Computational Statistics and Data Analysis 167 (2022) 107345

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Bayesian beta regression for bounded responses with unknown supports

Haiming Zhou [a,*], Xianzheng Huang [b]

[a] Department of Statistics and Actuarial Science, Northern Illinois University, DeKalb, IL 60115, USA
[b] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

ARTICLE INFO

Article history:
Received 19 January 2021
Received in revised form 29 August 2021
Accepted 30 August 2021
Available online 10 September 2021

Keywords:
Four-parameter beta distribution
g-Prior
Mean
Mode
Model criterion

ABSTRACT

A new Bayesian regression framework is presented for the analysis of continuous response data with support restricted to an unknown finite interval. A four-parameter beta distribution is assumed for the response conditioning on covariates, with the mean or mode depending linearly on covariates through a known link function. An informative g-prior is proposed to incorporate the prior distribution for the marginal mean or mode of the response. Byproducts of the Markov chain Monte Carlo sampling for implementing the proposed method lead to model criteria useful for model selection. Goodness-of-fit of the model is assessed using Cox-Snell residual plots. The methodology is illustrated in simulations and demonstrated in two real-life data applications. An R package, `betaBayes`, is developed for easy implementation of the proposed regression methodology.

1. Introduction

Researchers in a wide range of fields encounter bounded data in their studies. For example, environmental scientists monitor the proportion of hygienic waste in residential solid waste. Asset allocations in a portfolio and the share of household income spent on food are bounded data of interest in economics. Psychologists analyze confidence ratings and bounded scores from cognitive tests administered to study subjects. Examples of bounded data in the biomedical field include prevalence rates and death rates of the coronavirus disease 2019 (COVID-19), and body fat percentages of athletes. Different from unbounded data, central tendency measures, skewness, and other features of the underlying distribution for bounded data are inextricable from the support of the distribution. Consequently, more caution is necessary when drawing inference for these features based on bounded data, especially when the support is unknown.

Existing approaches for analyzing bounded data typically assume a prefixed support such as $(0, 1)$, sometimes after scaling the raw data. The beta mean regression model proposed by Ferrari and Cribari-Neto (2004) probably has received the most attention for modeling response data bounded on the unit interval, where the mean parameter of the beta distribution depends linearly on covariates through a known link function. Model diagnostic methods for a beta mean regression were considered in Espinheira et al. (2008a,b), Ferrari et al. (2011), and Rocha and Simas (2011). The model has also been extended to allow the precision parameter to vary with covariates (Smithson and Verkuilen, 2006; Ferrari et al., 2011). An R package `betareg` (Cribari-Neto and Zeileis, 2010; Grün et al., 2012) is available on CRAN for fitting the beta mean regres-

* Corresponding author.
E-mail address: zhouh@niu.edu (H. Zhou).

https://doi.org/10.1016/j.csda.2021.107345
0167-9473/

sion model with varying precision and performing model diagnostics. This R package also allows for fitting a finite mixture of beta regression models (Verkuilen and Smithson, 2012). Time series analysis of bounded data via a beta mean regression is presented in Guolo et al. (2014), which incorporates a serial dependence between responses via a Gaussian copula. All the aforementioned works carry out frequentist inference, mostly based on maximum likelihood. Bayesian treatments for modeling the response data bounded on $(0, 1)$ include the Bayesian beta mean regression model (Branscum et al., 2007), a beta rectangular regression model based on a mixture of a beta distribution and a uniform distribution (Bayes et al., 2012), a mixed effects beta model (Figueroa-Zúñiga et al., 2013), and a flexible beta model based on a special mixture of two beta distributions (Migliorati et al., 2018). Unlike all the above regression models which focus on inferring the conditional mean of a bounded response, Bayes et al. (2017) developed quantile regression models for bounded responses built upon beta distributions. Barrientos et al. (2017) proposed a fully nonparametric Bayesian approach to model the covariates-dependent distribution of a bounded response. Recently, Zhou et al. (2020) considered a beta mode regression model where the mode of the response is related to covariates through a link function.

All existing works mentioned above assume that the response variable is bounded on a prefixed interval such as (0, 1), which may not be appropriate. For example, a human being's body fat percentage can never reach a value close to zero or one. Google results show that the lowest body fat percentage is 2% in a human being; although the highest body fat percentage is not available, it is probably much less than one. In cases like this, misspecifying the support can degrade inference for a central tendency measure of the response conditioning on covariates, for instance. In some applications, inferring the support is the focal point of interest. For example, an accurate prediction for the support of the prevalence rate of COVID-19, that is more refined than the unit interval, in an upcoming flu season is important to local health officials. Other examples where the support of a response is unknown yet is of practical interest include models for survival analysis to study the minimum possible life time (Smith, 1994), the job-search problem (Flinn and Heckman, 1982; Christensen and Kiefer, 1991), and the procurement-auction problem (Paarsch, 1992; Donald and Paarsch, 2002). In these and many other existing works on regression models with the support of the response depending on unknown parameters, the authors established some unusual, often unappealing, properties of maximum likelihood estimators for the support parameters and other model parameters (e.g., Donald and Paarsch, 1993; Smith, 1994). These theoretical findings motivated alternative estimators for parameters in these nonregular regression models, many of which were proposed in the Bayesian paradigm.

To allow for inference on the support along with other features of the response, we consider in this study the four-parameter beta distribution, which extends the beta distribution by introducing two parameters to define the support, in addition to the two shape parameters. As noted above, statisticians have long recognized that estimating the support creates a non-regular problem, where the maximum likelihood estimation may fail to yield consistent estimators (Smith, 1985; Cheng and Traylor, 1995). Existing methods for estimating the four-parameter beta distribution include the moment-based estimation (Johnson et al., 1995; McGarvey et al., 2002), the maximum likelihood estimation when both shape parameters are greater than two (Carnahan, 1989), the corrected maximum likelihood method when both shape parameters are greater than one (Cheng and Iles, 1987), and the penalized likelihood approach (Wang, 2005), among others. The penalized likelihood approach by Wang (2005) is applicable without restricting the shape parameters to be above one or two, but standard error estimation for estimators of the four parameters are not provided.

These existing works on four-parameter beta distributions are not in a regression context. In fact, we can find little research on the four-parameter beta distribution in a regression setting. In this article, we present a class of Bayesian regression models that permit an inference for the support boundaries by considering the four-parameter beta distribution supported on (θ_1, θ_2), and introducing either a mean or mode parameter that linearly depends on covariates through a known link function. To facilitate Bayesian inference, we adopt an informative g-prior on the regression coefficients that leads to more efficient posterior sampling, especially when the data provide relatively weak information on the conditional mode or when multicollinearity is present. With a careful choice of blocking, we develop a fully automated (no manual "tuning" is required) Markov chain Monte Carlo (MCMC) algorithm for the posterior sampling. A new variation of the Cox-Snell residual plot (Cox and Snell, 1968) is provided for gross assessment of the model fit. Furthermore, all methods developed in the paper can be easily implemented in a freely-available R package, `betaBayes`, calling complied C++. The ready availability of software allows researchers to empirically compare various competing beta regression models on their own data with a continuous bounded response.

The remaining of the article is organized as follows. Section 2 describes the four-parameter beta regression models, including prior development and posterior inference. We consider in Section 3 model selection criteria and model diagnostics. Section 4 presents simulations to illustrate the quality of inference results when comparing to relevant existing methods. Section 5 comprises two illustrative data analyses with software implementation. The paper is concluded in Section 6 where we summarize the contributions of our study and discuss future research directions.

2. Model formulation and inference

2.1. The regression models

Consider observed data consisting of n independent realizations of the response-covariates pair, $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$, where y_i is the response supported on an unknown interval (θ_1, θ_2), and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ is a vector of covari-

ates with the intercept. For a random variable Y that follows a four-parameter beta distribution, $Y \sim \text{beta4}(\alpha_1, \alpha_2, \theta_1, \theta_2)$ in short, its probability density function (pdf) is given by

$$f_{\text{beta}}(y; \alpha_1, \alpha_2, \theta_1, \theta_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{(y-\theta_1)^{\alpha_1-1}(\theta_2-y)^{\alpha_2-1}}{(\theta_2-\theta_1)^{\alpha_1+\alpha_2-1}}, \text{ for } y \in (\theta_1, \theta_2), \tag{1}$$

where $\Gamma(t)$ is the gamma function, $\alpha_1 > 0$ and $\alpha_2 > 0$ are two shape parameters, and θ_1 and θ_2 are two unknown support parameters. One can show that $W = (Y-\theta_1)/(\theta_2-\theta_1)$ follows a beta distribution with shape parameters α_1 and α_2. Let μ_w and μ_y denote the mean for W and Y, respectively. It is easy to show that $\mu_w = \alpha_1/(\alpha_1+\alpha_2)$ and $\mu_y = \mu_w(\theta_2-\theta_1)+\theta_1$. In preparation for formulating of a mean regression model, we set $\alpha_1 = \phi m$ and $\alpha_2 = \phi(1-m)$, for $0 < m < 1$ and $\phi > 0$, which leads to $\mu_w = m$, and that ϕ plays the role of a precision parameter, of which a larger value implies a lower variance of the beta or beta4 distribution. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ denote a vector of regression coefficients with β_0 being the intercept. We propose a Bayesian beta4 mean regression model specified by the following submodels,

$$\begin{aligned} y_i|m_i, \phi, \theta_1, \theta_2 &\sim \text{beta4}\left(\phi m_i,\ \phi(1-m_i), \theta_1, \theta_2\right), \\ m_i &\equiv m(\mathbf{x}_i) = h^{-1}(\boldsymbol{\beta}'\mathbf{x}_i), \\ p(\boldsymbol{\beta}, \phi, \theta_1, \theta_2) &= p(\boldsymbol{\beta})p(\phi)p(\theta_1)p(\theta_2), \end{aligned} \tag{2}$$

where $h(\cdot)$ is a link function such as logit, probit, and log-log, and each $p(\cdot)$ represents a prior density. The prior specification is discussed in Section 2.2. Under this beta4 mean model, we have

$$h\{(\text{Mean}[y_i|\mathbf{x}_i] - \theta_1)/(\theta_2-\theta_1)\} = \boldsymbol{\beta}'\mathbf{x}_i,$$

the left-hand side of which can be interpreted as a quantile-score for the position of $\text{Mean}[y_i|\mathbf{x}_i]$ within (θ_1, θ_2). A higher quantile-score leads to a higher mean response. Therefore, the interpretation of the regression coefficient β_j is that, for every one unit increase in x_{ij}, the quantile-score for the mean response increases by β_j units, for $j = 1, \ldots, p$.

With the mode instead of mean as the central tendency measure of interest, we revise the above mean regression model to construct a mode regression model. It is easy to show that, when $\alpha_1, \alpha_2 > 1$, there is a unique mode for W and Y given by $m_w \equiv \text{Mode}(W) = (\alpha_1-1)/(\alpha_1+\alpha_2-2)$ and $m_y \equiv \text{Mode}(Y) = m_w(\theta_2-\theta_1)+\theta_1$, respectively. Focusing on a unimodal beta or beta4 distribution, we set $\alpha_1 = 1+\phi m$ and $\alpha_2 = 1+\phi(1-m)$, for $0 < m < 1$ and $\phi > 0$, which leads to $m_w = m$ and that ϕ again can be interpreted as a precision parameter. Mimicking the beta4 mean regression model in (2), we propose the Bayesian beta4 mode regression model via the following hierarchical models,

$$\begin{aligned} y_i|m_i, \phi, \theta_1, \theta_2 &\sim \text{beta4}\left(1+\phi m_i,\ 1+\phi(1-m_i), \theta_1, \theta_2\right), \\ m_i &\equiv m(\mathbf{x}_i) = h^{-1}(\boldsymbol{\beta}'\mathbf{x}_i), \\ p(\boldsymbol{\beta}, \phi, \theta_1, \theta_2) &= p(\boldsymbol{\beta})p(\phi)p(\theta_1)p(\theta_2). \end{aligned} \tag{3}$$

Under this beta4 mode regression model, we have

$$h\{(\text{Mode}[y_i|\mathbf{x}_i] - \theta_1)/(\theta_2-\theta_1)\} = \boldsymbol{\beta}'\mathbf{x}_i,$$

with the interpretation of regression coefficients similar to that under the beta4 mean regression model.

2.2. Prior specification

In what follows, we specify prior distributions for the support parameters, the precision parameter, and the regression coefficients successively.

Prior for θ_1 and θ_2

Suppose one has some natural prior information that the support of the response is within the interval $(a_{\theta_1}, b_{\theta_2})$, where $a_{\theta_1} < \theta_1$ and $b_{\theta_2} > \theta_2$ are known. For instance, the body fat percentage of a human being is naturally bounded within $(a_{\theta_1}, b_{\theta_2}) = (0, 1)$, but its true support should be strictly within and much narrower than $(0, 1)$. Given such prior information and data $\mathcal{D}$, we consider the following priors on θ_1 and θ_2,

$$\theta_1 \sim \text{unif}(a_{\theta_1}, b_{\theta_1}),\ \theta_2 \sim \text{unif}(a_{\theta_2}, b_{\theta_2}), \tag{4}$$

where $\text{unif}(a, b)$ refers to the uniform distribution on (a, b), $b_{\theta_1} < y_{(1)}$, and $a_{\theta_2} > y_{(n)}$, in which $y_{(k)}$ refers to the kth order statistic of $\{y_1, \ldots, y_n\}$. When the prior information on bounds is not available, we consider the following default choices: $a_{\theta_1} = y_{(1)} - \Delta$, $b_{\theta_1} = y_{(1)} - 10^{-15}$, $a_{\theta_2} = y_{(n)} + 10^{-15}$, $b_{\theta_2} = y_{(n)} + \Delta$, where $\Delta > 10^{-15}$ is used to control the prior precision with larger Δ values indicating more vague priors. The choice of $\Delta = 2s_y$ has been shown in our simulation studies to perform well, where s_y is the sample standard deviation of y_i's.

Prior for ϕ

The precision parameter $\phi > 0$ controls the variability of y_i given $\mathbf{x}_i$ with a larger value implying lower variability. In most applications where bounded data are of interest, observed data typically provide enough information to infer ϕ. Hence, we consider a commonly used gamma prior, $\Gamma(a_\phi, b_\phi)$, on ϕ with $a_\phi = b_\phi = 0.001$ as the defaults, where a_ϕ is the shape parameter and b_ϕ is the rate parameter.

Informative g-prior on $\boldsymbol{\beta}$

A common choice for the prior on $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ in beta-related mean regression models has been the usual flat normal prior $N_{p+1}(\mathbf{0}, 10^5\mathbf{I}_{p+1})$ (Bayes et al., 2012; Migliorati et al., 2018). However, for beta4 mode regression, we find that the flat prior does not always yield reasonable posterior results when data provide relatively weak information on the conditional mode, e.g., when ϕ is small, or α_1 and α_2 are both less than one leading to nonexistence of mode. In these cases, the likelihood function often approaches a constant that is free of data, leading to non-uniqueness of the maximum likelihood estimate for $\boldsymbol{\beta}$. Similar phenomenon can emerge from linear models with a strong multicollinearity, for which a well-received strategy is to include a penalty term in the likelihood function within the frequentist framework, or to impose an informative prior on $\boldsymbol{\beta}$ in the Bayesian paradigm. As for the beta4 mean mode, although a flat prior for $\boldsymbol{\beta}$ works well in most cases, an informative prior is also preferable in the presence of multicollinearity. We next propose an informative prior on $\boldsymbol{\beta}$ in both beta4 mean and mode models along the same vein of this Bayesian strategy.

Denote by $m \in (0, 1)$ the central tendency measure of $(y_i - \theta_1)/(\theta_2 - \theta_1)$ in a proposed regression model, which is the mean in the beta4 mean model, and it is the mode in the beta4 mode model. Consider the situation where a subject-matter expert has information on the marginal distribution of m, which can be well-characterized by beta(a_m, b_m), where $a_m > 0$ and $b_m > 0$ are known. Here we use $m \sim$ beta(a_m, b_m) because the beta prior is quite flexible for modeling $(0, 1)$-supported parameter and brings mathematical convenience in the prior development. Our goal is to formulate a prior on $\boldsymbol{\beta}$ that takes advantage of this prior information while adjusting for covariates. For this purpose, we consider the following g-prior (Zellner, 1986),

$$\boldsymbol{\beta} \sim N_{p+1}\left(b\mathbf{e}_1, gn(\mathbf{X}'\mathbf{X})^{-1}\right), \tag{5}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$ is of length $p + 1$, b is a prior mean for the intercept, and $g > 0$ is a scaling constant. Suppose covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ arise independently from a population $H(\cdot)$ with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}$. With $\mathbf{x}_i$ including the intercept in the first element, the first element of $\boldsymbol{\mu}$ is one, and entries in the first row and those in the first column of $\boldsymbol{\Sigma}$ are all zeros. For any new subject with covariates $\mathbf{x} \sim H$ and response y, we have the mean or mode of $(y - \theta_1)/(\theta_2 - \theta_1)$ equal to $m(\mathbf{x}) = h^{-1}(\boldsymbol{\beta}'\mathbf{x})$. Given the data $\mathbf{X}$, assuming $\mathbf{x}$ and $\boldsymbol{\beta}$ are mutually independent, one has $E(\boldsymbol{\beta}'\mathbf{x}) = E_{\mathbf{x}}\{E_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x}|\mathbf{x})\} = E_{\mathbf{x}}(b\mathbf{e}_1'\mathbf{x}) = E_{\mathbf{x}}(b) = b$, by the law of iterated expectations. In addition, by the law of total variance, one has

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\beta}'\mathbf{x}) &= E_{\mathbf{x}}\{\mathrm{Var}_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x}|\mathbf{x})\} + \mathrm{Var}_{\mathbf{x}}\{E_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x}|\mathbf{x})\} \\
&= E_{\mathbf{x}}\{gn\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\} + \mathrm{Var}_{\mathbf{x}}(b) \\
&= g \cdot \mathrm{trace}\left\{n(\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')\right\} \\
&\xrightarrow{p} g \cdot \mathrm{trace}\left\{(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')^{-1}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')\right\} = g(p+1),
\end{aligned}$$

where $\xrightarrow{p}$ denotes "converge in probability," and the limiting statement originates from the fact that $n(\mathbf{X}'\mathbf{X})^{-1} \xrightarrow{p} (\boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Sigma})^{-1}$ (Vershynin, 2012). Hence, given $\mathbf{X}$, the g-prior in (5) implies that $\boldsymbol{\beta}'\mathbf{x}$ has a variance approximately equal to $g(p + 1)$ for any covariate $\mathbf{x}$ randomly drawn from its population $H(\cdot)$. Hanson et al. (2014) found that $\boldsymbol{\beta}'\mathbf{x}$ also often approximately follows a normal distribution, and this approximation is very good for a variety of H considered in their simulations, even when some covariates are categorical. Therefore, it is reasonable to assume that $\boldsymbol{\beta}'\mathbf{x}$ approximately follows $N(b, g(p + 1))$.

Motivated by the above findings, we choose values of b and g in the g-prior (5) so that the induced distribution of $m(\mathbf{x}_i) = h^{-1}(\boldsymbol{\beta}'\mathbf{x}_i)$ matches the marginal prior distribution $m \sim$ beta(a_m, b_m). More specifically, we minimize the Kullback-Leibler divergence from the distribution of $m(\mathbf{x}) = h^{-1}(\boldsymbol{\beta}'\mathbf{x})$ to beta(a_m, b_m), yielding $b = E\{h(m)\}$ and $g = \mathrm{Var}\{h(m)\}/(p+1)$ for $m \sim$ beta(a_m, b_m). When $h(\cdot)$ is the logit link, explicit expressions of the above mean and variance can be obtained, leading to $b = \delta(a_m) - \delta(b_m)$ and $g = \{\delta'(a_m) + \delta'(b_m)\}/(p + 1)$, where $\delta(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. For other link functions, we use approximations $b \approx \widehat{E\{h(m)\}}$ and $g = \widehat{\mathrm{Var}\{h(m)\}}/(p+1)$, where $\widehat{E\{h(m)\}}$ and $\widehat{\mathrm{Var}\{h(m)\}}$ are the sample mean and variance of a random sample from $m \sim$ beta(a_m, b_m).

When the values for a_m and b_m are not available, we use $a_m = b_m = 1$ as the defaults, yielding relatively weak prior information on the location of m.

2.3. Block MCMC

Given data $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$, the likelihood function of parameters $\Omega = (\boldsymbol{\beta}, \phi, \theta_1, \theta_2)$ associated with model (2) or (3) is

$$L(\mathcal{D}|\Omega) = \prod_{i=1}^{n} \frac{\Gamma(\alpha_{i1}+\alpha_{i2})(y_i-\theta_1)^{\alpha_{i1}-1}(\theta_2-y_i)^{\alpha_{i2}-1}}{\Gamma(\alpha_{i1})\Gamma(\alpha_{i2})(\theta_2-\theta_1)^{\alpha_{i1}+\alpha_{i2}-1}}, \tag{6}$$

where $\alpha_{i1} = \phi m_i$ and $\alpha_{i2} = \phi(1-m_i)$ for the beta4 mean mode in (2), and $\alpha_{i1} = 1+\phi m_i$ and $\alpha_{i2} = 1+\phi(1-m_i)$ for the beta4 mode model in (3). The posterior density is

$$\begin{aligned} p(\boldsymbol{\beta}, \phi, \theta_1, \theta_2|\mathcal{D}) \propto\ & L(\mathcal{D}|\Omega) \\ & \times \exp\left\{-\frac{1}{2gn}(\boldsymbol{\beta}-b\mathbf{e}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-b\mathbf{e}_1)\right\} \\ & \times I(a_{\theta_1} < \theta_1 < b_{\theta_1}) I(a_{\theta_2} < \theta_2 < b_{\theta_2}) \\ & \times \phi^{a_\phi - 1}\exp(-b_\phi\phi). \end{aligned}$$

Posterior sampling is carried out through adaptive Metropolis samplers (Haario et al., 2001). As commented earlier, central tendency measures of bounded data typically entangle with the support of the underlying distribution. For example, according to the beta4 mean model in (2), the conditional mean of y_i given $\mathbf{x}_i$, $\text{Mean}[y_i|\mathbf{x}_i] = h^{-1}(\boldsymbol{\beta}'\mathbf{x}_i)(\theta_2-\theta_1)+\theta_1$, depends on the support (θ_1, θ_2). Consequently, the posterior distribution of $\boldsymbol{\beta}$ and $(\theta_1, \theta_2)'$ are often highly correlated. We thus update them in a single block to effectively eliminate problematic MCMC mixing. It has been well documented in the literature that block sampling can improve MCMC efficiency relative to updating each parameter independently (Liu et al., 1994; Roberts and Sahu, 1997; Sargent et al., 2000). The algorithm for posterior sampling that incorporates block sampling is described next, where the $d = p+3$ dimensional vector $\boldsymbol{\xi} = (\boldsymbol{\beta}', z_1, z_2)'$ is introduced, with $z_i = \log\{(\theta_i - a_{\theta_i})/(b_{\theta_i} - \theta_i)\}$, $i = 1, 2$.

Step 1: Update $\boldsymbol{\xi}$.
Because z_i follows a standard logistic distribution, the full conditional distribution for $\boldsymbol{\xi}$ is

$$p(\boldsymbol{\xi}|\text{else}) \propto L(\mathcal{D}|\Omega)\exp\left\{-\frac{1}{2gn}(\boldsymbol{\beta}-b\mathbf{e}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-b\mathbf{e}_1)\right\}\frac{e^{z_1+z_2}}{(1+e^{z_1})^2(1+e^{z_2})^2},$$

where "else" denotes all other model parameters and the data. The vector $\boldsymbol{\xi}$ is updated using adaptive Metropolis samplers (Haario et al., 2001). More specifically, suppose we have sampled the states $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(l-1)}$, now in iteration l, we generate $\boldsymbol{\xi}^*$ from $N_d(\boldsymbol{\xi}^{(l-1)}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{(l)})$ and accept it with probability

$$\min\left\{1, \frac{p(\boldsymbol{\xi}^*|\text{else})}{p(\boldsymbol{\xi}^{(l-1)}|\text{else})}\right\},$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{(l)} = \begin{cases} \boldsymbol{\Sigma}_{0\boldsymbol{\xi}}, & l \le l_0, \\ \dfrac{2.4^2}{d}(\mathcal{C}_l + 10^{-10}\mathbf{I}_d), & l > l_0, \end{cases}$$

in which l_0 is the length of an initial period (e.g., $l_0 = 1000$), $\mathcal{C}_l$ is the sample variance of $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(l-1)}$, $\boldsymbol{\Sigma}_{0\boldsymbol{\xi}} = \text{diag}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}, \pi^2/3, \pi^2/3)$ is an initial diagonal covariance matrix of $\boldsymbol{\xi}$, $\pi^2/3$ is the variance of a standard logistic distribution, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ is the covariance estimate for $\boldsymbol{\beta}$ when using the R package `betareg` to fit a beta mean regression model $\{(y_i - \hat{\theta}_{10})/(\hat{\theta}_{20} - \hat{\theta}_{10})\} \sim \text{beta}(\phi h^{-1}(\mathbf{x}_i'\boldsymbol{\beta}), \phi(1-h^{-1}(\mathbf{x}_i'\boldsymbol{\beta})))$, with $\hat{\theta}_{10} = y_{(1)} - s_y/\sqrt{n}$ and $\hat{\theta}_{20} = y_{(n)} + s_y/\sqrt{n}$ being the ad hoc estimates of θ_1 and θ_2. These ad hoc estimates of (θ_1, θ_2) are borrowed from Turnbull and Ghosh (2014) for modeling bounded data with unknown support. The current choices for these initial estimates work well in our extensive simulation studies, although other choices can be used as well and usually have little impact on posterior inferences (as long as they are not too small or large).

Step 2: Update ϕ.
The full conditional distribution for $\log\phi$ is

$$p(\log\phi|\text{else}) \propto L(\mathcal{D}|\Omega)\phi^{a_\phi}\exp(-b_\phi\phi).$$

The logarithm of the precision parameter $\log\phi$ is updated via adaptive Metropolis samplers with normal proposal $\log\phi^* \sim N_1(\log\phi^{(l-1)}, \boldsymbol{\Sigma}_\phi^{(l)})$, where $\boldsymbol{\Sigma}_\phi^{(l)}$ is defined similarly as $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{(l)}$ above but with $\boldsymbol{\Sigma}_{0\boldsymbol{\xi}}$ replaced by $\boldsymbol{\Sigma}_{0\phi} = 1/n$, and the acceptance probability is

$$\min\left\{1, \frac{p(\log\phi^*|\text{else})}{p(\log\phi^{(l-1)}|\text{else})}\right\}.$$

To determine the running length of an MCMC run, one may first run a short chain without thinning, then use R package `coda` (Plummer et al., 2006) for convergence diagnostics and effective sample size calculations. Specifically, we use `raftery.diag` (Raftery and Lewis, 1992) to determine the burn-in period, the thinning interval, and the total number of iterations, use `heidel.diag` (Heidelberger and Welch, 1983) to ascertain stationarity, and use `effectiveSize` to calculate the effective sample size for each parameter. The mixing of the chains is evaluated through trace and auto-correlation plots. These MCMC diagnostics are demonstrated in the sample R code provided in supplementary Appendix A.

As pointed out by a referee, besides the block sampling algorithm implemented here, the Hamiltonian Monte Carlo (HMC) sampling algorithm (Duane et al., 1987) is another well-accepted MCMC algorithm designed to reduce the correlation between successive sampled states. Although algorithmic performance comparison is not the focus of our study, we compare these two algorithms in the context of a real-life data application in Section 5.1.

3. Model comparison and diagnostics

To compare different regression models that one may fit to the same data set, we adopt three model criteria described next, all of which are readily computed from the MCMC output. This computational convenience partly motivates our choice among many existing and well-accepted model criteria that may be used here (Mills and Prasad, 1992; Claeskens, 2016). To set the notations, denote by $\mathcal{D}_i$ the ith data point, and by $\mathcal{D}_{-i}$ the data set with $\mathcal{D}_i$ removed, for $i = 1, \ldots, n$. Let $L_i(\cdot|\Omega)$ be the likelihood contribution based on $\mathcal{D}_i$.

The first model criterion is the deviance information criterion (DIC, Spiegelhalter et al., 2002), which is a generalization of the Akaike information criterion (AIC, Akaike, 1998), and commonly used for comparing complex hierarchical models for which the asymptotic justification of AIC is not appropriate. The DIC is defined as

$$\mathrm{DIC} = -2\log L(\mathcal{D}|\hat{\Omega}) + 2p_D,$$

where

$$p_D = 2\left\{\log L(\mathcal{D}|\hat{\Omega}) - \frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\log L(\mathcal{D}|\Omega^{(l)})\right\}$$

is referred to as the effective number of parameters measuring the model complexity. Similar to AIC, a smaller value of DIC indicates a better fit of the model.

The second model criterion is the Watanabe-Akaike information criterion (WAIC, Watanabe, 2010) that has gained popularity in recent years due to its stability compared to DIC (Gelman et al., 2014; Vehtari and Gelman, 2014). The WAIC is defined as

$$\mathrm{WAIC} = -2\sum_{i=1}^{n}\log\left\{\frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}L_i(\mathcal{D}_i|\Omega^{(l)})\right\} + 2p_W,$$

where

$$p_W = \sum_{i=1}^{n}\left[\frac{1}{\mathcal{L}-1}\sum_{l=1}^{\mathcal{L}}\left\{\log L_i(\mathcal{D}_i|\Omega^{(l)}) - \frac{1}{\mathcal{L}}\sum_{k=1}^{\mathcal{L}}\log L_i(\mathcal{D}_i|\Omega^{(k)})\right\}^2\right]$$

is the effective number of parameters. A smaller value of WAIC indicates a better fit of the model.

The third model criterion is the log pseudo marginal likelihood (LPML, Geisser and Eddy, 1979). The definition of LPML is based on the conditional predictive ordinate (CPO) statistic, which is defined by, for data point $\mathcal{D}_i$,

$$\mathrm{CPO}_i = f(\mathcal{D}_i|\mathcal{D}_{-i}) = \int L_i(\mathcal{D}_i|\Omega)p_{\mathrm{post}}(\Omega|\mathcal{D}_{-i})d\Omega,$$

where $p_{\mathrm{post}}(\cdot|\mathcal{D}_{-i})$ is the posterior density of Ω give $\mathcal{D}_{-i}$. The data points with relative low CPO values indicate that they are not well fitted by the model. Therefore, it can be used to detect potential outliers given the model (Congdon, 2005). As noted by Gelfand and Dey (1994), one can use importance sampling to estimate CPO_i by

$$\left\{\frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\frac{1}{L_i(\mathcal{D}_i|\Omega^{(l)})}\right\}^{-1}.$$

However, these estimates may be unstable since the weights $\omega_{i,l} = 1/L_i(\mathcal{D}_i|\Omega^{(l)})$ can have infinite variance (Epifani et al., 2008), depending on the tail behavior of $p_{\mathrm{post}}(\Omega|\mathcal{D}_{-i})$ relative to $L_i(\mathcal{D}_i|\Omega)$ as a function of Ω. To stabilize the weights, Vehtari and Gelman (2014) suggest replacing $\omega_{i,l}$ with $\tilde{\omega}_{i,l} = \min\{\omega_{i,l}, \sqrt{\mathcal{L}}\bar{\omega}_i\}$, where $\bar{\omega}_i = \sum_{l=1}^{\mathcal{L}}\omega_{i,l}/\mathcal{L}$, leading to the stabilized CPO statistic given by

$$\widehat{\text{CPO}}_i = \frac{\sum_{l=1}^{\mathcal{L}} L_i(\mathcal{D}_i|\Omega^{(l)})\tilde{\omega}_{i,l}}{\sum_{l=1}^{\mathcal{L}} \tilde{\omega}_{i,l}}.$$

Finally, the LPML is defined as

$$\text{LPML} = \sum_{i=1}^{n} \log \widehat{\text{CPO}}_i.$$

A larger value of LPML suggests a better fit. Among the three model criteria, DIC and WAIC place emphasis on the relative quality of model fitting, while LPML focuses on the predictive performance of a model.

Another important issue to consider in any parametric regression analysis is model diagnostics. To address this issue, we employ the Cox-Snell plots for a general residual (Cox and Snell, 1968) defined by $r_i(\Omega) = -\log\{1 - F_{\mathbf{x}_i}(y_i; \Omega)\}$, for $i = 1, \ldots, n$, where $F_{\mathbf{x}_i}(y_i; \Omega)$ is the cumulative distribution function of y_i given $\mathbf{x}_i$. Although Cox-Snell residuals had primarily used in survival models in the literature, they were originally proposed as a general definition of residuals for various regression models such as linear models with non-normal errors and Poisson models (Cox and Snell, 1968). Alternatively, the deviance residuals and standardized ordinary residuals used for the $(0, 1)$-supported beta mean regression model in Ferrari and Cribari-Neto (2004) can also be defined for the beta4 models. We choose Cox-Snell residuals mainly because they allow us to assess uncertainty based on a posterior sample of residuals $r_i(\Omega)$ as described below.

By the probability integral transform, when evaluated at the cumulative distribution function with the true Ω that characterizes the data generating process, $r_i(\Omega)$ follows a standard exponential distribution. Therefore, if the model is "correct," the residuals r_i's are expected to behave like a random sample from the standard exponential distribution, and thus the curve for $\Lambda(t|\Omega) = -\log\{1 - n^{-1}\sum_{i=1}^{n} I(r_i(\Omega) \le t)\}$ versus t should be approximately straight with a slope equal to one. As a function of t, $\Lambda(t|\Omega)$ is unknown but can be estimated by $\Lambda(t|\hat{\Omega})$, where $\hat{\Omega}$ is an estimate of Ω. We use $\hat{\Omega} = \sum_{l=1}^{\mathcal{L}} \Omega^{(l)}/\mathcal{L}$ in our study, where $\{\Omega^{(1)}, \ldots, \Omega^{(\mathcal{L})}\}$ (e.g. $\mathcal{L} = 5,000$) are random draws from posterior distribution $[\Omega|\mathcal{D}]$ at the convergence of the MCMC algorithm in Section 2.3. Instead of plotting the single curve of $\Lambda(t|\hat{\Omega})$ as in a typical Cox-Snell plot, our version of the plot incorporates c curves in $\{\Lambda(t|\Omega^{(l)}), l = 1, \ldots, c < \mathcal{L}\}$, along with the equal-tailed 95% point-wise credible intervals based on the whole posterior sample $\{\Lambda(t|\Omega^{(1)}), \ldots, \Lambda(t|\Omega^{(\mathcal{L})})\}$, contrasting with a 45° reference line, where the equal-tailed credible interval is chosen so that the percentage of posterior sample below the interval is the same as the percentage above it. We set $c = 30$, which is large enough to allow visual assessment of the uncertainty due to estimating Ω, but not too large such that the plot becomes overly crowded. The 95% credible intervals that severely deviate from the 45° line provide evidence of an overall lack-of-fit of the model. As evidenced in the simulation study in Section 4.3, Cox-Snell residuals are sometimes conservative in that they may almost lie on a straight line when the assumed model departures from the true model as described in Baltazar-Aban and Pena (1995) and O'Quigley and Xu (2005).

4. Simulation studies

We design three simulation experiments to illustrate the implementation of the proposed regression methodology, to demonstrate the performance of the posterior inference, and to evaluate the effectiveness of model selection via DIC, WAIC, and LPML, and the graphical diagnostic method.

4.1. Simulation I: parameter estimation

The first simulation study aims to inspect posterior inference for Ω in the beta4 mean and mode models from the MCMC algorithm described in Section 2.3, implemented using the R package `betaBayes` available on CRAN.

To generate a random sample from a regression model, we first simulate $x_{i1} \overset{iid}{\sim} N(0, 1)$ independent of $x_{i2} \overset{iid}{\sim}$ Bernoulli(0.5) to create covariates $\mathbf{x}_i = (1, x_{i1}, x_{i2})'$, for $i = 1, \ldots, n$, where "iid" stands for "independent and identically distributed." Given a sample of covariates, we follow the beta4 mean model in (2) or the beta4 mode model in (3) to generate responses supported on $(\theta_1, \theta_2) = (0, 2)$, with regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (1, 1, 1)'$, and the link function $h(t) = \log\{t/(1 - t)\}$. In addition, we consider two samples sizes, $n = 100$, 500, and two conditional precision parameters, $\phi = 10$, 50, yielding four cases in total under each regression model. Under the current model settings, response data near the lower bound of the support are much more scarce than those near the upper bound, as one can see from scatter plots under each regression model with $\phi = 10$ and $\phi = 50$ in supplementary Figure S1. Later in this subsection, we consider a smaller sample size at $n = 67$, along with other model settings, such as that for the covariates distribution, to create scenarios similar to a data set considered in a real-life application presented in Section 5.

Under each simulation setting, 300 Monte Carlo (MC) replicate data sets are generated. We fit the regression models using the default priors introduced in Section 2.2. For each MCMC run, 2000 scans are thinned from 10000 after a burn-in period of 5000 iterations, which we confirm by convergence diagnostics to be more than adequate. Table 1 summarizes results regarding estimation of $\boldsymbol{\beta}$, ϕ, θ_1 and θ_2, including the MC averages of the posterior mean point estimate and the posterior standard deviation (PSD) of each point estimate, the standard deviation (across 300 MC replicates) of the point estimate, and the coverage probability of the 95% credible interval. All credible intervals are equal-tailed in this article. When the underlying beta distribution is less variable (with ϕ set at 50 as opposed to 10), with a moderate sample size at

Table 1
Results from Simulation I. These include averaged posterior mean point estimate (Est) and posterior standard deviation (PSD) of each point estimate, the standard deviation (across 300 MC replicates) of the point estimate (SD-Est), and the coverage probability (CP) of the 95% credible interval under the beta4 *mean* model (in the upper half of the table) and those under the beta4 *mode* model (in the lower half of the table).

Parameter	Est	PSD	SD-Est	CP	Est	PSD	SD-Est	CP
	For the beta4 *mean* model							
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.965	0.179	0.164	0.987	1.014	0.086	0.084	0.950
$\beta_1 = 1$	0.999	0.093	0.087	0.960	0.997	0.041	0.040	0.943
$\beta_2 = 1$	1.008	0.155	0.148	0.967	0.996	0.068	0.069	0.937
$\phi = 10$	10.056	2.070	1.980	0.957	10.104	0.953	0.946	0.940
$\theta_1 = 0$	0.001	0.185	0.180	0.950	−0.024	0.094	0.089	0.953
$\theta_2 = 2$	2.000	0.001	0.001	0.950	2.000	0.000	0.000	0.953
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.943	0.121	0.147	0.880	1.005	0.066	0.067	0.947
$\beta_1 = 1$	1.013	0.070	0.071	0.937	0.999	0.029	0.029	0.953
$\beta_2 = 1$	1.022	0.093	0.090	0.957	0.998	0.039	0.039	0.943
$\phi = 50$	50.469	10.260	11.624	0.897	50.353	4.466	4.386	0.963
$\theta_1 = 0$	0.037	0.142	0.174	0.860	−0.014	0.083	0.082	0.953
$\theta_2 = 2$	2.002	0.011	0.010	0.940	2.000	0.002	0.002	0.950
	For the beta4 *mode* model							
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.914	0.214	0.227	0.927	1.002	0.110	0.113	0.940
$\beta_1 = 1$	1.071	0.202	0.195	0.957	1.008	0.083	0.084	0.953
$\beta_2 = 1$	1.070	0.267	0.248	0.970	1.008	0.107	0.113	0.923
$\phi = 10$	9.521	2.706	2.815	0.927	10.084	1.476	1.346	0.967
$\theta_1 = 0$	0.071	0.172	0.184	0.933	−0.022	0.123	0.118	0.947
$\theta_2 = 2$	2.005	0.026	0.020	0.947	2.001	0.006	0.006	0.947
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.919	0.126	0.151	0.847	1.008	0.069	0.070	0.950
$\beta_1 = 1$	1.041	0.095	0.095	0.947	0.995	0.042	0.041	0.940
$\beta_2 = 1$	1.043	0.116	0.117	0.953	0.992	0.050	0.051	0.933
$\phi = 50$	48.282	10.603	11.355	0.920	51.557	5.538	5.636	0.933
$\theta_1 = 0$	0.084	0.141	0.174	0.823	−0.023	0.092	0.092	0.943
$\theta_2 = 2$	2.000	0.017	0.017	0.947	2.002	0.006	0.006	0.947

$n = 100$, we see some under-coverages for estimating β_0, ϕ, and θ_1 under both regression models. This can be explained by the scarceness of data information to infer the lower bound of the support when the sample is not large enough while the underlying distribution is more concentrated around the mode. The deficiency in such data information results in high uncertainty in inferring θ_1, the uncertainty that is underestimated by the PSD, and thus leads to low coverage probabilities of the credible interval for θ_1. Due to the inextricable links between the support and location/variability measures of a distribution, inferences for β_0 and ϕ are somewhat compromised in this case as domino effects. Fortunately, the ripple effects have little impact on the covariate effects estimation. When the sample size is larger, say, $n = 500$, or when the underlying distribution is less concentrated around the mode (by setting ϕ at 10 in place of 50), point estimates for all parameters are greatly improved and are much closer to the truth under each regression model, with PSDs closer to the corresponding empirical standard deviations, yielding credible intervals with coverage probabilities matching the nominal value more closely.

We carry out additional simulation experiments under settings that more closely mimic those in the real-life data applications in Section 5. In particular, we generate random samples, each of size $n = 67$, from the beta4 mean model with $\phi = 50$ and from the beta4 mode model with $\phi = 5$. Each model contains $p = 5$ covariates, whose values are simulated from a multivariate normal distribution $N_p(\mathbf{0}, 0.16\mathbf{R_x})$, where $\mathbf{R_x}$ has diagonal entries being 1 and off-diagonal entries being 0.5. Like seen under earlier simulation settings, our inference procedure produces satisfactory point estimation, standard deviation estimation, and interval estimation under these additional settings. Summary statistics of these estimates are in supplementary Table S1.

Existing beta regression models that are highly relevant to our proposed regression models include the beta mean model considered in Ferrari and Cribari-Neto (2004) and the beta mode model developed by Zhou et al. (2020). Our proposed regression models in (2) and (3) differ from theirs in that, first, support parameters are inferred along with other model parameters instead of assumed known, and second, our models are formulated for Bayesian regression analysis whereas theirs are designed for frequentist approaches. To investigate impacts on inference of using a pre-fixed support within the Bayesian framework, we carried out Bayesian regression analysis based on the aforementioned existing beta mean model and beta mode model, but with the boundary parameters (θ_1, θ_2) fixed at $(y_{(1)} - s_y/\sqrt{n},\ y_{(n)} + s_y/\sqrt{n})$ following the

Table 2
Results from Simulation I. These include averaged posterior mean point estimate (Est) and posterior standard deviation (PSD) of each point estimate, the standard deviation (across 300 MC replicates) of the point estimate (SD-Est), and the coverage probability (CP) of the 95% credible interval under the beta *mean* model (in the upper half of the table) and the beta *mode* model (in the lower half of the table) when boundaries (θ_1, θ_2) are prefixed at $(y_{(1)} - s_y/\sqrt{n},\ y_{(n)} + s_y/\sqrt{n})$ for each MC replicate data set (and thus PSD and CP associated with boundary parameters are not available).

Parameter	Est	PSD	SD-Est	CP	Est	PSD	SD-Est	CP
	For the beta4 *mean* model							
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.657	0.096	0.162	0.170	0.834	0.042	0.084	0.147
$\beta_1 = 1$	0.927	0.082	0.081	0.830	0.925	0.036	0.037	0.410
$\beta_2 = 1$	0.898	0.144	0.137	0.900	0.928	0.063	0.066	0.780
$\phi = 10$	10.127	1.463	2.043	0.833	10.606	0.682	1.009	0.740
$\theta_1 = 0$	0.238	NA	0.154	NA	0.105	NA	0.076	NA
$\theta_2 = 2$	2.041	NA	0.003	NA	2.019	NA	0.001	NA
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.543	0.055	0.198	0.013	0.756	0.022	0.091	0.000
$\beta_1 = 1$	1.075	0.050	0.074	0.647	1.037	0.020	0.036	0.537
$\beta_2 = 1$	1.055	0.085	0.100	0.870	1.029	0.035	0.044	0.807
$\phi = 50$	35.019	5.054	8.817	0.300	42.586	2.701	4.800	0.380
$\theta_1 = 0$	0.386	NA	0.169	NA	0.226	NA	0.090	NA
$\theta_2 = 2$	2.027	NA	0.008	NA	2.015	NA	0.002	NA
	For the beta4 *mode* model							
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.593	0.157	0.250	0.390	0.808	0.061	0.107	0.273
$\beta_1 = 1$	1.221	0.169	0.146	0.723	1.093	0.061	0.082	0.597
$\beta_2 = 1$	1.212	0.259	0.283	0.840	1.076	0.097	0.111	0.857
$\phi = 10$	6.955	1.227	1.648	0.370	8.337	0.627	1.031	0.360
$\theta_1 = 0$	0.322	NA	0.150	NA	0.175	NA	0.084	NA
$\theta_2 = 2$	2.014	NA	0.017	NA	2.008	NA	0.006	NA
	$n = 100$				$n = 500$			
$\beta_0 = 1$	0.566	0.066	0.195	0.040	0.763	0.025	0.097	0.013
$\beta_1 = 1$	1.208	0.065	0.109	0.153	1.103	0.024	0.053	0.127
$\beta_2 = 1$	1.199	0.105	0.129	0.527	1.090	0.040	0.058	0.420
$\phi = 50$	30.236	4.610	7.665	0.147	38.482	2.542	5.035	0.140
$\theta_1 = 0$	0.405	NA	0.150	NA	0.250	NA	0.092	NA
$\theta_2 = 2$	2.008	NA	0.016	NA	2.005	NA	0.006	NA

suggestion in Turnbull and Ghosh (2014). Table 2 presents results from this comparative experiment, which clearly suggest that all model parameters are poorly estimated if one does not carefully estimate the unknown support (θ_1, θ_2) along with other model parameters.

4.2. Simulation II: model selection

The second simulation experiment is designed to evaluate the performance of the three model criteria defined in Section 3. Here, data are generated from each of the two proposed beta regression models, in conjunction with three link functions, the logit, probit, and log-log link. These are also the six candidate models from which a model criterion chooses the "best" model based on a simulated data set. Configurations for covariates and regression coefficients are the same as those described in the second paragraph in Section 4.1, except for that we now focus on the setting with $\phi = 50$ and $n = 500$. For each of 300 MC replicate data sets, despite the true model used to generate the data, we fit all six candidate models using the default priors. Table 3 presents the average across 300 MC replicates for each of the three model criteria evaluated at a candidate model when data are generated from each of the six true models.

According to the summarized results in Table 3, all three considered model criteria tend to choose the true model as the best or the second best model based on a data set generated from the true model. In addition to this pattern, three other observations are worth pointing out. First, when evaluated at data generated from the beta4 mean model with the probit link, the considered model criteria can often choose the beta4 mean model with the logit link. This is not surprising considering the high similarity between the two link functions. Second, when a candidate model disagrees with the true model only in the link function, the candidate model that assumes the logit link tends to yield a better fit according to the model criteria. Based on this observation, we recommend using a logit link in the proposed beta4 regression models in practice, unless a model criterion strongly supports a different link. A third interesting observation is that, when data are generated from the beta4 mean model with the probit link, the model criteria suggest substantially worse fit of the beta4

Table 3
Results from Simulation II. Averages of each of the three model criteria across 300 Monte Carlo replicates evaluated at a candidate model based on data from different models. Numbers in parentheses are Monte Carlo standard errors associated with the averages. The true models are listed as the row titles: the beta4 mode model with the logit link (mo-logit), the probit link (mo-probit), and the log-log link (mo-loglog), and the beta4 mean model with the logit link (me-logit), the probit link (me-probit), and the log-log link (me-loglog). These are also the six candidate models listed as the column titles. Entries in bold in each row indicate that their corresponding candidate models outperform all other candidate models.

	mo-logit	mo-probit	mo-loglog	me-logit	me-probit	me-loglog
	For negative DIC					
mo-logit	**890** (2.0)	887 (2.0)	881 (2.0)	887 (2.0)	871 (2.1)	888 (2.1)
mo-probit	1293 (2.5)	**1310** (2.5)	1224 (2.6)	1225 (2.7)	1127 (2.8)	1247 (2.8)
mo-loglog	876 (2.0)	856 (2.0)	**895** (2.0)	861 (2.1)	823 (2.2)	896 (2.2)
me-logit	934 (2.1)	941 (2.1)	913 (2.0)	**950** (2.1)	947 (2.3)	938 (2.2)
me-probit	1516 (3.2)	1618 (3.5)	1372 (3.0)	**3958** (19.4)	**3962** (19.3)	3816 (18.8)
me-loglog	924 (2.1)	915 (2.1)	936 (2.1)	930 (2.2)	906 (2.3)	**981** (3.0)
	For negative WAIC					
mo-logit	**890** (2.0)	887 (2.0)	881 (2.0)	886 (2.0)	869 (2.1)	883 (2.2)
mo-probit	1293 (2.5)	**1310** (2.5)	1223 (2.6)	1217 (2.8)	1108 (3.0)	1180 (4.0)
mo-loglog	877 (2.0)	856 (2.0)	**895** (2.0)	859 (2.0)	817 (2.2)	872 (2.9)
me-logit	934 (2.1)	941 (2.1)	913 (2.0)	**948** (2.1)	934 (2.3)	928 (2.3)
me-probit	1515 (3.2)	1616 (3.5)	1369 (3.0)	**3828** (18.2)	**3827** (18.1)	3643 (18.9)
me-loglog	925 (2.0)	915 (2.1)	936 (2.1)	926 (2.1)	882 (2.6)	**946** (3.4)
	For LPML					
mo-logit	**445** (1.0)	443 (1.0)	440 (1.0)	443 (1.0)	435 (1.0)	442 (1.0)
mo-probit	646 (1.3)	**655** (1.3)	612 (1.3)	610 (1.3)	558 (1.4)	611 (1.3)
mo-loglog	438 (1.0)	428 (1.0)	**447** (1.0)	430 (1.0)	410 (1.0)	443 (1.0)
me-logit	467 (1.0)	471 (1.0)	456 (1.0)	**474** (1.1)	470 (1.1)	466 (1.0)
me-probit	757 (1.6)	808 (1.8)	685 (1.5)	**1914** (9.1)	**1914** (9.0)	1836 (9.0)
me-loglog	462 (1.0)	458 (1.0)	468 (1.0)	463 (1.1)	447 (1.1)	**482** (1.4)

mode model. A closer inspection on such data reveals that, under the current parameter settings in a beta4 mean model, the distribution of the response is of J-shaped, a feature that cannot be captured by a unimodal beta distribution one assumes in the beta4 mode regression model. This explains the lack of fit reflected in the model criteria. These observations in turn suggest the effectiveness of the model criteria in identifying the best model based on observed data.

The third observation above brings up the important issue of drawing statistical inference based on a misspecified model, e.g., fitting a beta4 mode regression model to data from a model without a well-defined mode. Well-established results regarding inference based on misspecified models, especially likelihood-based frequentist inference (White, 1982), suggest that a sensible inference procedure is expected to conclude an inferred model in the misspecified family of models that is closest to the true model according to the Kullback-Leibler divergence criterion. With prior information of (misspecified) model parameters incorporated in a Bayesian inference procedure, posterior inference can be dominated by these prior information when data information severely contradict with the assumed model, e.g., when fitting a unimodal distribution to data that exhibit some striking multi-cluster structure. In our context, if one fits the beta4 mode regression model to, for instance, multimodal data, posterior inference for the support parameters may still be sensible, thanks to the prior formulation for θ_1 and θ_2. But all posterior inference results should be interpreted with caution when there is little update in the posterior inference from the prior distributions for the assumed model parameters.

4.3. Simulation III: model diagnostics

In the third simulation experiment we evaluate the performance of the Cox-Snell residual plot for model diagnostics. Here, despite the true model used to generate data $\mathcal{D}$, we fit a beta4 mean model in (2) with $h(\cdot)$ being the logit link, and $\boldsymbol{\beta}'\mathbf{x}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Covariate data (x_{i1}, x_{i2}), for $i = 1, \ldots, n$, are generated in the same way as in Simulation I (see the second paragraph in Section 4.1). We then generate response data from each of the following beta4 mean models, with $\phi = 50$ in (2), that are different in regard to the linear predictor specification or in the link function:

(C1) the model in (2) with $\boldsymbol{\beta}'\mathbf{x}_i = 1 + x_{i1} + x_{i2}$, and $h(\cdot)$ being the logit link,
(C2) the model in (2) with $\boldsymbol{\beta}'\mathbf{x}_i = 1 + x_{i1} + x_{i2} + x_{i1}^2$, and $h(\cdot)$ being the logit link,
(C3) the model in (2) with $\boldsymbol{\beta}'\mathbf{x}_i = 1 + x_{i1} + x_{i2}$, and $h(\cdot)$ being the probit link,
(C4) the model in (2) with $\boldsymbol{\beta}'\mathbf{x}_i = 1 + x_{i1} + x_{i2}$, and $h(\cdot)$ being the log-log link.

Under (C1), the assumed model coincides with the true model. Under (C2), the assumed model misspecifies the linear predictor; under (C3) and (C4), the assumed model involves a misspecified link function.

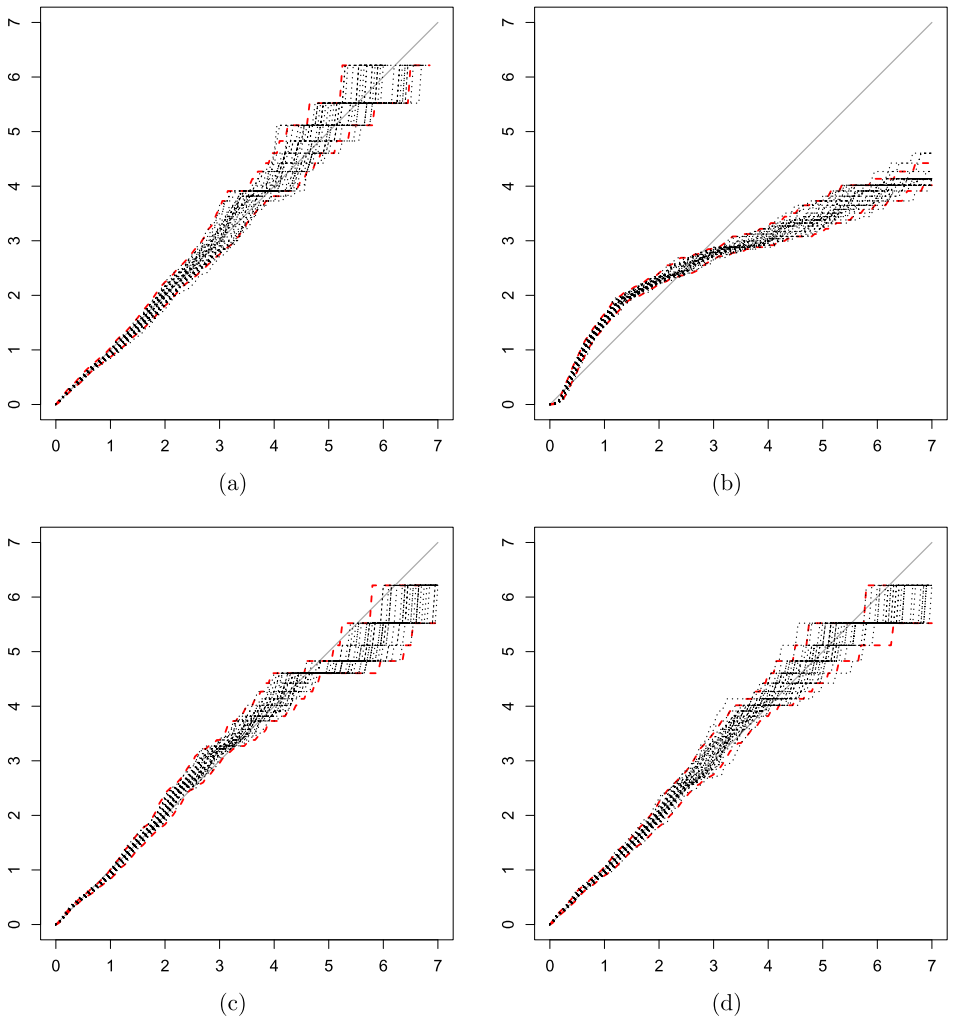

Fig. 1. Results from Simulation III. Cox-Snell residual plots with 95% point-wise credible intervals (thick dashed lines) for data of size $n = 500$ generated from a beta4 mean model. Panels (a)–(d) correspond to cases (C1)–C(4), respectively. A beta4 mean model with a linear predictor and the logit link is fit for all cases.

Fig. 1 presents the Cox-Snell residual plots obtained based on one data set of size $n = 500$ generated from each of (C1)–(C4). We can see that the Cox-Snell plot is very sensitive to linear predictor misspecification, but not as sensitive to the link misspecification. The same pattern was observed when one assumes a beta4 mode model for data generated from beta4 mode models that disagree with the assumed model in the linear predictor specification or in the link function (see supplementary Figure S2). In fact, link misspecification in a regression model has been shown to be notoriously difficult to detect by many well-accepted goodness-of-fit tests (e.g., Hosmer et al., 1997). Huang (2016) and Yu and Huang (2019) provided some insight on the reason behind this phenomenon, and proposed diagnostic tools deviating from the residual-based theme that are more powerful against link misspecification. Besides these frequentist goodness-of-fit tests, one may also consider the Bayesian goodness-of-fit test recently proposed by Barrientos and Canale (2021) that produces a Bayes factor following estimating the distribution of universal residuals (Brockwell, 2007). Besides fitting the posited regression model, that is, a beta4 regression model in our context, this Bayesian testing procedure also requires fitting a mixture normal model to the transformed residuals. Without looking into other diagnostics methods, here we recommend using the DIC, WAIC and LPML model criteria along with the residual plot in order to avoid less subtle link misspecification, such as assuming a logit link when the truth is a log-log link.

5. Real-life data applications

In this section we apply the proposed Bayesian regression methodology to analyze two data sets from real-life applications. A sample R code for implementing the proposed models using the provided R package `betaBayes` is available in supplementary Appendix A.

Table 4
Summary statistics of the COVID-19 data.

Variable	Minimum	Median	Mean	Maximum
Incidence rate	0.017	0.034	0.039	0.143
Death rate	0.006	0.020	0.021	0.046
MaleP	47.30	49.50	51.37	65.10
BlackP	2.90	11.70	14.54	55.80
Age65plusP	11.40	19.80	21.22	55.60
PovertyP	8.40	15.20	16.59	29.50
RUCC	1	2	3.19	9

Table 5
The values of DIC, WAIC, and LPML associated with four models for the COVID-19 data. The four considered models are the beta4 mean model (beta4-mean), the beta4 mode model (beta4-mode), and the regular beta mean (beta-mean) and beta mode (beta-mode) models supported on (0, 1).

Response	beta4-mean	beta-mean	beta4-mode	beta-mode
	negative DIC			
incidence	**437**	429	436	429
death	466	461	**475**	461
	negative WAIC			
incidence	**432**	424	431	424
death	464	459	**475**	459
	LPML			
incidence	**216**	212	215	212
death	232	229	**237**	229

5.1. Covid-19 data

To demonstrate the informativeness of inference results from the proposed regression models, we analyze a COVID-19 data set to examine the association between several county-level characteristics and the cumulative numbers of confirmed cases and deaths in the state of Florida. There are $n = 67$ counties in Florida. For each county, we collect the following variables: the cumulative number of confirmed cases and the cumulative number of deaths as of October 13, 2020, the total population census estimate, the percentage of people who are male (MaleP), the percentage of people who are black or African American (BlackP), the percentage of people who are 65 years and over (Age65plusP), the percentage of people whose income in the past 12 months is below poverty (PovertyP), and the 2013 Rural–Urban Continuum Code (RUCC). The RUCC varies from 1 to 9 (https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/), with a higher value indicating a more rural county. It distinguishes metropolitan counties by the population size of their metro area, and differentiates nonmetropolitan counties by the degree of urbanization and adjacency to a metro area. Although RUCC is an ordinal variable, it is statistically valid to treat RUCC as a continuous covariate in regression analysis (Yaghjyan et al., 2019). The COVID-19 case count and the death count are downloaded from https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. Based on these two counts, we define two county-level response variables: (1) the incidence rate, as the ratio of the cumulative number of confirmed cases to the total population; (2) the death rate, as the ratio of the cumulative number of deaths to the cumulative number of confirmed cases. County-level covariates listed above are based on the 2018 ACS 5-year estimates available at www.census.gov/data. Table 4 presents several summary statistics of the data.

5.1.1. Regression analysis

The two response variables defined above are naturally bounded within $(0, 1)$, assuming that a county has a zero probability of getting an incidence or death rate equal to zero or one. Without this assumption, one may need to consider a zero-inflated or one-inflated model, which is beyond the scope of the current study. For each response variable, we fit the beta4 mean model and the beta4 mode model using the default prior values given in Section 2.2, except that we set $a_{\theta_1} = 0$ and $b_{\theta_2} = 1$ to acknowledge the natural bound of $(0, 1)$. For the purpose of comparison, we also fit the regular $(0, 1)$-supported beta mean and mode models using the same prior values given in Section 2.2, except that we set $a_{\theta_1} = b_{\theta_1} = 0$ and $a_{\theta_2} = b_{\theta_2} = 1$ to fix (θ_1, θ_2) at $(0, 1)$. A logit link is used throughout. For each MCMC run we retain 5,000 scans thinned from 500,000 after a burn-in period of 20,000 iterations; convergence diagnostics deem this more than adequate. Table 5 lists values of DIC, WAIC, and LPML associated with all fitted models. Based on these model criteria, we conclude that the beta4 mean model outperforms all others when regressing the incidence rate on the considered covariates, and the beta4 mode model is the best when regressing the death rate on these covariates. Fig. 2 reports the Cox-Snell residual plots under the two best models, where we do not see severe deviation from the 45° line, indicating a goodness-of-fit under each chosen model.

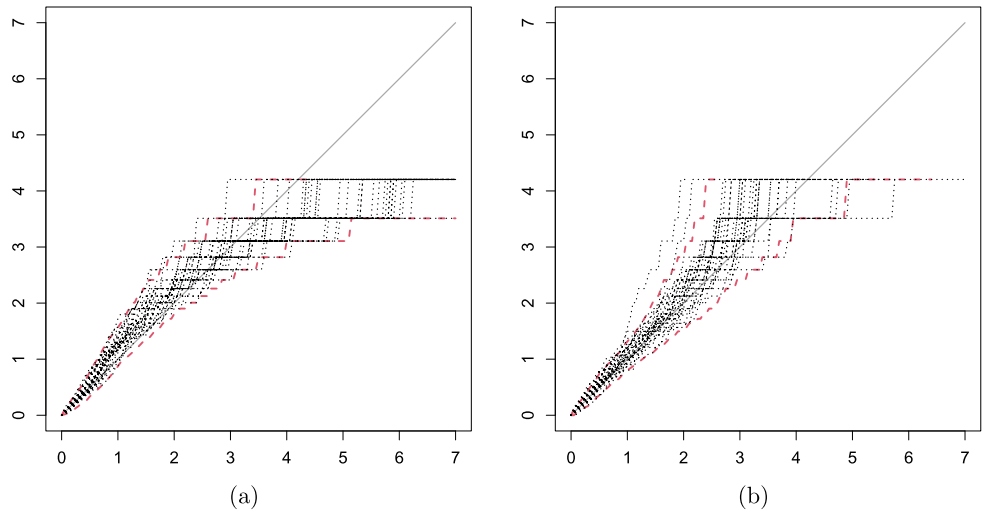

Fig. 2. COVID-19 data for Florida. Cox-Snell residual plots from (a) fitting the beta4 mean model for incidence and (b) from fitting the beta4 mode model for death rate.

Table 6
Estimates for model parameters based on the COVID-19 data. These include the posterior mean of covariate effects, the posterior mean of boundary parameters, and that of the precision parameter from fitting the beta4 mean model for the incidence rate, and all counterparts estimates when fitting the beta4 mode model for the death rate. The 95% credible interval associated with each parameter is given in parentheses following each posterior mean. We use * to highlight a statistically significant covariate effect.

	Incidence	Death
(Intercept)	-4.088 (-6.447, -1.646)	-2.041 (-13.16, 9.389)
MaleP	0.035 (-0.008, 0.078)	-0.035 (-0.246, 0.158)
BlackP	0.022 (0.008, 0.036)*	-0.013 (-0.070, 0.039)
Age65plusP	-0.044 (-0.072, -0.020)*	0.155 (0.072, 0.258)*
PovertyP	0.000 (-0.028, 0.028)	0.052 (-0.080, 0.192)
RUCC	0.164 (0.070, 0.266)*	-0.371 (-0.865, 0.064)
θ_1	0.016 (0.011, 0.017)	0.005 (0.003, 0.006)
θ_2	0.273 (0.173, 0.509)	0.049 (0.046, 0.058)
ϕ	51.12 (20.93, 120.6)	4.346 (2.286, 7.771)

Table 6 reports the covariate effects and boundary estimates under the beta4 mean model for the incidence rate and the beta4 mode model for the death rate. The posterior mean estimate of θ_2 indicates that the upper bound of county-level incidence rate for Florida is 0.273 with the 95% credible interval being $(0.173, 0.509)$, and the maximum county-level death rate for Florida is 0.049 with the 95% credible interval being $(0.046, 0.058)$. These estimates can be helpful information for local health officials when it comes to allocating medical supplies and health care professionals across the state. As for the covariate effects, we find that counties with a higher percentage of black people, a lower percentage of age 65+ people, and less degree of urbanization tend to have higher incidence rates. However, for the death rate, only the covariate Age65plusP is statistically significant, indicating that counties with a higher percentage of age 65+ people tend to have higher death rates. For example, for every one percent increase in Age65plusP, the quantile-score for the mode position within the true range (θ_1, θ_2) of the county-level death rate will increase by 0.155 on average, holding other covariates constant. These inference results regarding covariates effects can provide guidelines for state officials when planning for more targeted mitigation measures to control the disease spread and lower the death toll.

5.1.2. Additional analysis

Noting that some covariates are not statistically significant, one may perform variable selection based on posterior inference for covariate effects. Take the covariate MaleP as an example, one may use its posterior mean 0.035 and posterior standard deviation 0.022 to construct a score similar to a z-score for testing this covariate effect. One may then compute the probability under the normality assumption for the posterior distribution, $P(|Z| > |0.035/0.022|) = 0.112$, and use it like a p-value, along with the so-defined p-values associated with other covariates, to implement backward selection with the stopping rule that all p-values associated with the remaining covariates are less than 0.05. By adopting this strategy, we end up with a beta4 mean model for the incidence rate with three covariates (BlackP, Age65plusP, and RUCC), and a beta4 mode model for the death rate with only two covariates (Age65plusP and RUCC). The updated parameter estimates

Table 7
Effective sample sizes and running times (in seconds) of two algorithms implemented in `betaBayes` and `rstan`, respectively, applied to the COVID-19 data. The beta4 mean model is fit for the incidence rate, and the beta4 mode model is fit for the death rate. The MCMC setting for our `betaBayes`: burn-in=20,000, thinning=100, saved=5,000. The MCMC setting for `rstan`: burn-in=20,000, thinning=1, saved=5,000.

	Incidence		Death	
	`betaBayes`	`rstan`	`betaBayes`	`rstan`
Running time	30.47	161.80	39.80	173.37
(Intercept)	4666	5000	5000	5000
MaleP	4695	5000	5000	5000
BlackP	5000	5000	5000	5000
Age65plusP	4671	5000	4653	5359
PovertyP	5000	5479	4702	5000
RUCC	4769	5000	5000	5000
θ_1	3480	4761	4700	5000
θ_2	2492	4755	4173	5373
ϕ	2339	5000	3802	5000

are reported in the supplementary Table S2, along with the DIC, WAIC, and LPML values of these final models, all of which indicate improvement over the original models with all five covariates. As a follow-up research direction, one may develop more formal Bayesian variable selection procedures in the context of the proposed beta4 regression models by introducing a random indicator vector to indicate inclusion/exclusion of covariates, along with a g-prior (Liang et al., 2008; Guan and Stephens, 2011; Fisher and Mehta, 2014; Wang et al., 2015) or a spike-and-slab prior (Ročková and George, 2014; Chen et al., 2019; Zhang et al., 2021) for $\boldsymbol{\beta}$. Currently, the backward selection procedure based on posterior information makes a convenient tool data analysts can easily apply to look into other covariates of interest, such as the population density and the vaccination rate among adults in a county.

Finally, we compare our MCMC algorithm implemented in `betaBayes` with the HMC algorithm implemented using the software Stan (Stan Development Team, 2021) via the R package `rstan` (Stan Development Team, 2020). A sample R code for fitting a beta4 mean regression model using `rstan` is given in supplementary Appendix A. Table 7 provides the effective sample sizes and running times of these algorithms applied to the COVID-19 data. This comparison reveals that our proposed algorithm implemented in `betaBayes` has much lower effective sample sizes for most parameters than those in the HMC algorithm implemented in `rstan`, especially for (θ_1, θ_2) and ϕ. However, fitting a regression model is much faster when using `betaBayes` than when using `rstan`. Looking more closely at Table 7, one may argue that the choice of burn-in $= 20,000$ can be conservative and unfair to `rstan` in terms of running time comparison. To address this concern, besides what are presented in Table 7, we also set burn-in $= 200$ for `rstan`, which is comparable to the setting with burn-in $= 20,000$ and thinning $= 100$. Under this additional setting, the running times of `rstan` to fit the regression model for incidence rate and that for death rate are 53.94 and 43.17 seconds, respectively, both longer than the running times of `betaBayes`. Despite these and other differences between the two algorithms, they yield almost identical estimates for all model parameters, as one can see from the supplementary Table S3.

5.2. The Australian institute of sport data

Using their proposed regression models for a response assumed to be supported on $(0, 1)$, Bayes et al. (2012) and Migliorati et al. (2018) predicted an athlete's body fat percentage using the lean body mass based on a data set of $n = 37$ rowing athletes in the Australian Institute of Sport (AIS). Recall that Bayes et al. (2012) used a beta rectangular model based on a mixture of a beta distribution and a uniform distribution, and Migliorati et al. (2018) used a flexible beta model based on a special mixture of two beta distributions. In both existing works, the authors identified two outliers in the data, and inspected robustness of their inference results to the outliers. This is also the aspect of interest here in our analysis of the same data set, available in the R package `GLMsData` (Dunn and Smyth, 2018). Unlike their analyses, we do not assume (0, 1) as the support of the body fat percentage as we believe the actual support to be much narrower, although unknown.

5.2.1. Impact of outliers on model parameters estimation

To look into sensitivity of model parameters estimation to outliers, we first choose a regression model. Using the same prior settings as Section 5.1, we fit the beta4 mean, beta4 mode, beta mean, and beta mode models with the logit link to the full data. The DIC values associated with these four models are -155, -156, -136, and -136, respectively, indicating that the beta4 models perform much better than the latter two models. Between the former two models, we pick the beta4 mean model as the final model for further discussions on sensitivity of inference to outliers. The Cox-Snell residual plot shown in panel (a) of Fig. 3 indicates an overall goodness-of-fit of this chosen model, with the fitted curve depicted in panel (b) of this figure. The left half portion of Table 8 presents the corresponding model parameters estimates.

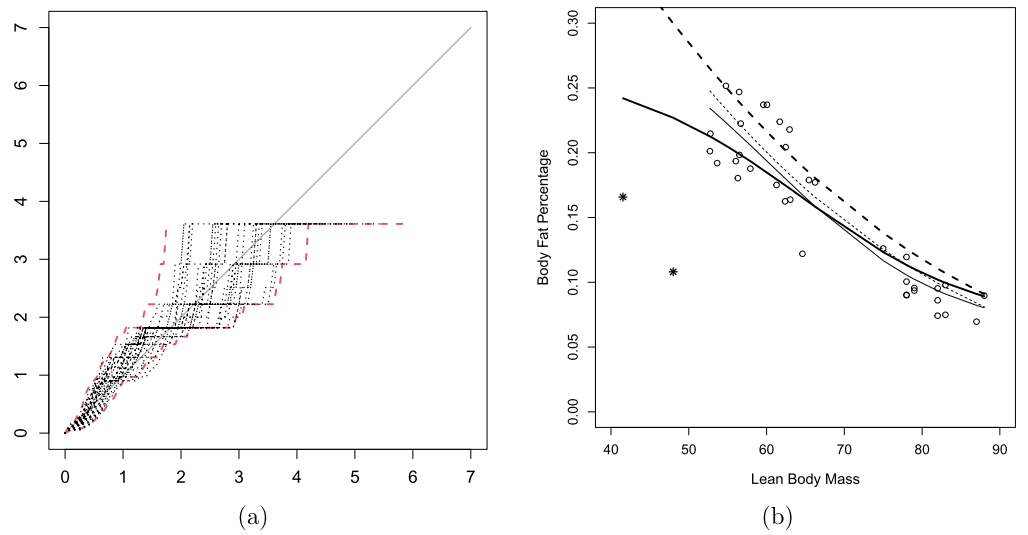

Fig. 3. Panel (a): Cox-Snell residual plots from fitting the beta4 mean model base on the AIS data with outliers. Panel (b): fitted regression curves for the mean of body fat percentage under the beta4 mean (solid) and the beta rectangular (dashed), based on the AIS data with (thick) and without (thin) outliers. The two potential outliers are marked as $*$.

Table 8
Estimates for model parameters based on the AIS data with and without outliers. The 95% credible interval associated with each parameter is given in parentheses following each posterior mean. We use * to highlight a statistically significant covariate effect associated with the lean body mass (LBM).

	With outliers	Without outliers
(Intercept)	5.531 (2.920, 7.933)	5.239 (1.598, 8.674)
LBM	-0.086 (-0.120, -0.051)*	-0.088 (-0.131, -0.045)*
θ_1	0.064 (0.029, 0.070)	0.054 (0.008, 0.070)
θ_2	0.273 (0.252, 0.355)	0.378 (0.256, 0.827)
ϕ	6.853 (3.309, 15.63)	36.52 (8.403, 140.5)

After fitting the beta4 mean model to the full data, we compute the CPO statistic defined in Section 3 for each data point. Two data points stand out with CPO values equal to 0.017 and 0.503, in contrast to other data points' CPO values that range from 4.91 to 33.68. We thus claim these two data points as outliers, marked as $*$ in panel (b) of Fig. 3. They are the same two outliers identified in Bayes et al. (2012) and Migliorati et al. (2018) via visual inspection of the scatter plot. We then remove the two outliers from the data and refit the beta4 mean model. The resultant parameter estimates are provided in the right half portion of Table 8, and the fitted regression curve is shown in panel (b) of Fig. 3.

One can see in Table 8 that the covariate effect estimates before and after removing outliers are very close, with the 95% credible interval of one estimate containing the other estimate. Hence, in this application, inference on the covariate effect based on the beta4 mean model is fairly robust to outliers. Not surprisingly, outliers do have a strong impact on the precision parameter estimation. Although less strong, outliers also have a noticeable influence on the boundary parameter estimation.

5.2.2. Impact of outliers on overall fit

We now turn to sensitivity to outliers of the overall fit for non-outlier observations based on the proposed regression models in comparison with other candidate models. For this purpose, we use the residual sum of squares (RSS) associated with the 35 non-outlier observations to assess the overall goodness-of-fit. Under a regression model, denote by $\hat{\Omega}_1$ the estimated model parameters based on the data excluding the two outliers, and by $\hat{\Omega}_2$ the counterpart estimates based on the full data in regression analysis. Following fitting a regression model, we compute $\text{RSS}_k = \sum_i (\hat{y}_i^{(k)} - y_i)^2$, where the sum is over all non-outlier observations, $\hat{y}_i^{(k)} = E(y_i|\mathbf{x}_i, \hat{\Omega}_k)$ for the mean regression, and $\hat{y}_i^{(k)} = \text{Mode}(y_i|\mathbf{x}_i, \hat{\Omega}_k)$ for the mode regression, for $k = 1, 2$. If, under a regression model, RSS_1 does not change much when comparing with RSS_2, then we say that the corresponding regression methodology is more robust to outliers in terms of the overall fit for non-outlier observations, despite potential non-robustness of parameter estimation to outliers reflected in the comparison between $\hat{\Omega}_1$ and $\hat{\Omega}_2$.

Besides the beta4 mean and beta4 mode models, we consider three other candidate models in this sensitivity analysis: the beta mean model, the beta mode model, and the beta rectangular model, all assuming the support being (0, 1). To

Table 9
Results regarding the residual sum of squares (RSS) based on the AIS data. These include the values of RSS under each considered model using the data without outliers (RSS_1), and the counterpart values when the complete data set is used (RSS_2). The % increase is defined as $(RSS_2 - RSS_1)/RSS_1$. The five considered models are the beta4 mean, the beta4 mode, the beta rectangular model (beta-rect) proposed in Bayes et al. (2012), and the regular beta mean and beta mode models supported on (0, 1).

	beta4-mean	beta-mean	beta-rect	beta4-mode	beta-mode
RSS_1	0.0197	0.0209	0.0234	0.0201	0.0213
RSS_2	0.0229	0.0294	0.0391	0.0243	0.0348
% increase	16%	41%	67%	21%	63%

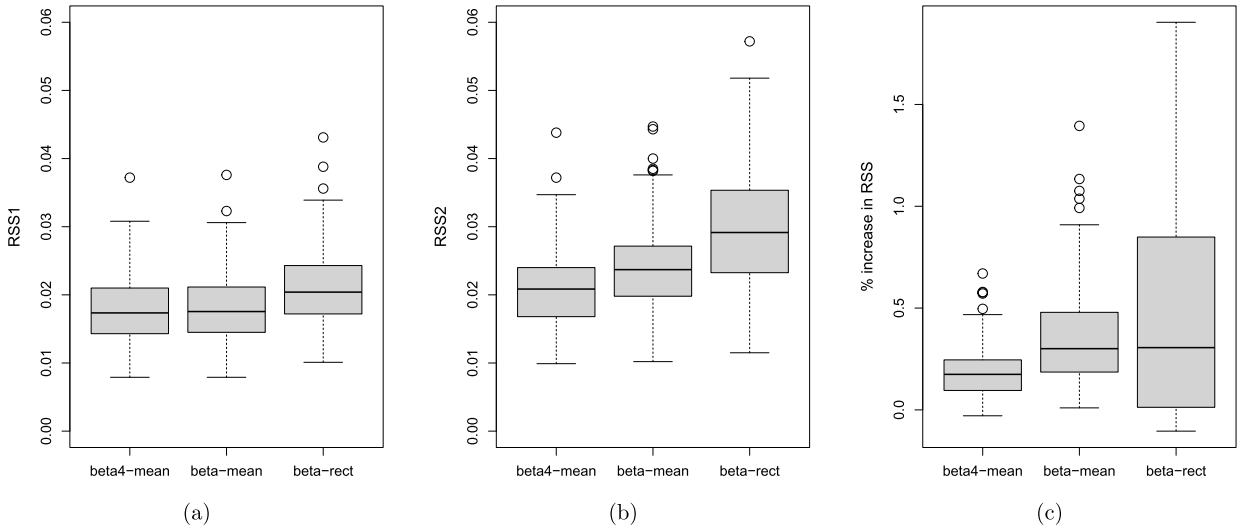

Fig. 4. Boxplots (across 300 MC replicates) of RSS_1 (in (a)), RSS_2 (in (b)), and % increase in RSS given by $(RSS_2 - RSS_1)/RSS_1$ (in (c)).

fit the beta rectangular model to the data, we use the statistical software Just Another Gibbs Sampler (JAGS, Plummer, 2003) via the R package `R2jags` (Su and Yajima, 2020). Table 9 reports the percent increase in RSS_2 relative to RSS_1, i.e., $(RSS_2 - RSS_1)/RSS_1$. One can see that the RSS value increases by 67% under the beta rectangular regression model, whereas there are merely a 16% and 21% increase under the beta4 mean and mode models, respectively. This indicates that, when it comes to goodness-of-fit for non-outlier observations, the proposed beta4 models are more robust to outliers than the beta rectangular model. Finally, for each $k \in \{1, 2\}$, the values of RSS_k associated with the beta4 models are the lowest among the five candidate models, suggesting better overall fit for the non-outlier observations whether or not outliers are excluded when estimating Ω. This last observation is reassuring especially given that, as seen in Section 5.2.1, outliers do have non-negligible impact on the estimation of some parameters in Ω that are involved in prediction.

5.2.3. A simulation study on sensitivity to outliers

To confirm that the robustness phenomenon of the new regression models to outliers in terms of overall fit is more than a coincidence observed in the application to the AIS data, we carry out a simulation study where data are generated from the $(0, 1)$-supported beta mean regression model. In particular, we use the covariate values from the AIS data in the beta4 mean model with $(\theta_1, \theta_2) = (0, 1)$ to generate response data, with $\boldsymbol{\beta} = (0.837, -0.038)'$ and $\phi = 229.32$, which are the posterior means of these parameters resulting from fitting a beta mean model to the AIS data after removing outliers. We then subtract 0.15 from the two response values corresponding to the lowest covariates to create outliers. Based on each of 300 MC replicate data sets, each of size 37, we repeat the sensitivity analysis presented in Section 5.2.2.

Fig. 4 reports the boxplots of 300 realizations of RSS_1, RSS_2, and the percentage increase of the latter relative to the former. As observed in the case study of the AIS data, the beta4 mean model indeed tends to be more robust to outliers than the beta mean model and the beta rectangular model in regard to goodness-of-fit for non-outlier observations in this simulation experiment. Also consistent with the comparison shown in Table 9, the beta4 mean model offers the best fit for non-outlier observations among the three candidate models whether or not outliers are excluded when estimating Ω, despite the fact that estimation of some parameters in Ω can be sensitive to outliers. Although related to prediction for non-outlier observations, RSS tends to be overly optimistic in assessing the quality of prediction. Hence, if one aims to assess robustness of predictions for non-outlier observations instead of overall fit for these observations, a different criterion, such as leave-one-out prediction error, should be used.

Migliorati et al. (2018) analyzed the same AIS data by fitting a flexible beta mean regression model derived from a special mixture of two beta distributions supported on $(0, 1)$. Their DIC values are -219 and -169 for the data without and with outliers, respectively, which are smaller than our DIC values of -179 and -155 under the beta4 mean model. This may not be surprising given the fact that their flexible beta distribution contains four parameters to characterize a distribution, thus offering greater flexibility for capturing various density shapes including bimodal ones. In contrast, the four-parameter beta distribution used to formulate our regression models only has two parameters to characterize the density shape, with the remaining two parameters used to define the support. As a future direction, it would be interesting to extend their flexible beta distribution by replacing the two beta mixture components with four-parameter beta distributions supported on (θ_1, θ_2). One downside of formulating the mixture in this way is that the resultant distribution cannot be easily used to perform mode regression.

6. Discussion

We propose a class of four-parameter beta regression models for studying the association between a continuous response bounded on an unknown interval and covariates via inferring either the conditional mean or mode of the response. Almost all existing approaches for analyzing bounded data assume a prefixed interval, which may not be accurate in many applications. To the best of our knowledge, the proposed regression models in this paper are the first regression framework allowing for an inference on the support boundaries along with inference for other model parameters. Moreover, this is also the first regression framework within which the regression function encompasses two central tendency measures, the mean and the mode. Besides offering more flexibility and shedding more light upon the association between a response and covariates, the benefit of unifying mean regression and mode regression in one parametric framework is that model comparison using likelihood-based model criteria becomes more convenient and meaningful. For each proposed model, we have developed efficient block-adaptive MCMC algorithms free of manual tuning for posterior sampling and a graphical model diagnostic tool to detect inadequate parametric assumptions. We have also provided a freely available R package `betaBayes` for fitting both proposed models and several competing models considered in this study.

We envision four directions of generalizing the proposed regression models upon completion of the current study. First, besides allowing the mode or mean parameter to depend on covariates, one may consider covariate-dependent precision parameter $\phi(\mathbf{x})$ to expand the class of beta4 regression models. Second, a more flexible family of regression models can be formulated via mixing a four-parameter beta with a uniform distribution by mimicking the construction of beta rectangular distributions (Bayes et al., 2012), or a mixture of two special four-parameter beta distributions similar to the construction of the flexible beta distribution (Migliorati et al., 2018). These mixture distributions will allow inclusion of distributions with heavier tails than those of four-parameter beta distributions. Third, one may consider a zero-inflated four-parameter beta regression model with an known upper bound of the response to account for an excess of zero values in, for example, the death rate of COVID-19 across counties in a state. Fourth, also motivated by the case study of COVID-19, one may allow the unknown support depend on covariates, such as the population density of a county when modeling the incidence rate. This last generalization presents more algorithmic and technical challenges (Chernozhukov and Hong, 2004; Hirano and Porter, 2003) that we plan to address in our follow-up study.

Acknowledgements

The authors wish to thank the Co-Editor, anonymous Associate Editor, and two referees for their insightful comments and suggestions that greatly improved the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107345.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer, pp. 199–213.
Baltazar-Aban, I., Pena, E.A., 1995. Properties of hazard-based residuals and implications in model diagnostics. J. Am. Stat. Assoc. 90 (429), 185–197.
Barrientos, A.F., Canale, A., 2021. A Bayesian goodness-of-fit test for regression. Comput. Stat. Data Anal. 155, 107104.
Barrientos, A.F., Jara, A., Quintana, F.A., 2017. Fully nonparametric regression for bounded data using dependent Bernstein polynomials. J. Am. Stat. Assoc. 112 (518), 806–825.
Bayes, C.L., Bazán, J.L., De Castro, M., 2017. A quantile parametric mixed regression model for bounded response variables. Stat. Interface 10 (3), 483–493.
Bayes, C.L., Bazán, J.L., García, C., et al., 2012. A new robust regression model for proportions. Bayesian Anal. 7 (4), 841–866.
Branscum, A.J., Johnson, W.O., Thurmond, M.C., 2007. Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. Aust. N. Z. J. Stat. 49 (3), 287–301.
Brockwell, A., 2007. Universal residuals: a multivariate transformation. Stat. Probab. Lett. 77 (14), 1473–1478.
Carnahan, J., 1989. Maximum likelihood estimation for the 4-parameter beta distribution. Commun. Stat., Simul. Comput. 18 (2), 513–536.
Chen, S., Walker, S.G., et al., 2019. Fast Bayesian variable selection for high dimensional linear models: marginal solo spike and slab priors. Electron. J. Stat. 13 (1), 284–309.
Cheng, R., Iles, T., 1987. Corrected maximum likelihood in non-regular problems. J. R. Stat. Soc. B 49 (1), 95–101.
Cheng, R., Traylor, L., 1995. Non-regular maximum likelihood problems. J. R. Stat. Soc. B 57 (1), 3–24.

Chernozhukov, V., Hong, H., 2004. Likelihood estimation and inference in a class of nonregular econometric models. Econometrica 72 (5), 1445–1480.
Christensen, B.J., Kiefer, N.M., 1991. The exact likelihood function for an empirical job search model. Econom. Theory, 464–486.
Claeskens, G., 2016. Statistical model choice. Annu. Rev. Stat. Appl. 3, 233–256.
Congdon, P., 2005. Bayesian Models for Categorical Data. John Wiley & Sons.
Cox, D.R., Snell, E.J., 1968. A general definition of residuals. J. R. Stat. Soc. B 30 (2), 248–275.
Cribari-Neto, F., Zeileis, A., 2010. Beta regression in R. J. Stat. Softw. 34 (2), 1–24.
Donald, S.G., Paarsch, H.J., 1993. Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. Int. Econ. Rev. 34 (1), 121–148.
Donald, S.G., Paarsch, H.J., 2002. Superconsistent estimation and inference in structural econometric models using extreme order statistics. J. Econom. 109 (2), 305–340.
Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. Phys. Lett. B 195 (2), 216–222.
Dunn, P.K., Smyth, G.K., 2018. GLMsData: generalized linear model data sets. R package version 1.0.0.
Epifani, I., MacEachern, S.N., Peruggia, M., 2008. Case-deletion importance sampling estimators: central limit theorems and related results. Electron. J. Stat. 2, 774–806.
Espinheira, P.L., Ferrari, S.L., Cribari-Neto, F., 2008a. Influence diagnostics in beta regression. Comput. Stat. Data Anal. 52 (9), 4417–4431.
Espinheira, P.L., Ferrari, S.L., Cribari-Neto, F., 2008b. On beta regression residuals. J. Appl. Stat. 35 (4), 407–419.
Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. J. Appl. Stat. 31 (7), 799–815.
Ferrari, S.L., Espinheira, P.L., Cribari-Neto, F., 2011. Diagnostic tools in beta regression with varying dispersion. Stat. Neerl. 65 (3), 337–351.
Figueroa-Zúñiga, J.I., Arellano-Valle, R.B., Ferrari, S.L., 2013. Mixed beta regression: a Bayesian perspective. Comput. Stat. Data Anal. 61, 137–147.
Fisher, C.K., Mehta, P., 2014. Fast Bayesian feature selection for high dimensional linear regression in genomics via the ising approximation. Preprint. arXiv:1407.8187.
Flinn, C., Heckman, J., 1982. New methods for analyzing structural models of labor force dynamics. J. Econom. 18 (1), 115–168.
Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Am. Stat. Assoc. 74 (365), 153–160.
Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. J. R. Stat. Soc. B 56 (3), 501–514.
Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24 (6), 997–1016.
Grün, B., Kosmidis, I., Zeileis, A., 2012. Extended beta regression in R: shaken, stirred, mixed, and partitioned. J. Stat. Softw. 48 (11), 1–25.
Guan, Y., Stephens, M., 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. 5 (3), 1780–1815.
Guolo, A., Varin, C., et al., 2014. Beta regression for time series analysis of bounded data, with application to Canada Google® flu trends. Ann. Appl. Stat. 8 (1), 74–88.
Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli 7 (2), 223–242.
Hanson, T.E., Branscum, A.J., Johnson, W.O., 2014. Informative g-priors for logistic regression. Bayesian Anal. 9 (3), 597–612.
Heidelberger, P., Welch, P.D., 1983. Simulation run length control in the presence of an initial transient. Oper. Res. 31 (6), 1109–1144.
Hirano, K., Porter, J.R., 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. Econometrica 71 (5), 1307–1338.
Hosmer, D.W., Hosmer, T., Le Cessie, S., Lemeshow, S., 1997. A comparison of goodness-of-fit tests for the logistic regression model. Stat. Med. 16 (9), 965–980.
Huang, X., 2016. Dual model misspecification in generalized linear models with error in variables. In: New Developments in Statistical Modeling, Inference and Application. Springer, pp. 3–35.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions. Vol. 2, 2nd edition. Wiley.
Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. 103 (481), 410–423.
Liu, J.S., Wong, W.H., Kong, A., 1994. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika 81 (1), 27–40.
McGarvey, R.G., Del Castillo, E., Cavalier, T.M., Lehtihet, E.A., 2002. Four-parameter beta distribution estimation and skewness test. Qual. Reliab. Eng. Int. 18 (5), 395–402.
Migliorati, S., Di Brisco, A.M., Ongaro, A., et al., 2018. A new regression model for bounded responses. Bayesian Anal. 13 (3), 845–872.
Mills, J.A., Prasad, K., 1992. A comparison of model selection criteria. Econom. Rev. 11 (2), 201–234.
O'Quigley, J., Xu, R., 2005. Goodness of fit in survival analysis. In: Encyclopedia of Biostatistics. John Wiley & Sons, Ltd.
Paarsch, H.J., 1992. Deciding between the common and private value paradigms in empirical models of auctions. J. Econom. 51 (1–2), 191–215.
Plummer, M., 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March 20–22, Vienna, Austria. ISSN 1609-395X.
Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6 (1), 7–11.
Raftery, A.E., Lewis, S.M., 1992. Comment: One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. Stat. Sci. 7 (4), 493–497.
Roberts, G.O., Sahu, S.K., 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. J. R. Stat. Soc., Ser. B, Stat. Methodol. 59 (2), 291–317.
Rocha, A.V., Simas, A.B., 2011. Influence diagnostics in a general class of beta regression models. Test 20 (1), 95–119.
Ročková, V., George, E.I., 2014. EMVS: the EM approach to Bayesian variable selection. J. Am. Stat. Assoc. 109 (506), 828–846.
Sargent, D.J., Hodges, J.S., Carlin, B.P., 2000. Structured Markov chain Monte Carlo. J. Comput. Graph. Stat. 9 (2), 217–234.
Smith, R.L., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika 72 (1), 67–90.
Smith, R.L., 1994. Nonregular regression. Biometrika 81 (1), 173–183.
Smithson, M., Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol. Methods 11 (1), 54.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B 64 (4), 583–639.
Stan Development Team 2020. RStan: the R interface to Stan. R package version 2.21.2.
Stan Development Team, 2021. Stan User's Guide. Version 2.26. http://mc-stan.org/.
Su, Y.-S., Yajima, M., 2020. R2jags: using R to Run 'JAGS'. R package version 0.6-1.
Turnbull, B.C., Ghosh, S.K., 2014. Unimodal density estimation using Bernstein polynomials. Comput. Stat. Data Anal. 72, 13–29.
Vehtari, A., Gelman, A., 2014. WAIC and cross-validation in Stan.
Verkuilen, J., Smithson, M., 2012. Mixed and mixture regression models for continuous bounded responses using the beta distribution. J. Educ. Behav. Stat. 37 (1), 82–113.
Vershynin, R., 2012. How close is the sample covariance matrix to the actual covariance matrix? J. Theor. Probab. 25 (3), 655–686.
Wang, J.Z., 2005. A note on estimation in the four-parameter beta distribution. Commun. Stat., Simul. Comput. 34 (3), 495–501.
Wang, M., Sun, X., Lu, T., 2015. Bayesian structured variable selection in linear regression models. Comput. Stat. 30 (1), 205–229.
Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica J. Econ. Soc. 50 (1), 1–25.

Yaghjyan, L., Cogle, C.R., Deng, G., Yang, J., Jackson, P., Hardt, N., Hall, J., Mao, L., 2019. Continuous rural-urban coding for cancer disparity studies: is it appropriate for statistical analysis? Int. J. Environ. Res. Public Health 16 (6), 1076.
Yu, S., Huang, X., 2019. Link misspecification in generalized linear mixed models with a random intercept for binary responses. Test 28 (3), 827–843.
Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. North-Holland/Elsevier, pp. 233–243.
Zhang, H., Huang, X., Han, S., Rezwan, F.I., Karmaus, W., Arshad, H., Holloway, J.W., 2021. Gaussian Bayesian network comparisons with graph ordering unknown. Comput. Stat. Data Anal. 157, 107156.
Zhou, H., Huang, X., Initiative, A.D.N., 2020. Parametric mode regression for bounded responses. Biom. J. 62 (7), 1791–1809.