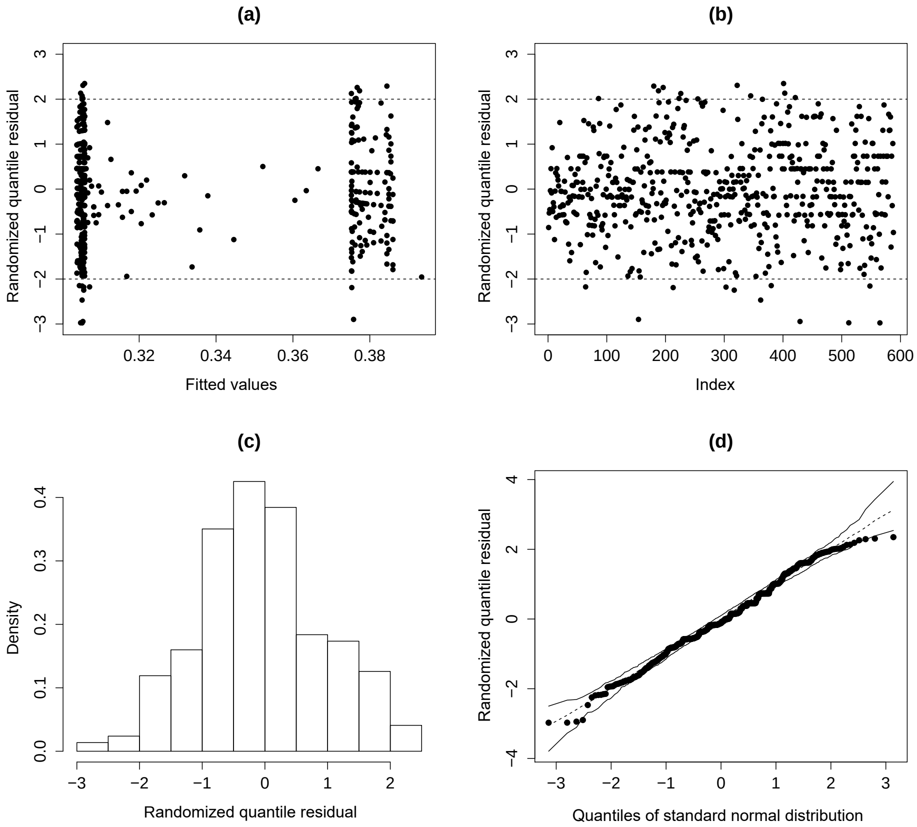

Figure 4: Residual plots for the ZOAB model.

Table 6: Parameter estimates, standard errors, 95% confidence intervals, test statistic and p-value for the ZOABR final model.

Parameter	Estimate	SE	CI(95%)	Statistic	p-value
μ_1	-0.828	.123	[-1.069; -0.587]	-6.732	< 0.001
$(\beta_2)_2$	.494	.153	[.194; .794]	3.229	0.001
$(\beta_2)_4$	.245	.121	[.008; .482]	2.025	0.043
μ_2	-1.729	.139	[-2.001; -1.457]	-12.439	< 0.001
$(\delta_2)_2$	.782	.233	[.325; 1.239]	3.356	0.001
$(\delta_2)_4$	.578	.215	[.157; .999]	2.688	0.007
α	.450	.166	[.125; .775]	2.711	0.007
ρ_0	-2.418	.265	[-2.937; -1.899]	-9.125	< 0.001
ρ_1	.024	.008	[.008; .040]	3	0.003
ψ_0	-4.548	.505	[-5.538; -3.558]	-9.006	< 0.001
ψ_1	.044	.012	[.020; .0675]	3.667	< 0.001

8 Concluding remarks

We developed a zero-and/or-one augmented rectangular beta regression model as a natural extension of rectangular beta regression model proposed by Bayes et al. (2012). The proposed model has practical applicability in modeling of proportions, rates or fractions data in the

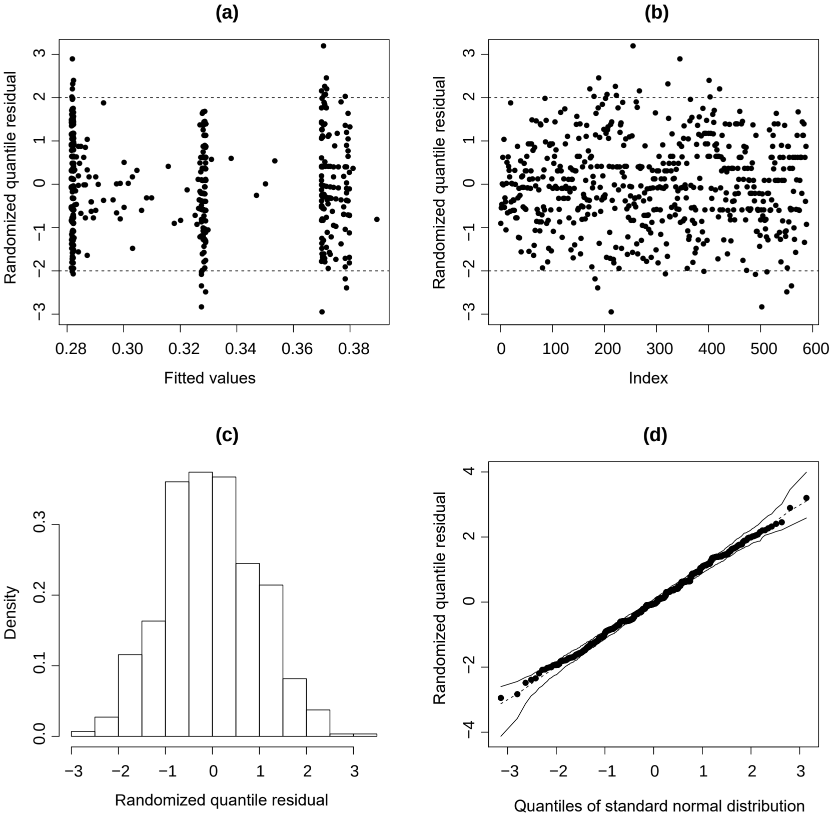

Figure 5: Residual plots for the ZOABR model.

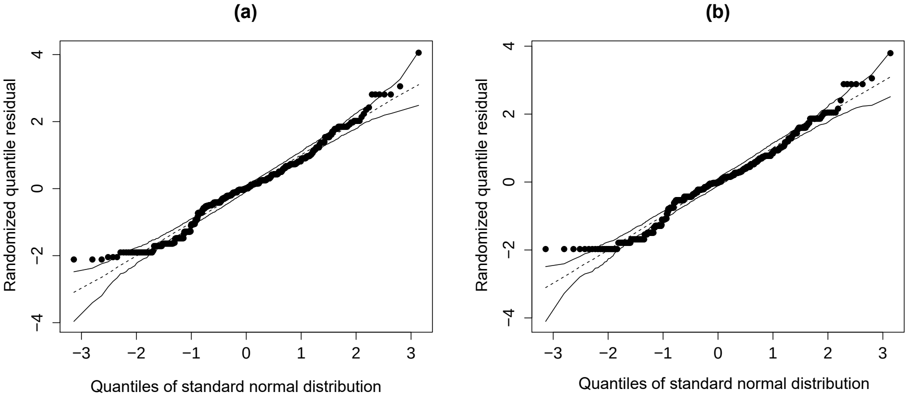

Figure 6: Residual plots for the (a) beta and (b) rectangular beta models.

presence of zeros and ones. In this work, we model the mean, the precision parameter and the probabilities of occurrences of ones and zeros, through linear predictors, using appropriate link functions.

Different to other approaches, we developed a two-step algorithm for maximum likelihood estimation, where the parameters of the discrete part where estimated by using a Fisher scoring algorithm, whereas those related to the continuous part were estimated through the EM algorithm. The respective standard errors were also obtained. For testing the significance of the regression parameter, we present hypothesis testing tools based on the Wald's statistic.

In addition, we propose randomized quantile residuals (RQR) for our model. Also, we conducted simulation studies which showed that: (i) the model and the estimation algorithm recovery the parameters properly, (ii) the statistics of model comparison (AIC and BIC) indicate with great accuracy the true underlying model, (iii) the RQR perform well in the model fit assessment and (iv) less accurate results are obtained when the non augmented models are used in the transformed data (replacing zero/one by convenient values).

The analysis of real data set indicates that our model accommodates properly the observations in the tails of distributions and better than the beta regression model. Also, it shows that misleading inference can be obtained, when the non-augmented models are used in the transformed data.

As future research we suggest developing local influence analysis and the use of other distributions for the continuous part, as the simplex distribution. Also, Item Response Theory (IRT) models for continuous-limited responses, augmented in zeros and ones, can be developed by using the results presented in this work.

Acknowledgement(s)

We gratefully acknowledge São Paulo Research Foundation (FAPESP), for the financial support of this project, through a Master's scholarship, grant number 2013/07850-0, granted to the first author under the guidance of the second.

References

Akaike, H. (1974). A new look at the statistical model identification, *Automatic Control, IEEE Transactions on*, **19**, 6, 716–723.

Atkinson, A. C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis.* Clarendon Press Oxford.

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

Bayes, C. L. and Bazán, J. L. (2014). An EM algorithm for Beta-Rectangular Regression Models, *Personal Communication.*

Bayes, C. L. and Valdivieso, L.. (2016). A beta inflated mean regression model for fractional response variables, *Journal of Applied Statistics*, **43**, 10, 1814–1830.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific Computing*, **16**, 5, 1190–1208.

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**, 3, 385-392.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R, *J Stat Software*, **34**, 1–24.

Cribari-Neto, F. and Souza, T. C. (2012). Testing inference in variable dispersion beta regressions, *Journal of Statistical Computation and Simulation*, **82**, 12, 1827–1843.

Cribari-Neto, F. and Pereira, T. L. (2014). Detecting model misspecification in inflated beta regressions, *Communications in Statistics-Simulation and Computation*, **43**, 3, 631–656.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 3, 236–244.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions, *Journal odf applied Statistics*, **31**, 7, 799–815.

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

García, C. B., Pérez, J. G., and Van Dorp, J. R. (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support, *Statistical Methods & Applications*, **20**, 4, 463–486.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**, 2, 226–233.

Ma, Z. and Leijon, A. (2011). Bayesian Estimation of Beta Mixture Models with Variational Inference, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 11, 2160–2173.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *Journal of the Royal Statistical Society. Series B (Methodological)*, **51**, 1, 127–138.

Nogarotto, D. C., Azevedo, C. L. N., and Bazán, J.L. (2015). *Bayesian estimation, residual analysis and prior sensitivity study for zero-one augmented beta regression model with an application to psychometric data.* Technical Report, `http://www.ime.unicamp.br/sites/default/files/rp14_2016.pdf`, University of Campinas.

Ospina, R. (2008). *Inflated Beta Regression Models. Doctoral's Thesis (In Portuguese)* IME-USP.

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

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

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.

Pereira, T. L. (2010). *Inflated beta regression momdels: inference and applications.* PhD thesis (In Portuguese), Universidade Federal de Pernambuco.

Pereira, G. H. A. (2012). *Truncated and inflated beta regression models.* PhD thesis (In Portuguese), Universidade de São Paulo.

Schwarz, G. (1978). Estimating the dimension of a model, *The annals of statistics*, **6**, 2, 461–464.

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

Souza, T. C., Cribari-Neto, F., and Lima, M. C. V. (2016). Detecting model misspecification in inflated beta regressions, *Communications in Statistics-Simulation and Computation*, **45**, 2, 625–642.

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

Wang, J. and Luo, S. (2015). Bayesian multivariate augmented Beta rectangular regression models for patient-reported outcomes and survival data, *Statistical methods in medical research*, 1–20.

A Score vector for discrete components of the model

In this section, we present the score function for discrete components of the ZOABR model. From (3.1) the system of equations $(\zeta_{0t}, \zeta_{1t}) = (h_0(p_{0t}, p_{1t}), h_1(p_{0t}, p_{1t}))$ defines an one by one transformation in such way that we can solve, the equations $\zeta_{0t} = h_0(p_{0t}, p_{1t})$ and $\zeta_{1t} = h_1(p_{0t}, p_{1t})$ in terms of p_{0t} and p_{1t}. We denote this inverse transformation by $p_{0t} = h_0^*(\zeta_{0t}, \zeta_{1t}), p_{1t} = h_1^*(\zeta_{0t}, \zeta_{1t})$.

Consider the parameter $\boldsymbol{\varphi} = (\boldsymbol{\rho}^\top, \boldsymbol{\psi}^\top)^\top$, the elements of the score vector are given by

$$\begin{aligned} U_R &= \frac{\partial \ell_1(\boldsymbol{\rho}, \boldsymbol{\psi})}{\partial \rho_R} = \sum_{t=1}^{n} \frac{\partial \ell_t(p_{0t}, p_{1t})}{\partial p_{0t}} \frac{\partial p_{0t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_R} + \frac{\partial \ell_t(p_{1t}, p_{1t})}{\partial p_{1t}} \frac{\partial p_{1t}}{\partial \zeta_{0t}} \frac{\partial \zeta_{0t}}{\partial \rho_R} \\ &= \sum_{t=1}^{n} \left\{ \frac{z_t^*(1-y_t)}{p_{0t}} - \frac{(1-z_t^*)}{1-p_{0t}-p_{0t}} \right\} \frac{\partial p_{0t}}{\partial \zeta_{0t}} v_{tR} + \sum_{t=1}^{n} \left\{ \frac{z_t^* y_t}{p_{1t}} - \frac{(1-z_t^*)}{1-p_{0t}-p_{0t}} \right\} \frac{\partial p_{1t}}{\partial \zeta_{0t}} v_{tR} \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} U_S &= \frac{\partial \ell_1(\boldsymbol{\rho}, \boldsymbol{\psi})}{\partial \psi_S} = \sum_{t=1}^{n} \frac{\partial \ell_t(p_{0t}, p_{1t})}{\partial p_{0t}} \frac{\partial p_{0t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \psi_S} + \frac{\partial \ell_t(p_{1t}, p_{1t})}{\partial p_{1t}} \frac{\partial p_{1t}}{\partial \zeta_{1t}} \frac{\partial \zeta_{1t}}{\partial \psi_S} \\ &= \sum_{t=1}^{n} \left\{ \frac{z_t^*(1-y_t)}{p_{0t}} - \frac{(1-z_t^*)}{1-p_{0t}-p_{0t}} \right\} \frac{\partial p_{0t}}{\partial \zeta_{1t}} z_{tS} + \sum_{t=1}^{n} \left\{ \frac{z_t^* y_t}{p_{1t}} - \frac{(1-z_t^*)}{1-p_{0t}-p_{0t}} \right\} \frac{\partial p_{1t}}{\partial \zeta_{1t}} z_{tS}, \end{aligned} \tag{A.2}$$

where $R = 1, \ldots, k_0$ and $S = 1, \ldots, k_1$.

B Expected Fisher information matrix for discrete components of the model

In this section, we present the expected Fisher information matrix for discrete components of the ZOABR model. For $S, S' = 1, \dots, k_1$ and $R, R' = 1, \dots, k_0$, we have

$$\begin{aligned} K_{RR'} &= \sum_{t=1}^{n}\left[\frac{1}{p_{0t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\left(\frac{\partial p_{0t}}{\partial \zeta_{0t}}\right)^2 v_{tR}v_{tR'} + 2\sum_{t=1}^{n}\left[\frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{1t}}{\partial \zeta_{0t}}\frac{\partial p_{0t}}{\partial \zeta_{0t}} v_{tR}v_{tR'} \\ &+ \sum_{t=1}^{n}\left[\frac{1}{p_{1t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\left(\frac{\partial p_{1t}}{\partial \zeta_{0t}}\right)^2 v_{tR}v_{tR'}, \end{aligned} \tag{B.1}$$

$$\begin{aligned} K_{SS'} &= \sum_{t=1}^{n}\left[\frac{1}{p_{0t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\left(\frac{\partial p_{0t}}{\partial \zeta_{1t}}\right)^2 z_{tS}z_{tS'} + 2\sum_{t=1}^{n}\left[\frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{0t}}{\partial \zeta_{1t}}\frac{\partial p_{1t}}{\partial \zeta_{1t}} z_{tS}z_{tS'} \\ &+ \sum_{t=1}^{n}\left[\frac{1}{p_{1t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\left(\frac{\partial p_{1t}}{\partial \zeta_{1t}}\right)^2 z_{tS}z_{tS'} \end{aligned} \tag{B.2}$$

and

$$\begin{aligned} K_{RS} &= \sum_{t=1}^{n}\left[\frac{1}{p_{0t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{0t}}{\partial \zeta_{1t}}\frac{\partial p_{0t}}{\partial \zeta_{0t}} z_{tS}v_{tR} + \sum_{t=1}^{n}\left[\frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{1t}}{\partial \zeta_{1t}}\frac{\partial p_{0t}}{\partial \zeta_{0t}} z_{tS}v_{tR} \\ &+ \sum_{t=1}^{n}\left[\frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{0t}}{\partial \zeta_{1t}}\frac{\partial p_{1t}}{\partial \zeta_{0t}} z_{tS}v_{tR} + \sum_{t=1}^{n}\left[\frac{1}{p_{1t}} + \frac{1}{(1-p_{0t}-p_{1t})}\right]\frac{\partial p_{1t}}{\partial \zeta_{1t}}\frac{\partial p_{1t}}{\partial \zeta_{0t}} z_{tS}v_{tR}. \end{aligned} \tag{B.3}$$

C Observed empirical information matrix

In this section, we present the observed empirical information matrix for continuous components of the ZOABR model.

$$\begin{aligned} s_{\boldsymbol{\beta}}(y_t|\boldsymbol{\vartheta}) &= (1-z^*)\left\{\frac{2\alpha(1-2\gamma_t)}{(1-\theta_t)}\left[\frac{\widehat{u}_t}{\theta_t} - \frac{(1-\widehat{u}_t)}{(1-\theta_t)}\right] + (1-\widehat{u}_t)\right. \\ &\times \left.(1-\alpha)\frac{\phi_t}{(1-\theta_t)^3}(y_t^* - \mu_t^*)\right\}\frac{1}{g'(\gamma_t)}\mathbf{x}_t, \end{aligned} \tag{C.1}$$

$$s_{\boldsymbol{\delta}}(y_t|\boldsymbol{\vartheta}) = -(1-z_t^*)(1-\widehat{u}_t)[\mu_t(y_t^* - \mu_t^*) + (y_t^+ - \mu_t^+)]\frac{1}{g'(\phi_t)}\mathbf{w}_t \tag{C.2}$$

and

$$\begin{aligned} s_{\alpha}(y_t|\boldsymbol{\vartheta}) &= (1-z_t^*)\left\{\frac{2\gamma_t(1-\gamma_t)}{(1-\theta_t)}\left[\frac{\widehat{u}_t}{\theta_t} - \frac{(1-\widehat{u}_t)}{(1-\theta_t)}\right]\right. \\ &- \left.\left[\frac{\phi_t\gamma_t(1-\gamma_t)(1-2\gamma_t)}{(1-\theta_t)^3}\right](y_t^* - \mu_t^*)\right\}, \end{aligned} \tag{C.3}$$

where $\widehat{u}_t$ is given by the Equation (4.4), $\dfrac{1}{g'(\gamma_t)} = \gamma_t(1-\gamma_t)$ and $\dfrac{1}{g'(\phi_t)} = \phi_t$, $\mathbf{x}_t = (x_{t1}, ...x_{tp})^\top$ and $\mathbf{w}_t = (w_{t1}, ...w_{tk})^\top$ are p and k covariable vectors, respectively.