

Augmented Beta rectangular regression models: A Bayesian perspective

Jue Wang and Sheng Luo*

Department of Biostatistics, The University of Texas Health Science Center at Houston, 1200 Pressler St, Houston, TX 77030, USA

Received 20 October 2014; revised 4 May 2015; accepted 12 May 2015

Mixed effects Beta regression models based on Beta distributions have been widely used to analyze longitudinal percentage or proportional data ranging between zero and one. However, Beta distributions are not flexible to extreme outliers or excessive events around tail areas, and they do not account for the presence of the boundary values zeros and ones because these values are not in the support of the Beta distributions. To address these issues, we propose a mixed effects model using Beta rectangular distribution and augment it with the probabilities of zero and one. We conduct extensive simulation studies to assess the performance of mixed effects models based on both the Beta and Beta rectangular distributions under various scenarios. The simulation studies suggest that the regression models based on Beta rectangular distributions improve the accuracy of parameter estimates in the presence of outliers and heavy tails. The proposed models are applied to the motivating Neuroprotection Exploratory Trials in Parkinson's Disease (PD) Long-term Study-1 (LS-1 study, $n = 1741$), developed by The National Institute of Neurological Disorders and Stroke Exploratory Trials in Parkinson's Disease (NINDS NET-PD) network.

Keywords: Augmented Beta; Beta rectangular distribution; GAMLSS family; Longitudinal data; Markov chain Monte Carlo; Proportional data.



Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

In many clinical trials or biomedical studies, researchers often collect some outcomes in the form of percentages, proportions, fractions, and rates measured in the open unit interval $(0, 1)$ (referred to as proportional data) (Kieschnick and McCullough, 2003). Examples include the Alzheimer's disease assessment scale (range from 0 to 70, which can be transformed into the unit interval) (Rogers et al., 2012), microbial data expressed as percentages (Zhao et al., 2001), and proportion of parasitized eggs in a biological control assay (Vieira et al., 2000). Beta regression models (Ferrari and Cribari-Neto, 2004) are increasingly used by researchers from various fields to directly model the covariate effects on the proportional response through a generalized linear model (GLM) framework. When the outcomes are measured longitudinally, Beta regression models with random effects have been proposed to account for the within-subject correlation (Verkuilen and Smithson, 2012; Figueroa-Zúniga et al., 2013). However, because the support of the Beta distributions is between zero and one, boundary values zero and one are not allowed in Beta regression models. Several previous studies have discussed this issue through various approaches. Smithson and Verkuilen (2006) introduced a Lemon-squeezer

*Corresponding author: e-mail: sheng.t.luo@uth.tmc.edu; Phone: +1-713-500-9554

(LS) transformation that first linearly transforms the variable from the original scale to the open unit interval $(0, 1)$ and then compresses the range to avoid zeros and ones. The rescaling might work nicely for small proportions of zeros and/or ones, but the parameter estimates can be sensitive with higher proportions (Galvis et al., 2014). Such inefficiency may be even worse with the presence of within-subject correlation or multilevel clustering in typical longitudinal data. Ospina and Ferrari (2010) proposed a mixed continuous-discrete distribution to capture the probabilities at zero and one, by using a continuous distribution on the open unit interval $(0, 1)$ and a degenerate distribution that assigns values of zero and one with nonnegative probabilities. This approach allows one to directly model the variable without transformation. Using the same idea, Galvis et al. (2014) proposed a generalized linear-mixed model framework by augmenting the probabilities of zeros and ones to the Beta regression model via a zero-one-augmented Beta (ZOAB) model. They termed the model as “augmented” model rather than “inflated” model because the Beta distribution does not include zero and one in its support, similar in spirit to Hatfield et al. (2011, 2012).

The Beta regression models are based on Beta distributions, which can take on a variety of shapes to account for nonnormality and skewness in proportional data by considering different values of its parameters (Johnson et al., 1994). However, as noted by Hahn (2008), García et al. (2011), and Bayes et al. (2012), the Beta distribution considers neither the tail area events nor the outlying events, fails to represent excess variability and overoccurrence of tail-area events, which could limit its applications for modeling the proportional data. Various methods have been proposed to address the issue of outlying and tail area events. Rigby and Stasinopoulos (2005) proposed a generalized additive models for location, scale, and shape (GAMLSS) family and developed an R package GAMLSS (Stasinopoulos and Rigby, 2007), which allows all parameters of the response variable’s distribution to be modeled as linear and nonlinear functions of the covariates. In the current context, the zero and one inflated Beta regression model (BEINF) in the GAMLSS family can be used for modeling proportional data in $[0, 1]$. Additionally, Hahn (2008) proposed a Beta rectangular distribution, which is a mixture distribution consisting of a Beta distribution and a uniform (rectangular) distribution between 0 and 1. The Beta rectangular distribution reduces to the Beta distribution when the mixture probability is 0. Comparing to the Beta distribution, the Beta rectangular distribution assigns more weight to extremal tail-area events, and more probability to the outliers and extremal events. Bayes et al. (2012) proposed a Beta rectangular regression model for cross-sectional data and obtained more robust inference against outlying observations than the Beta regression. However, the Beta rectangular model accounts for neither the within-subject correlation in longitudinal data nor the boundary values zeros and ones.

In this article, we generalize the model by Bayes et al. (2012) and develop an augmented Beta rectangular regression model to account for the occurrence of boundary values 0 and 1 in the closed unit interval $[0, 1]$. Moreover, we account for the within-subject correlation in the longitudinal data by introducing random effects under the generalized linear-mixed model framework. The rest of the article proceeds as follows. In Section 2, we describe a motivating clinical trial, the proportional outcome variable of interest, and the issue of outlying observations. In Section 3, we briefly review the Beta and Beta rectangular distributions and their regression models, and develop the augmented Beta rectangular regression model. In Section 4, we discuss the Bayesian inference and Bayesian model selection criteria. In Section 5, we conduct an extensive simulation study with three settings to compare the performance of various models. In Section 6, we apply the proposed model to the motivating clinical trial. Concluding remarks and discussions are given in Section 7.

2 A motivating clinical trial

This methodological development is motivated by the Parkinson’s Disease (PD) Long-term Study-1 (LS-1 study, $n = 1741$), developed by The National Institute of Neurological Disorders and Stroke Exploratory Trials in Parkinson’s Disease (NINDS NET-PD) network. The LS-1 study is a multi-center, double-blind, phase III study of creatine in patients with early treated PD to assess whether

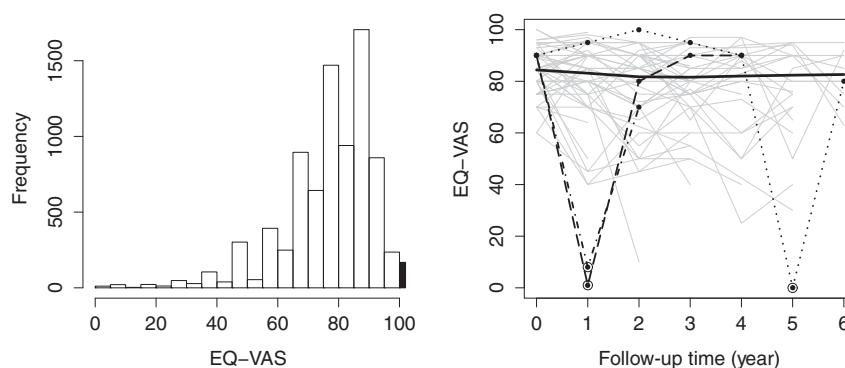


Figure 1 Histogram of the EQ-VAS scores (left panel) and the longitudinal profiles (right panel) of the EQ-VAS scores of 50 randomly selected subjects (gray lines) and three subjects with outlying observations (black dashed, dotdashed, and dotted lines) from the LS-1 study and the lowest smooth curve (black solid line).

creatine slows PD clinical decline defined by a combination of cognitive, physical, and quality of life measures. A total of 1741 patients with early PD were randomly assigned to receive either placebo or creatine. Participants were followed until the last enrolled participant has completed 5 years of observation. Thus, many participants had extended follow-up, to a maximum of 6 years. In-person evaluations were conducted at baseline and then annually beginning at 12 months. The LS-1 study represents the largest cohort of patients with early treated PD ever enrolled in a clinical trial (Elm, 2012; Kieburtz et al., 2015). The detailed description of the design of the LS-1 study can be found in Elm (2012).

The primary outcome of interest in this article is standardized generic instrument EuroQol vertical visual analog scale (EQ-VAS), which is widely used in PD related research (Kieburtz et al., 2015). EQ-VAS is a patient-reported outcome (PRO) that takes values between 100 (best imaginable health) and 0 (worst imaginable health), on which patients provide a global assessment of their health. Participants of LS-1 study are early PD patients with diagnosis within 2 years. Because PD is a slow progression disease, it is likely that early PD patients in the LS-1 study were still in a very good health condition and they considered themselves as in best imaginable health. Each patient contributed between 1 and 7 (mean = 4.73, SD = 1.70) EQ-VAS measurements. There are 24 intermittent missing values observed in EQ-VAS and during the follow-up, 78 and 323 individuals died and dropped out the study, respectively. The intermittent missing data and the missing data after loss of follow-up or death are assumed to be missing at random (MAR) in this article. Figure 1 (left panel) displays the histogram based on all EQ-VAS observations. Because there is only one patient reporting zero value for EQ-VAS, we add 0.1 to that observation. However, the presence of a substantial number of 100's (168 out of 8227 observations, 2.04%), if unaccounted for, is a potential issue for Beta or Beta rectangular regression models.

Figure 1 (right panel) displays the longitudinal profiles of the EQ-VAS scores of 50 randomly selected subjects. Because PD is a slow progression disease, it is unexpected to observe sudden value change in the outcome variables such as EQ-VAS, as indicated by the nearly horizontal lowest smooth curve (black solid line) (Cleveland, 1979). However, compared to baseline measurements, two patients (denoted by dashed and dot dashed lines) have a sudden value drop at their year 1 follow-up visit, and return to a higher level at their year 2 visit. Similarly, another patient's EQ-VAS score (denoted by dotted line) at the fifth year is significantly lower than the two adjacent years. Hence, these three observations are potential outliers. We divide the EQ-VAS variable by 100 to rescale it to the interval (0, 1]. We are interested in examining the effect of outliers, as well as the boundary values of 1, on the inference of regression models based on the Beta and Beta rectangular distributions.

3 Model and estimation

3.1 Beta distribution

In this section, we briefly review the reparameterization of Beta distribution (Ferrari and Cribari-Neto, 2004). A random variable Y follows a Beta distribution if the probability density function (pdf) in terms of its mean μ and precision parameter ϕ is given by,

$$f_B(Y = y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad (1)$$

where $0 < y < 1$, $0 < \mu < 1$, and $\phi > 0$. Then we have $E(Y) = \mu$ and $\text{Var}(Y) = \mu(1-\mu)/(1+\phi)$. We adopt the notation $Y \sim \text{Beta}(\mu, \phi)$.

The Beta regression model can be defined by linking the mean and covariates of interest under a GLM framework as $\text{logit}(\mu_i) = X_i\beta$, where X_i is the covariates vector of interest for subject i . The precision parameter ϕ can be either regressed on covariates after logarithm transformation or considered as a constant among different subjects. By specifying different values of μ and ϕ , the Beta distribution is flexible to allow different shapes and skewness.

3.2 Beta rectangular distribution

The Beta rectangular (BR) distribution is a mixture distribution consisting of a Beta distribution and a uniform (rectangular) distribution between 0 and 1. Its probability density function with support $(0, 1)$ is given by

$$f_{BR}(Y = y|\mu, \phi, q) = q + (1-q)f_B(y|\mu, \phi),$$

where $0 \leq q \leq 1$ is the mixture probability, and $f_B(y|\mu, \phi)$ is the pdf of the Beta distribution as in (1). We denote the BR distribution as $Y \sim \text{BR}(\mu, \phi, q)$. Obviously, if $q = 1$, the BR distribution reduces to the uniform (rectangular) distribution between 0 and 1 and if $q = 0$, it reduces to the Beta distribution $f_B(y|\mu, \phi)$. The mean and variance of the BR distribution are $E(Y) = (1-q)\mu + \frac{q}{2}$ and $\text{Var}(Y) = \frac{\mu(1-\mu)}{1+\phi}(1-q)[1+q(1+\phi)] + \frac{q}{12}(4-3q)$.

However, a regression analysis typically models the mean of the response (Ferrari and Cribari-Neto, 2004). To obtain a more appropriate regression structure for the mean of the BR distribution, Bayes et al. (2012) let $\gamma = \frac{q}{2} + (1-q)\mu$ and $\alpha = \frac{q}{1-(1-q)2\mu-1}$, whose parameter space is $\{0 \leq \gamma \leq 1, 0 \leq \alpha \leq 1\}$. Under this reparameterization, the pdf of the BR distribution is

$$f_{BR}(Y = y|\gamma, \phi, \alpha) = \alpha(1 - |2\gamma - 1|) + (1 - \alpha(1 - |2\gamma - 1|)) \times \\ \times f_B\left(y \left| \frac{\gamma - 0.5\alpha(1 - |2\gamma - 1|)}{1 - \alpha(1 - |2\gamma - 1|)}, \phi \right.\right). \quad (2)$$

We denote the reparameterized BR distribution as $Y \sim \text{BR}(\gamma, \phi, \alpha)$, where γ is the mean, ϕ is the precision parameter, and α is a shape parameter controlling the thickness of the tails. When the mixture probability $q = 1$, then $\alpha = 1$ and $\gamma = 0.5$, the BR distribution reduces to the uniform (rectangular) distribution between 0 and 1. When $q = 0$, then $\alpha = 0$ and $\gamma = \mu$, the BR distribution reduces to the Beta distribution. In general, when $0 < q < 1$, then $0 < \alpha < 1$, the BR distribution has heavier tails than its Beta distribution counterpart. To visualize this, Fig. 2 displays the density functions of various BR distributions with different values of γ , ϕ , and α . It suggests that when $\alpha > 0$, the BR distribution has a heavier tail than the corresponding Beta distribution.

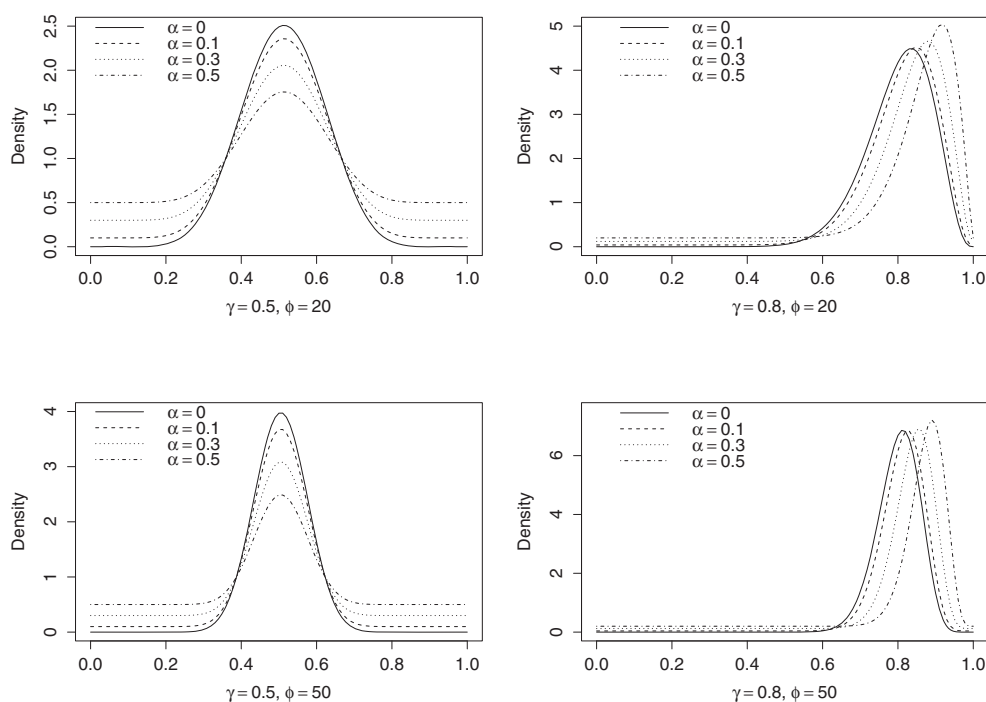


Figure 2 The density functions of various Beta rectangular distributions with different values of γ , ϕ , and α . $\alpha = 0$ (solid line), $\alpha = 0.1$ (dashed line), $\alpha = 0.3$ (dotted line), $\alpha = 0.5$ (dotdash line).

Similar to the Beta regression model, the Beta rectangular regression model can be defined as $\text{logit}(\gamma_i) = \mathbf{X}_i\boldsymbol{\beta}$ and the precision parameter ϕ can be either regressed on covariates or considered as a constant among different subjects.

3.3 One-augmented Beta rectangular random effects model

In this section, we generalize the Beta rectangular regression model to account for the longitudinal data structure and the boundary values zero and one. For the ease of illustration, we only consider the boundary value of one (rescaled from 100, as in Fig. 1 left panel). We illustrate how to extend the model to account for both zero and one in Web Section 2. Let y_{ij} be the observed outcome (e.g. EQ-VAS) at visit j ($j = 1, \dots, J_i$, where $j = 1$ is baseline and J_i is the number of visits for subject i) from subject i ($i = 1, \dots, I$, where I is the number of subjects). Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ_i})'$ be the outcome vector for subject i and let $\mathbf{y} = (y_1, \dots, y_I)'$ be the observed outcome matrix. We then propose a one-augmented BR (OABR) model, denoted by $Y \sim \text{OABR}(p_{1ij}, \gamma_{ij}, \phi_{ij}, \alpha)$, whose probability density function follows:

$$f(Y_{ij} = y_{ij} | p_{1ij}, \gamma_{ij}, \phi_{ij}, \alpha) = \begin{cases} p_{1ij} & \text{if } y_{ij} = 1 \\ (1 - p_{1ij})f_{BR}(Y_{ij} = y_{ij} | \gamma_{ij}, \phi_{ij}, \alpha) & \text{if } y_{ij} \in (0, 1), \end{cases} \quad (3)$$

where $p_{1ij} = P(Y_{ij} = 1)$, $f_{BR}(Y_{ij} = y_{ij} | \gamma_{ij}, \phi_{ij}, \alpha)$ is the reparameterized BR density function given in (2), $\gamma_{ij} = E[Y_{ij}]$, and ϕ_{ij} is the precision parameter of subject i at visit j . Next, we propose the OABR regression model by regressing the covariates onto p_{1ij} , γ_{ij} , and ϕ_{ij} , which are transformed by some

link functions:

$$\begin{aligned}\text{logit}[p_{1ij} = P(y_{ij} = 1|u_{i0})] &= \mathbf{X}_{i0}\boldsymbol{\omega} + u_{i0} \\ \text{logit}(\gamma_{ij}|u_{i1}) &= \mathbf{X}_{i1}\boldsymbol{\beta} + u_{i1} \\ \text{log}(\phi_{ij}|u_{i2}) &= \mathbf{X}_{i2}\boldsymbol{\eta} + u_{i2},\end{aligned}\tag{4}$$

where the covariate vectors \mathbf{X}_{i0} , \mathbf{X}_{i1} , and \mathbf{X}_{i2} can be identical or different and they include covariates of interest (e.g. treatment assignment) and potential confounding variables (e.g. subjects characteristics and socioeconomic status) from subject i . We adopt the logit link function for both p_{1ij} and γ_{ij} , while other link functions (e.g. probit and complementary log-log) can also be used. We assume that the random effects vector $\mathbf{u}_i = (u_{i0}, u_{i1}, u_{i2})'$ follows a multivariate normal distribution $N_3(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{Bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{Bmatrix}.\tag{5}$$

The proposed OABR regression model can be modified to accommodate various features in the data. For example, the one-inflated Beta regression model (BEOI) in the GAMLSS family can be obtained by replacing the BR density $f_{BR}(Y_{ij} = y_{ij}|\gamma_{ij}, \phi_{ij}, \alpha)$ with the Beta density in (1) or equivalently by setting $\alpha = 0$. If the outcome matrix y contains both zeros and ones, the OABR regression model can be generalized to the zero-one augmented BR (ZOABR) regression model by adding the probability $p_{0ij} = P(Y_{ij} = 0)$ to model (3) representing the probability of observing 0 and regressing covariates onto p_{0ij} transformed by some link function. Note that the ZOABR regression model requires an additional constraint: $0 < p_{0ij} + p_{1ij} < 1$ and we illustrate in details how to impose this constraint in Web Section 2. On the other hand, if there are no zeros or ones observed, we can let $p_{0ij} = p_{1ij} \equiv 0$, then the OABR regression model reduces to the mixed effects BR regression model. Moreover, we only include random intercepts in model (4) for simplicity, more random effects (e.g. random slope) can be easily included in the model.

Let the parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\eta}, \Sigma, \alpha)$. Conditional on the random effects \mathbf{u}_i , all measurements of each subject are assumed to be independent. We have the full likelihood of subject i as follows:

$$L(\boldsymbol{\Theta}, \mathbf{u}_i; \mathbf{y}_i) = \left[\prod_{j=1}^{J_i} p(y_{ij}|\mathbf{u}_i) \right] p(\mathbf{u}_i) = \left[\prod_{j=1}^{J_i} p_{1ij}^{I(y_{ij}=1)} \{(1 - p_{1ij})f_{BR}(y_{ij}|\gamma_{ij}, \phi_{ij}, \alpha)\}^{1-I(y_{ij}=1)} \right] p(\mathbf{u}_i),\tag{6}$$

where $I(\cdot)$ denotes the indicator function, $f_{BR}(y_{ij}|\gamma_{ij}, \phi_{ij}, \alpha)$ is the density function of the BR distribution, and $p(\mathbf{u}_i)$ is the density function of the random effects \mathbf{u}_i .

4 Bayesian inference

We adopt Bayesian methods based on Markov chain Monte Carlo (MCMC) algorithms to obtain statistical inference. The fully Bayesian inference has many advantages. First, MCMC algorithms can be used to estimate exact posterior distributions of the parameters, while likelihood-based estimation only produces a point estimate of the parameters, with asymptotic standard errors (Dunson, 2007). Second, Bayesian inference provides better performance in small samples compared to likelihood-based estimation (Lee and Song, 2004). In addition, it is more straightforward to deal with more complicated models using Bayesian inference via MCMC.

4.1 Prior specification

To make inference on the unknown parameter vector Θ , we use Bayesian inference based on MCMC posterior simulations. We use vague prior distributions on all elements in the parameter vector Θ . The prior distributions of all elements in ω , β , and η are $N(0, \tau^2)$. We use the prior distribution Inverse-Gamma (λ_1, λ_2) for σ 's to ensure positivity. Specifically, we let $\tau = 10$ and $\lambda_1 = \lambda_2 = 0.01$. The prior distribution for ρ 's is $\rho \sim \text{Uniform}(-1, 1)$. We have investigated other selections of vague prior distributions with various hyper-parameters (e.g. τ , λ 's) and obtained very similar results. Please refer to Web Section 3 for details of the sensitivity analysis results. The posterior samples are obtained from the full conditional of each unknown parameter using Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1994) and No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014). The HMC and NUTS samplers are implemented in Stan (Stan Development Team, 2014, version 2.6.0), which is a probabilistic programming language implementing statistical inference with HMC and NUTS samplers. The model fitting is performed in Stan by specifying the full likelihood function and the prior distributions of all unknown parameters. For large datasets, Stan may be more efficient than BUGS language in achieving faster convergence and requiring smaller number of samples. To facilitate easy reading and implementation of the proposed OABR regression model, the Stan codes have been posted in the Web Supplement.

To monitor Markov chain convergence, we use history plots and view the absence of apparent trend in the plots as evidence of convergence. We run multiple chains with diffuse initial values and ensure the scale reduction \hat{R} of all parameters are smaller than 1.1 (Gelman et al., 2013).

4.2 Bayesian model selection and influence diagnostics

There are a variety of model selection criteria in Bayesian inference. The widely used criteria conditional predictive ordinate (CPO) (Geisser, 1993; Dey et al., 1997; Sinha and Dey, 1997; Carlin and Louis, 2009; Ghosh and Hanson, 2010) is adopted to assess model fit and selection. The CPO for the (ij) th observation (observation j from subject i) is defined as

$$\text{CPO}_{ij} = p(y_{ij}|y_{(ij)}) = \int p(y_{ij}|\Theta)p(\Theta|y_{(ij)})d\Theta, \quad (7)$$

where y_{ij} denotes the full data and $y_{(ij)}$ denotes the data after deleting the (ij) th observation. CPO is a form of cross-validation with high value indicating that the data for observation (ij) can be accurately predicted by a model based on the data from all other observations. Hence, a model with larger CPO_{ij} for all observations suggests a better fit. Although the close form of CPO_{ij} is not available for our proposed model, a Monte Carlo estimator of CPO_{ij} can be obtained by MCMC samples $\{\Theta^{(t)}\}_{t=1}^M$ from posterior distribution $p(\Theta|y)$, with M being the total number of post burn-in samples. Because $p(y_{ij}|y_{(ij)}) = p(y)/p(y_{(ij)}) = 1/\int p(\Theta|y)/p(y_{ij}|y_{(ij)}, \Theta)d\Theta$, a harmonic-mean approximation of CPO_{ij} is $\widehat{\text{CPO}}_{ij} = \left(\frac{1}{M} \sum_{t=1}^M \frac{1}{p(y_{ij}|y_{(ij)}, \Theta^{(t)})}\right)^{-1} = \left(\frac{1}{M} \sum_{t=1}^M \frac{1}{p(y_{ij}|\Theta^{(t)})}\right)^{-1}$ (Dey et al., 1997). A summary statistics of $\widehat{\text{CPO}}_{ij}$ for all individuals is the log pseudo-marginal likelihood (LPML) defined as $\text{LPML} = \sum_{i=1}^I \sum_{j=1}^J \log(\widehat{\text{CPO}}_{ij})$. A larger value of LPML indicates better fit of the model. Moreover, we adopt a pseudo-Bayes factor for comparing two models defined as $\text{PBF}_{21} = \exp(\text{LPML}_2 - \text{LPML}_1)$ (Ghosh and Hanson, 2010).

To detect the occurrence of outliers and extremal events around the tail-area, we consider the Kullback–Leibler (K–L) divergence defined as $K\{p(\Theta|y), p(\Theta|y_{(ij)})\} = \int p(\Theta|y) \log\left(\frac{p(\Theta|y)}{p(\Theta|y_{(ij)})}\right)d\Theta$. Peng and Dey (1995) pointed out that $K\{p(\Theta|y), p(\Theta|y_{(ij)})\} = \log E_{\Theta|y}[\{p(y_{ij}|\Theta)\}^{-1}] + E_{\Theta|y}[\log\{p(y_{ij}|\Theta)\}] = -\log(\text{CPO}_{ij}) + E_{\Theta|y}[\log\{p(y_{ij}|\Theta)\}]$, where $E_{\Theta|y}(\cdot)$ denotes the expectation with

respect to the joint posterior distribution $p(\Theta|y)$. Cancho et al. (2011) proposed the Monte Carlo estimate of the K-L divergence as

$$K\{p(\Theta|y), \widehat{p(\Theta|y_{(ij)})}\} = -\log(\widehat{\text{CPO}}_{ij}) + \frac{1}{M} \sum_{t=1}^M \log\{p(y_{ij}|\Theta^{(t)})\}. \quad (8)$$

5 Simulation studies

In this section, we conduct an extensive simulation study with three settings to compare the performance of the one-inflated Beta (BEOI) regression model in the GAMLSS family and the proposed OABR regression model. In all three settings, we generate 200 datasets with sample size $N = 1200$ subjects and seven visits (baseline and six follow-up visits, $J_i = 7$) for each subject. The data structure is similar to the motivating LS-1 study and the continuous proportional outcome is restricted in the interval $(0, 1]$. We consider one covariate x_i taking value 0 or 1 each with probability 1/2 to mimic the treatment assignment. The time vector $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{i7})' = (0, 1, 2, 3, 4, 5, 6)'$, which is the same as the motivating LS-1 study.

5.1 Simulation Settings I and II: Data simulated from either the BEOI or the OABR regression model

The aims of simulation Settings I and II are to demonstrate that under model overparameterization, the OABR regression model can reduce to the BEOI regression model, while the BEOI regression model does not perform well when data are generated from the OABR regression model. In these two simulation settings, we assume that there are boundary value ones and generate the datasets from the models

$$\begin{aligned} \text{logit}[p_{1ij} = P(y_{ij} = 1|u_{i0})] &= \omega_0 + \omega_1 x_i + u_{i0} \\ \text{logit}(y_{ij}|u_{i1}) &= \beta_0 + \beta_1 x_i + \beta_2 t_{ij} + \beta_3 x_i t_{ij} + u_{i1} \\ \text{log}(\phi_{ij}|u_{i2}) &= \eta_0 + \eta_1 x_i + \eta_2 t_{ij} + \eta_3 x_i t_{ij} + u_{i2}, \end{aligned} \quad (9)$$

where the random effects $(u_{i0}, u_{i1}, u_{i2})' \sim N_3(\mathbf{0}, \Sigma)$, with $\sigma_1 = 1.2$, $\sigma_2 = 0.6$, $\sigma_3 = 0.4$, and $\rho_{12} = 0.4$, $\rho_{13} = 0.1$, $\rho_{23} = 0.4$ as the components of the covariance matrix Σ shown in (5). We set the regression coefficients $\omega = (\omega_0, \omega_1)' = (-1.5, -0.5)'$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (1.5, -0.5, -0.1, 0.2)'$, and $\eta = (\eta_0, \eta_1, \eta_2, \eta_3)' = (2.5, 0.2, 0.1, 0.1)'$. We set either $\alpha = 0$ (Setting I) or $\alpha = 0.2$ (Setting II), then the data are simulated from either the BEOI (Setting I) or the OABR (Setting II) regression models. We fit the BEOI and OABR regression models to the simulated datasets in both settings. Web Tables 1 and 2 display bias (the average of the posterior means minus the true values), standard deviation (SD, the standard deviation of the posterior means), coverage probabilities (CP) of 95% equal-tail credible intervals, and root mean squared error (RMSE) from the BEOI and OABR regression models in Settings I and II, respectively. The results suggest that when data are simulated from BEOI regression model as in Setting I, both the BEOI and OABR regression models generate comparable results with very small bias, RMSE close to SD, and the coverage probability being reasonably close to 0.95. Under model overparameterization, the estimate of the shape parameter α from the OABR regression model is correctly close to zero, suggesting that it is still a reasonable model in this simulation setting. When the data are simulated from the OABR regression model in Setting II ($\alpha = 0.2$), the OABR regression model can successfully recover all parameters, including α . In contrast, the BEOI regression model gives biased estimates and poor coverage probabilities for most parameters.

Table 1 Simulation results when data are simulated from the BEOI (Setting III) regression models and are contaminated by 1% outliers.

	BEOI regression model				OABR regression model			
	Bias	SD	SE	RMSE	Bias	SD	SE	RMSE
Setting III: Data simulated from the BEOI regression model with outliers								
$\omega_0 = -1.5$	-0.001	0.071	0.068	0.071	-0.007	0.070	0.068	0.070
$\omega_1 = -0.5$	0.005	0.102	0.095	0.102	0.005	0.099	0.096	0.098
$\beta_0 = 1.5$	-0.103	0.033	0.032	0.108	-0.063	0.034	0.032	0.072
$\beta_1 = -0.5$	0.015	0.044	0.043	0.047	0.019	0.045	0.043	0.049
$\beta_2 = -0.1$	0.007	0.006	0.005	0.009	0.004	0.006	0.005	0.007
$\beta_3 = 0.2$	-0.010	0.008	0.007	0.012	-0.008	0.008	0.007	0.011
$\eta_0 = 2.5$	-0.170	0.059	0.063	0.180	0.041	0.061	0.058	0.073
$\eta_1 = 0.2$	0.090	0.084	0.088	0.123	0.035	0.082	0.080	0.089
$\eta_2 = 0.1$	0.011	0.017	0.015	0.020	0.009	0.017	0.016	0.019
$\eta_3 = 0.1$	-0.058	0.024	0.021	0.063	-0.014	0.023	0.022	0.028
$\sigma_1 = 1.2$	0.000	0.057	0.055	0.057	0.003	0.055	0.055	0.055
$\sigma_2 = 0.6$	-0.043	0.016	0.015	0.045	-0.029	0.015	0.014	0.032
$\sigma_3 = 0.4$	0.420	0.017	0.027	0.420	-0.069	0.042	0.042	0.081
$\rho_{12} = 0.4$	-0.024	0.046	0.044	0.052	-0.006	0.042	0.042	0.042
$\rho_{13} = 0.1$	-0.133	0.065	0.060	0.148	-0.025	0.124	0.123	0.126
$\rho_{23} = 0.4$	-0.046	0.034	0.041	0.057	-0.058	0.079	0.078	0.098
$\alpha = 0$					0.074	0.003	0.007	0.074

Table 2 Model comparison statistics for the LS-1 dataset. LPML: log pseudo-marginal likelihood; PBF: pseudo-Bayes factor.

	LPML	PBF
BEOI	7494.85	Ref
OABR	7504.75	>> 100

5.2 Simulation Settings III: Data simulated from the BEOI regression model with outliers

The aims of simulation Setting III is to compare the performance of the BEOI and OABR regression models while data are simulated from the BEOI model but contaminated with outliers, in order to evaluate the influence of outliers and extremal events around the tail-area. Setting III are similar to Setting I, but we contaminate 1% of the randomly selected observations with high scores (between 0.9 and 1) by decreasing Δ units ($y_{ij}^* = y_{ij} - \Delta$, and $\Delta = 0.8$). To visualize the outliers, we randomly select one dataset in Setting III and plot it in Fig. 3. Figure 3 displays the longitudinal profiles (upper panels) of 50 randomly selected subjects and three contaminated subjects (denoted by black dashed, dotted, and dot dash lines), in addition to the boxplots of all subjects (lower panels) before (left panels) and after (right panels) the outlier contamination. Some observations from the three highlighted subjects are contaminated and have some potential outliers represented by the sudden value drops.

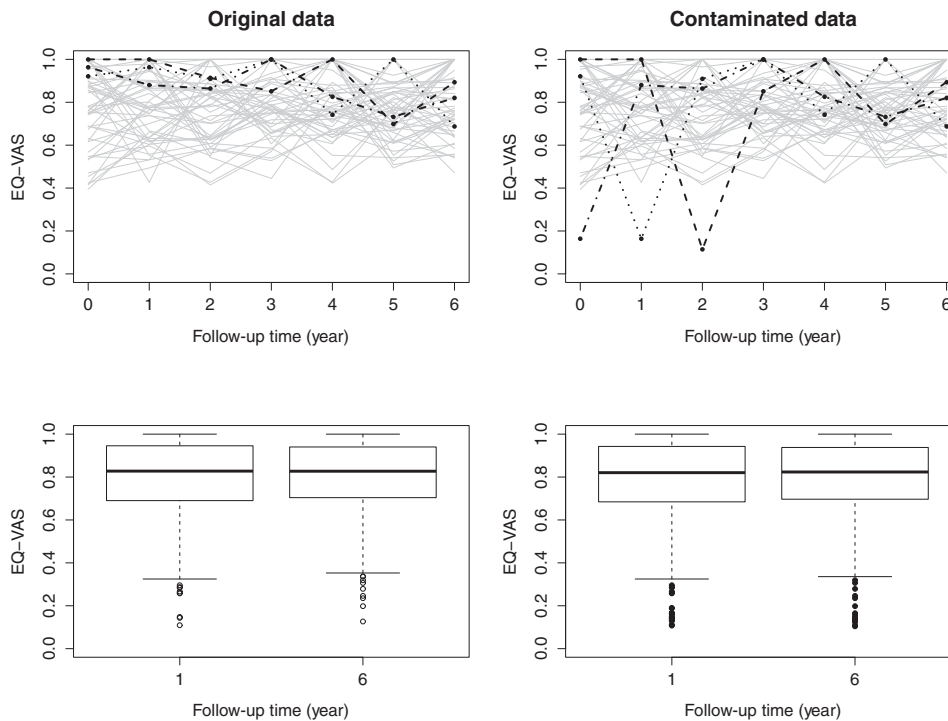


Figure 3 The longitudinal profiles (upper panels) of 50 randomly selected subjects and three contaminated subjects (black dashed, dotted, and dot dash lines), in addition to the boxplots of all subjects (lower panels) before (left panels) and after (right panels) the outlier contamination.

After contamination by outliers, the lower tails of the data become heavier and include more extremal values.

We then fit the BEOI and OABR regression models in Setting III. Table 1 displays the simulation results. In comparison to the BEOI regression model, the OABR regression model provides similar estimates for parameters β and ω , but much smaller bias (e.g. -0.170 vs. 0.041 for η_0 and 0.090 vs. 0.035 for η_1) and much smaller RMSE (e.g. 0.180 vs. 0.073 for η_0 and 0.123 vs. 0.089 for η_1) for parameters η , σ 's, and ρ 's. The estimates of the regression parameter vector ω from both models have small bias because the probability of being 1 is not contaminated. The estimate of the shape parameter α is 0.074 in the OABR regression model, assigning some probability to the occurrences of outliers and extremal events.

Figure 4 displays the K–L divergence measures of all observations from a randomly selected simulation dataset, estimated by the BEOI (left panel) and OABR (right panel) regression models. The BEOI regression model identifies many observations as potential outliers indicated by K–L divergence measures being larger than 3. However, using the OABR regression model, there seems to be no influential observations with all K–L divergence measures smaller than 2. This figure suggests that in comparison to the BEOI model, the OABR regression model can effectively control the potential outlying observations.

In conclusion, the simulation results suggest that when the data are simulated from the BEOI regression model, the overparameterized OABR regression model generates results comparable to the true BEOI regression model. However, when the data are simulated from the OABR regression model, the BEOI regression model provides parameter estimates with large bias and RMSE and poor coverage probabilities. When the data are simulated from the BEOI regression model, but are contaminated by

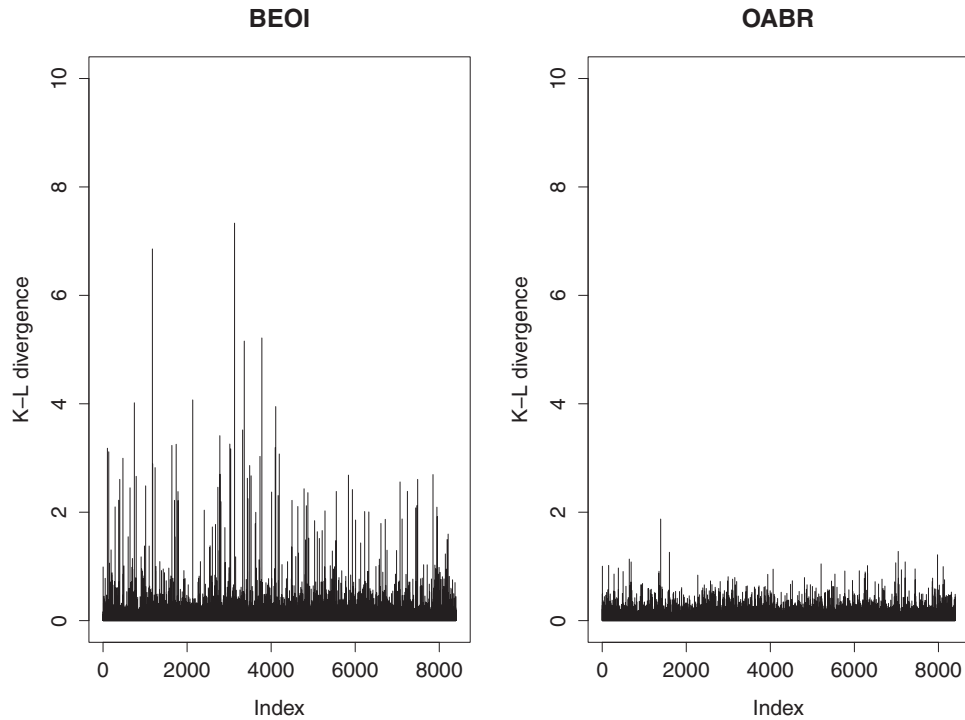


Figure 4 Estimated K–L divergence measures from the BEOI (left panel) and OABR (right panel) regression models for a randomly selected simulation dataset in Setting III.

some outliers, the OABR regression model provides parameter estimates with much smaller bias and RMSE, comparing to the BEOI regression model.

6 Application to the LS-1 study

In this section, we apply the BEOI and proposed OABR regression models and the Bayesian inference framework to the motivating LS-1 study. For all results in this section, we run two parallel MCMC chains with overdispersed initial values and run each chain for 2000 iterations. The first 1000 iterations are discarded as burn-in and the inference is based on the remaining 1000 iterations from each chain. Good mixing properties of the MCMC chains for all model parameters are observed in the trace plots. The scale reduction \widehat{R} of all parameters are smaller than 1.1.

In the data analysis, we consider the following covariates: the treatment assignment variable x_i (1 for treatment, and 0 for placebo), time t_{ij} , and time and treatment interaction. In model (4), we let $\mathbf{X}_{i0} = (\mathbf{1}, x_i)$, $\mathbf{X}_{i1} = (\mathbf{1}, x_i, t_i, x_i \times t_i)$, and $\mathbf{X}_{i2} = (\mathbf{1}, x_i, t_i, x_i \times t_i)$. We divide the EQ-VAS variable by 100 to rescale it to the interval (0, 1]. Table 2 compares the BEOI and OABR regression models using the model selection criteria discussed in Section 4.2. The OABR regression model performs better than the BEOI regression model with larger LPML. The PBF in favor of the OABR over the BEOI regression model is much greater than 100, indicating decisive evidence of choosing the OABR regression model as the final model.

To determine the presence of possible outlying observations, Fig. 5 displays the K–L divergence measures of all observations estimated from the BEOI (left panel) and OABR (right panel) regression models. The BEOI regression model identifies many observations as potential outliers indicated by

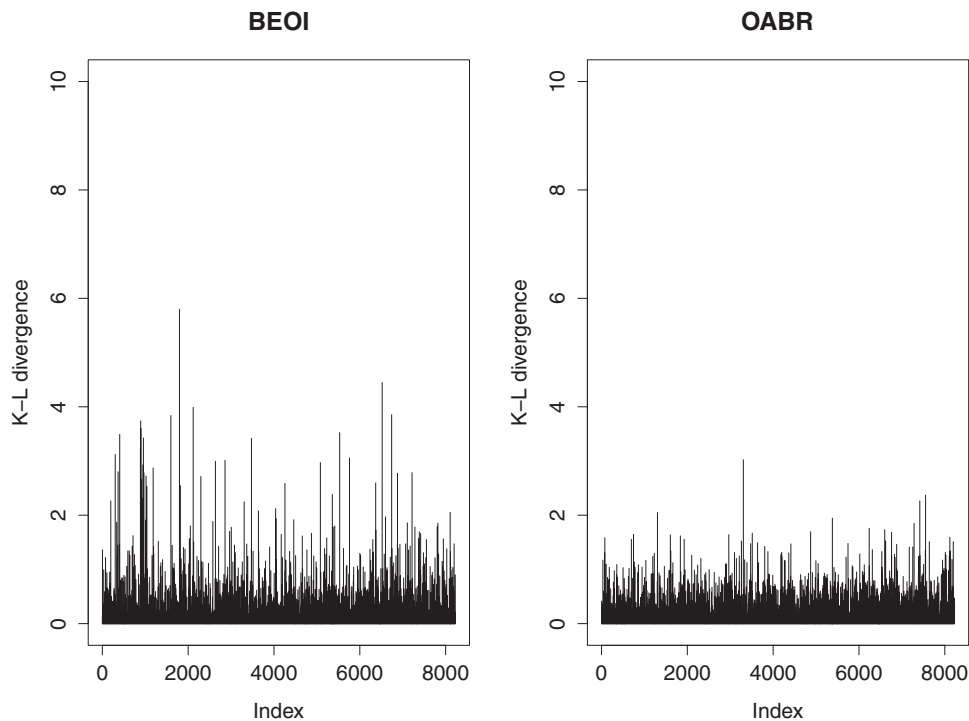


Figure 5 Estimated K–L divergence measures from the BEOI (left panel) and OABR (right panel) regression models for the LS-1 dataset.

the K–L divergence larger than 3. However, using the OABR regression model, there seems to be only one influential observation while all other observations have K–L divergence measures smaller than 3. This figure suggests that the OABR regression model can effectively control the potential outlying observations.

Table 3 displays the posterior mean, standard deviation (SD), and 95% equal-tail credible intervals from the BEOI and OABR regression models. Parameter estimates are noticeably different from two models. For creatine patients, the odds ratio of reporting 100 in EQ-VAS score is 0.604 ($\exp(-0.505)$, 95% CI: [0.348, 0.970]) comparing to the placebo patients, in the BEOI regression model, vs. 0.607 ($\exp(-0.500)$, 95% CI: [0.361, 1.065]) in the OABR regression model. The parameters in the second part of Table 3 represent the covariates effects on the mean EQ-VAS score conditional on not being one. Thus, negative parameters suggest deterioration in patients' global assessment of their health represented by EQ-VAS score. Conditional on other covariates and the random effects, parameter interpretation can be expressed in terms of the covariate effect on the odds $\frac{\gamma_{ij}}{1-\gamma_{ij}}$ (rescale) or $\frac{100\gamma_{ij}}{100-100\gamma_{ij}}$

in the original scale for the OABR regression model. Specifically, for creatine patients, the ratio between the expected EQ-VAS score γ_{ij} and the difference to perfect health $(1 - \gamma_{ij})$ is 0.941 ($\exp(-0.061)$, 95% CI: [0.872, 1.012]) times the ratio of placebo patients. For one year increase in time, the ratio between the expected EQ-VAS score and the difference to perfect health decreases by 7.5% ($1 - \exp(-0.078)$, 95% CI: [0.066, 0.085]). The regression parameters in the BEOI regression model can be interpreted in a similar way. The parameters in the third part of Table 3 represent the covariates effects on the precision parameter ϕ_{ij} . The results suggest that the precision parameter is not affected by treatment, but changes over time.

Table 3 Results of fitting the BEOI and OABR regression models in the LS-1 dataset.

	BEOI regression model				OABR regression model			
	Mean	SD	95% CI		Mean	SD	95% CI	
For probability of being one								
Int	-6.237	0.388	-7.053	-5.587	-6.367	0.369	-7.122	-5.687
Trt	-0.505	0.255	-1.056	-0.030	-0.500	0.273	-1.019	0.063
For conditional mean								
Int	1.573	0.036	1.508	1.635	1.599	0.026	1.547	1.649
Trt	-0.041	0.040	-0.121	0.024	-0.061	0.037	-0.137	0.012
Time	-0.076	0.005	-0.087	-0.066	-0.078	0.005	-0.089	-0.068
Time:Trt	0.001	0.008	-0.014	0.015	0.002	0.008	-0.013	0.017
For precision parameter								
Int	3.133	0.068	3.024	3.251	3.338	0.055	3.227	3.441
Trt	-0.018	0.077	-0.179	0.143	-0.075	0.076	-0.231	0.068
Time	-0.044	0.018	-0.081	-0.012	-0.057	0.019	-0.095	-0.019
Time:Trt	-0.004	0.023	-0.054	0.041	0.004	0.027	-0.049	0.056
σ_1	2.555	0.241	2.084	3.081	2.632	0.225	2.212	3.086
σ_2	0.670	0.013	0.646	0.697	0.654	0.014	0.627	0.681
σ_3	0.802	0.025	0.759	0.853	0.588	0.031	0.524	0.648
ρ_{12}	0.610	0.051	0.522	0.705	0.646	0.053	0.542	0.742
ρ_{13}	-0.118	0.092	-0.258	0.068	-0.041	0.096	-0.226	0.153
ρ_{23}	0.657	0.026	0.603	0.706	0.643	0.037	0.569	0.717
α					0.053	0.008	0.039	0.068

Note that the estimate of the correlation coefficient ρ_{12} between two random effects u_{i0} and u_{i1} is 0.646 (95% CI: [0.542, 0.742]). The significant positive correlation coefficient indicates that the patients with better global assessment of their health (higher u_{i1}) are more likely to report perfect health (higher u_{i0}). The facts that the estimate of the shape parameter α is 0.053 (95% CI: [0.039, 0.068]) and the PBF in favor of the OABR regression model over the BEOI regression model is much larger than 100 (as in Table 2) suggest the existence of potential outliers and heavier tails than what the Beta distributions can model.

7 Discussion

In this article, we propose the one-augmented Beta rectangular (OABR) regression model for responses measured in the interval $(0, 1]$, which is extended to data in the closed unit interval $[0, 1]$. This model accounts for not only the within-subject correlation and the occurrence of boundary values, but also the overoccurrence of tail-area events including heavy tails and outlying observations. We adopt Bayesian inference framework based on Markov Chain Monte Carlo (MCMC) simulation for parameter estimation. The extensive simulation study suggests that the proposed OABR regression model improves the accuracy of parameter estimates when outliers and heavy tails exist. In comparison, the regression model based on the Beta distribution (BEOI regression model in the GAMLSS family) provides parameter estimates with large bias and RMSE and poor coverage probabilities in the presence of heavy tails or outliers. We apply the proposed model to the motivating LS-1 study dataset. The OABR regression model has better fit than the BEOI regression model. The treatment creatine has

insignificant effects in either the probability of reporting perfect health (100 in the EQ-VAS score) or the mean EQ-VAS score. This finding is not surprising because the LS-1 study was terminated early for futility based on results of a planned interim analysis (Kiebert et al., 2015). However, the patient-reported global health assessment deteriorates along with time. The proposed model and Bayesian inference can be easily implemented by the publicly available software packages such as BUGS and Stan, and be easily accessible to by applied researchers.

There are some limitations in our proposed OABR regression model that we will address in our future study. First, the parameters of the augmented Beta rectangular distribution (e.g. p_{1ij} , γ_{ij} , and ϕ_{ij}) are modeled as linear functions of covariates. However, the linear assumption may not be realistic in some scenarios. As pointed out by one reviewer, the GAMLSS family allows nonlinear or smooth functions in the regression models and it can handle smooth effects of covariates. In our future research, we would like to investigate a class of varying-coefficient models (Sun and Wu, 2005) that incorporate the time-dependent covariate effects via penalized splines with a truncated polynomial basis and a fixed number of knots (Ruppert, 2002). Second, both the monotone (e.g. dropout) and nonmonotone (e.g. missed visit) missing exist in the LS-1 study. In this article, we assume they are all missing at random (MAR). However, monotone missing is likely to be caused by some terminal events such as dropout or death. The terminal events are often correlated with the health outcomes such as EQ-VAS and they create the issue of informative censoring. How to address the informative censoring issue in the proposed augmented Beta rectangular regression model is an interesting direction of future research. Moreover, when the nonmonotone missingness due to missed visit is associated with either the unobserved value or the underlying health status (e.g. sicker patients are more likely to miss visits), the nonmonotone missing data are missing not at random (MNAR) (Little and Rubin, 2002). Under the MNAR assumption, the missing data mechanism needs to be modeled simultaneously with the outcome variable to avoid biased parameter estimates (Diggle et al., 2002). In addition, we have chosen a multivariate normal distribution for the random effects vector because it is flexible in modeling the covariance structure within and between various types of recurrent events and it has meaningful interpretation on correlation. In generalized linear-mixed models, misspecification of random effects distribution has little impact on the parameters that are not associated with the random effects (Jacqmin-Gadda et al., 2007; Rizopoulos et al., 2008; McCulloch et al., 2011). The impact of random effects misspecification in the proposed modeling framework warrants further investigation. We will also investigate the effect of random effects misspecification and relax the normality assumption by considering Bayesian nonparametric (BNP) framework based on Dirichlet process mixture (Escobar, 1994).

Acknowledgments The project described was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant KL2 TR000370. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performing computing resources that have contributed to the research results reported within this article. URL: <http://www.tacc.utexas.edu>. The author thanks the editor, associate editor, and two anonymous referees for their reading and valuable comments.

Conflict of interest

The authors have declared no conflict of interest.

References

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

- Cancho, V. G., Dey, D. K., Lachos, V. H. and Andrade, M. G. (2011). Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: estimation and case influence diagnostics. *Computational Statistics and Data Analysis* **55**, 588–602.
- Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.
- Dey, D. K., Chen, M. H. and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**, 1239–1252.
- Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press.
- Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B* **195**, 216–222.
- Dunson, D. D. (2007). Bayesian methods for latent trait modelling of longitudinal data. *Statistical Methods in Medical Research* **16**, 399–415.
- Elm, J. J. (2012). Design innovations and baseline findings in a long-term Parkinson's trial: the national institute of neurological disorders and stroke exploratory trials in Parkinson's Disease Long-Term Study–1. *Movement Disorders* **27**, 1513–1521.
- Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.
- Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815.
- Figueroa-Zúniga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. P. (2013). Mixed beta regression: a Bayesian perspective. *Computational Statistics and Data Analysis* **61**, 137–147.
- 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.
- 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 and Applications* **20**, 463–486.
- Geisser, S. (1993). *Predictive Inference*. Vol. **55**. CRC Press, Boca Raton, FL.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*. London, UK: CRC press, Boca Raton, FL.
- Ghosh, P. and Hanson, T. (2010). A semiparametric Bayesian approach to multivariate longitudinal data. *Australian and New Zealand Journal of Statistics* **52**, 275–288.
- Hahn, E. D. (2008). Mixture densities for project management activity times: a robust approach to PERT. *European Journal of Operational Research* **188**, 450–459.
- Hatfield, L. A., Boye, M. E. and Carlin, B. P. (2011). Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of Biopharmaceutical Statistics* **21**, 971–991.
- Hatfield, L. A., Boye, M. E., Hackshaw, M. D. and Carlin, B. P. (2012). Multilevel Bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *Journal of the American Statistical Association* **107**, 875–885.
- Hoffman, 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.
- Jacqmin-Gadda, Hélène, Sibillot, Solenne, Proust, Cécile, Molina, Jean-Michel and Thiébaud, Rodolphe. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis* **51**, 5142–5154.
- Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. **2**. John Wiley & Sons, Hoboken, New Jersey.
- Kiebertz, K., Tilley, B. C., Elm, J. J., et al. (2015). Effect of creatine monohydrate on clinical progression in patients with parkinson disease: a randomized clinical trial. *JAMA* **313**, 584–593.
- Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling* **3**, 193–213.
- Lee, Sik-Yum and Song, Xin-Yuan. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research* **39**, 653–686.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. John Wiley & Sons, Hoboken, New York.

- McCulloch, C. E., Neuhaus, J. M., et al. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science* **26**, 388–402.
- Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics* **111**, 194–203.
- Ospina, R. and Ferrari, S. L. P. (2010). Inflated Beta distributions. *Statistical Papers* **51**, 111–126.
- Peng, F. and Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics* **23**, 199–213.
- Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C* **54**, 507–554.
- Rizopoulos, D., Verbeke, G. and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95**, 63–74.
- Rogers, J. A., Polhamus, D., Gillespie, W. R., Ito, K., Romero, K., Qiu, R., Stephenson, D., Gastonguay, M. R. and Corrigan, B. (2012). Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *Journal of Pharmacokinetics and Pharmacodynamics* **39**, 479–498.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.
- Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195–1212.
- Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54.
- Stan Development Team. (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.6.0*.
- Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* **23**, 1–46.
- Sun, Y. and Wu, H. (2005). Semiparametric time-varying coefficients regression model for longitudinal data. *Scandinavian Journal of Statistics* **32**, 21–47.
- Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the Beta distribution. *Journal of Educational and Behavioral Statistics* **37**, 82–113.
- Vieira, A. M. C., Hinde, J. P. and Demétrio, C. G. B. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics* **27**, 373–389.
- Zhao, L., Chen, Y. and Schaffner, D. W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. *Applied and Environmental Microbiology* **67**, 2129–2135.