Bayesian Analysis (2012) **7**, Number 4, pp. 841–866

A New Robust Regression Model for Proportions

Cristian L. Bayes*, Jorge L. Bazán† and Catalina García‡

Abstract. A new regression model for proportions is presented by considering the Beta rectangular distribution proposed by Hahn (2008). This new model includes the Beta regression model introduced by Ferrari and Cribari-Neto (2004) and the variable dispersion Beta regression model introduced by Smithson and Verkuilen (2006) as particular cases. Like Branscum, Johnson, and Thurmond (2007), a Bayesian inference approach is adopted using Markov Chain Monte Carlo (MCMC) algorithms. Simulation studies on the influence of outliers by considering contaminated data under four perturbation patterns to generate outliers were carried out and confirm that the Beta rectangular regression model seems to be a new robust alternative for modeling proportion data and that the Beta regression model shows sensitivity to the estimation of regression coefficients, to the posterior distribution of all parameters and to the model comparison criteria considered. Furthermore, two applications are presented to illustrate the robustness of the Beta rectangular model.

Keywords: Proportions, Beta regression, Bayesian estimation, link function, MCMC

1 Introduction

The Beta regression model was introduced by Ferrari and Cribari-Neto (2004) and it is adequate for situations where the variable of interest is continuous, restricted to the interval $(0,1)$ — such as percentages, proportions and fractions or rates (Kieschnick and McCullough, 2003) — and related to other variables through a regression structure.

Examples of these situations can be fluctuations in variables whose support is the standard unit interval, such as income concentration; unemployment rate; proportion of households subscribing to a cable TV service; proportion of household television viewing of various programming streams; proportion of votes for an incumbent president running for re-election, proportion of poor single mothers with children migrating from one state to another, etc.

In the Beta regression model, the regression parameters are interpretable in terms of the mean response, and in many aspects are similar to Generalized Linear Models (GLM). Estimation can be performed by maximum likelihood (Ferrari and Cribari-Neto 2004) or by Bayesian inference (Branscum et al. 2007). The Beta regression model

*Departamento de Ciencias, Pontificia Universidad Católica del Perú, Lima, Perú, cbayes@pucp.edu.pe
†Departamento de Ciencias, Pontificia Universidad Católica del Perú, Lima, Perú, jl-bazan@pucp.edu.pe
‡Departamento de Métodos Cuantitativos, Universidad de Granada, España, cbgarcia@ugr.es

 DOI:10.1214/12-BA728

is sufficiently documented by several publications such as Espinheira et al. (2008a,b); Ferrari et al. (2011); Cribari-Neto and Zeileis (2010) and several applications as, for example, Kelly et al. (2007) and Wallis et al. (2009).

However, the model proposed by Ferrari and Cribari-Neto (2004) does not model the dispersion but treats it solely as a nuisance parameter. As indicated by Smithson and Verkuilen (2006) this is a major oversight, as the ability to model dispersion can prove very useful. In addition, seemingly independently of Ferrari and Cribari-Neto (2004), Kieschnick and McCullough (2003) compared the performance of a Beta regression model for proportions, applied in economics and finance research, with several alternatives and conclude that it is often the best option.

Paolino (2001), by considering maximum likelihood, and Buckley (2002), by considering a Bayesian approach, use a Beta regression model which considers covariates between the mean and the dispersion parameters. This model, extensively presented in Smithson and Verkuilen (2006), is termed the variable dispersion Beta regression in Simas et al. (2010); Ferrari et al. (2011). In this case, the parameter that accounts for the precision of the data is not assumed to be constant across observations, as is the case with the Beta regression model, but is allowed to vary.

The Beta distribution can be considered flexible since the probability density function (pdf) can take different shapes by considering different values of its parameters. However, as was noted by Hahn (2008) and García et al. (2011), the Beta distribution neither considers tail-area events nor greater flexibility in the variance specification. We consider that this fact could limit its applications for modeling proportions. More discussion can be found in Kotz and van Dorp (2004).

It is important to consider diagnostic tools when estimating Beta regressions. The classical perspective of Ferrari and Cribari-Neto (2004) provided some guidelines for diagnostic analysis, including the use of two different residuals. Espinheira et al. (2008a) proposed two new Beta regression residuals and Espinheira et al. (2008b) proposed measures of influence analysis. By considering these measures, these authors, particularly Espinheira et al. (2008b), showed the influence of an observation in applications of the Beta regression model. However, no study has been carried out to show that the Beta regression model is not influenced by the estimation of regression parameters when there are outlying observations in the response variable. Thus, there is no robustness study of the Beta regression model from a Bayesian perspective by introducing perturbations in some observations to obtain artificial outliers and evaluating the impact of these perturbations on regression coefficients and on the measure of global fit. Following Pinheiro et al. (2001) and Barnett and Lewis (1995), we refer to an outlier as an observation (or set of observations) that appears to be inconsistent with the rest of the data.

In order to obtain some additional flexibility, we provide a regression model that permits varying amounts of dispersion and greater likelihood of more extreme tail-area events by considering the Beta rectangular distribution proposed by Hahn (2008). This new model includes the Beta regression model and the variable dispersion Beta regression model as particular cases. In addition, following a robust statistical modeling approach (Pinheiro et al. 2001) we will show that Beta rectangular regression is more robust than the Beta regression model.

The Beta rectangular distribution is just a mixture of a Beta distribution with a Uniform distribution and thus is a finite Beta mixture model as given by Bouguila et al. (2006). It is well-known that mixture distributions are more robust to outliers (comparatively large or influential values) since by including an extra distributional component in the model, the variability is better accounted for and the estimation of the "true" mean parameter is less affected, as indicated by Markatou (2000). As expected, we show that the Beta rectangular regression is a simple model and can be more "robust" compared to the usual Beta regression by considering simulation studies. In addition, we show an efficient way of simulating a posterior distribution.

The paper is organized as follows. In Section 2, we review the Beta rectangular distribution and a new parametrization is introduced. In Section 3, the Beta rectangular regression model is proposed, which can fit the mean and dispersion of the model. In the fourth section, a Bayesian inference approach is adopted using a Markov Chain Monte Carlo (MCMC) algorithm. Model comparison criteria for model selection such as Deviance Information Criterion (DIC), Expected Akaike Information Criterion (EAIC) and Expected Bayesian Information Criterion (EBIC) are also shown. In Section 5, two simulation studies are carried out to show that the new model is more robust than the Beta regression model in the presence of artificial outliers obtained by perturbing observations in the data by considering a sensitivity study on (a) the estimation of regression coefficients, (b) the posterior distribution of all parameters and (c) model comparison criteria. Section 6 presents two applications which are developed to show the utility of the Beta rectangular regression model. The conclusions are presented in Section 7.

2 Beta rectangular distribution

2.1 Beta distribution

As the Beta distribution is well-known, we will directly review the reparametrization proposed by Ferrari and Cribari-Neto (2004). Then, a random variable Y follows a Beta distribution if its probability density function (pdf) is given by:

$$b(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1 \tag{1}$$

where $0 < \mu < 1$ and $\phi > 0$. We consider the notation $Y \sim Beta(\mu, \phi)$. The mean and variance are expressed by:

$$E(y \mid \mu, \phi) = \mu \quad \text{and} \quad Var(y \mid \mu, \phi) = \frac{V(\mu)}{1+\phi} \tag{2}$$

where $V(\mu) = \mu(1-\mu)$, μ is the mean and ϕ can be interpreted as a precision parameter.

As indicated by Morris (1982) with this parametrization, the Beta distribution is a univariate exponential family but is not a natural exponential family. In addition, by considering that the quadratic variance function above matches that of the binomial, there is a failure to characterize the family within all exponential families.

2.2 Beta Rectangular distribution

The Beta distribution can initially be considered as flexible since the pdf can have different shapes by considering different values of μ and ϕ. However, as was noted by Hahn (2008) and García et al. (2011), the Beta distribution neither considers tail-area events nor greater flexibility in the variance specification. We consider that this fact could limit its application for modeling proportions. In order to get some additional flexibility, we provide a regression model which permits varying amounts of dispersion and greater likelihood of more extreme tail-area events by considering the Beta rectangular distribution proposed by Hahn (2008) whose pdf is given by:

$$f(y \mid \mu, \phi, \theta) = \theta + (1 - \theta) b(y \mid \mu, \phi), \tag{3}$$

where $0 \leq \theta \leq 1$ is a mixture parameter. Clearly the Uniform distribution is recovered when $\theta = 1$ and the Beta distribution is recovered when $\theta = 0$. In Figure 1 we show several probability density and cumulative distribution functions of the Beta rectangular distribution for different values of μ, ϕ and θ.

We consider the notation $Y \sim BR(\mu, \phi, \theta)$. Note that the Beta rectangular distribution is a mixture of two Beta variables, a $Beta\,(1/2, 2)$, or the Uniform distribution, and a $Beta(\mu, \phi)$. The mean and variance of this distribution are given by:

$$\begin{aligned} E(y \mid \mu, \phi, \theta) &= \frac{\theta}{2} + (1 - \theta)\mu \\ Var(y \mid \mu, \phi, \theta) &= \frac{V(\mu)}{1 + \phi}(1 - \theta)(1 + \theta(1 + \phi)) + \frac{\theta}{12}(4 - 3\theta). \end{aligned} \tag{4}$$

Note that when $\theta = 0$ the expressions given in (4) coincide with the expressions of the mean and variance of the Beta distribution. In a similar way when $\theta = 1$, these expressions coincide with the mean and variance of the Uniform distribution. In this case there is not a unique mode, which is of no particular difficulty since the mode is not a quantity of interest. When $0 < \theta < 1$ the mode is:

$$m = \frac{\mu\phi - 1}{\phi - 2} \tag{5}$$

which clearly is the same as that of the underlying Beta distribution.

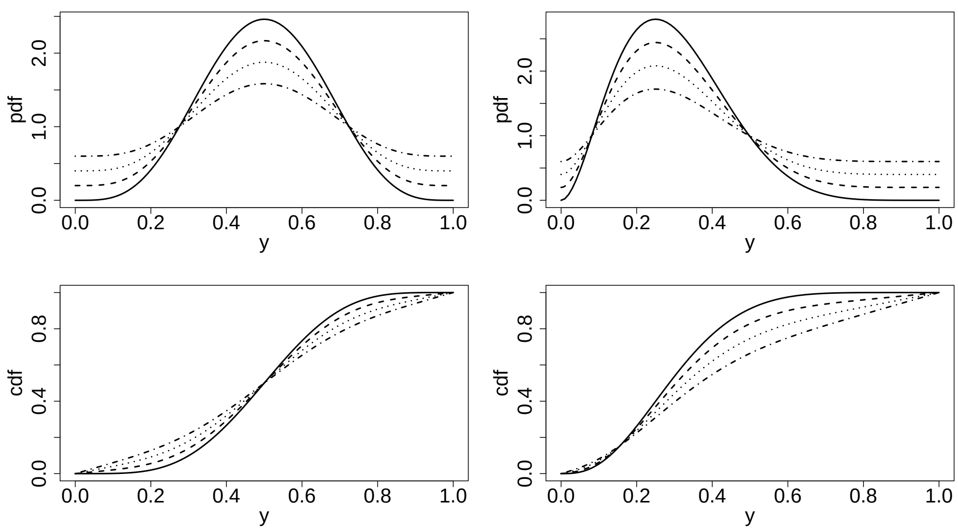

Figure 1: Examples of Beta rectangular pdf and cdf for $\mu = 0.5$ and $\phi = 10$ (left panels) and $\mu = 0.3$ and $\phi = 10$ (right panels) with different values of the mixture parameter θ: $\theta = 0$ (solid line), $\theta = 0.2$ (dashed line), $\theta = 0.4$ (dotted line) and $\theta = 0.6$ (dot dashed line).

2.3 A reparameterization of Beta Rectangular distribution

For a regression analysis, the mean of the response is typically modeled (Ferrari and Cribari-Neto 2004). However, the mean of the Beta rectangular distribution (4) is a function of the mixture parameter θ and μ. If we take $E(Y) = \dfrac{\theta}{2} + (1-\theta)\mu = \gamma$, we obtain that the parametric space of θ is restricted for the value of γ in the following way:

$$0 < \theta < 1 - |2\gamma - 1| \, .$$

In order to obtain a more appropriate regression structure for the mean of the Beta rectangular distribution, we take

$$\gamma = \frac{\theta}{2} + (1-\theta)\mu \quad \text{and} \quad \alpha = \frac{\theta}{1-(1-\theta)|2\mu - 1|} \tag{6}$$

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

Under this parametrization the pdf of the Beta rectangular distribution is given by:

$$\begin{aligned} g(y \mid \gamma, \phi, \alpha) &= \alpha\,(1-|2\gamma-1|) + (1-\alpha\,(1-|2\gamma-1|)) \\ &\quad \times b\left(y \,\middle|\, \frac{\gamma - 0.5\alpha\,(1-|2\gamma-1|)}{1-\alpha\,(1-|2\gamma-1|)}, \phi\right). \end{aligned} \tag{7}$$

We consider the following notation: $Y \sim BRr(\gamma, \phi, \alpha)$ with mean parameter γ. In addition, by considering Figure 2, which depicts the pdf (expression 7) of a BRr distribution for different values of γ, ϕ and α, we note that α is a shape parameter which is associated with the thickness of the tails of the distribution and ϕ is a parameter that seems to control the precision of the distribution; for larger values of ϕ we observe less dispersion. When $\alpha = 0$ the Beta distribution is recovered.

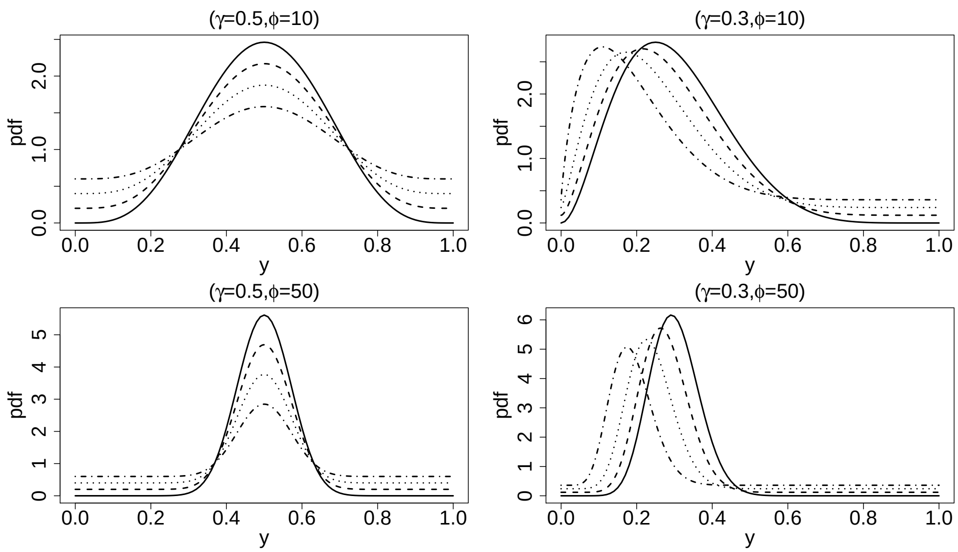

Figure 2: Examples of Beta rectangular pdf under reparametrization (6) for different values of γ, ϕ and α: $\alpha = 0$ (solid line), $\alpha = 0.2$ (dashed line), $\alpha = 0.4$ (dotted line) and $\alpha = 0.6$ (dot dashed line).

3 Beta Rectangular Regression

Let $\mathbf{Y} = (Y_1, ..., Y_n)^T$ be a vector of observed responses that takes values in $(0, 1)$. The Beta rectangular regression model is given by:

$$\begin{aligned} Y_i &\sim BRr(\gamma_i, \phi_i, \alpha), \\ F_1^{-1}(\gamma_i) &= \mathbf{x}_i^T \beta, \\ F_2^{-1}(\phi_i) &= -\mathbf{w}_i^T \delta \end{aligned} \tag{8}$$

where $\beta = (\beta_1, ..., \beta_k)$ and $\delta = (\delta_1, ..., \delta_l)$ are vectors of regression parameters, $\mathbf{x}_i = (x_{i1}, ..., x_{ik})$ and $\mathbf{w}_i = (w_{i1}, ..., w_{il})$ are vectors of k and l covariates (possibly overlapping or even identical) and $F_1^{-1}(\cdot)$ and $F_2^{-1}(\cdot)$ are strictly monotone and double differentiable functions on $(0, 1)$ to $\mathbf{R}$ in the first case and $\mathbf{R}^+$ to $\mathbf{R}$ in the second.

In general $F_1(\cdot)$ can be any cumulative distribution function corresponding to a continuous distribution where the inverse function is called the link function relating the mean parameter γ_i with covariates $\mathbf{x}_i$. Some examples of link functions for $F_1^{-1}(\cdot)$ are logit, probit and complementary log-log, but other links can be explored. $F_2^{-1}(\cdot)$ also is a link function relating the precision parameter ϕ_i with covariates $\mathbf{w}_i$. In addition, as the parameters ϕ_i must be strictly positive, the log link that can be used for $F_2^{-1}(\cdot)$ is $\log(\phi_i) = -\mathbf{w}_i^T\delta$ where we use the negative sign, as indicated by Smithson and Verkuilen (2006), to make the interpretation of the coefficients δ easier. Because ϕ_i is a precision parameter, a positive-signed δ_j indicates smaller variance, which is potentially confusing. It seems more natural to model dispersion rather than precision, and the negative sign enables us to do so.

The likelihood function for the model proposed can be written, alternatively, considering the parameterization given in (3) or the parameterization in (7) as:

$$L(\eta \mid \mathbf{Y}) = \prod_{i=1}^{n} f(y_i \mid \mu_i, \phi_i, \theta_i) = \prod_{i=1}^{n} g(y_i \mid \gamma_i, \phi_i, \alpha) \tag{9}$$

where $\eta = (\beta, \delta, \alpha)$, γ_i and ϕ_i are defined in (8) with $\theta_i = \alpha(1 - 2|\gamma_i - 1/2|)$ and $\mu_i = \dfrac{\gamma_i - \theta_i/2}{1 - \theta_i}$.

Note that in (8) when $\alpha = 0$ and ϕ_i is constant, we obtain the Beta regression model proposed by Ferrari and Cribari-Neto (2004) which is not an exponential family and so is not a GLM. Note also that, if solely $\alpha = 0$ then we obtain the variable dispersion Beta regression model introduced by Paolino (2001), Smithson and Verkuilen (2006) and recently by Simas et al. (2010). When other values of α are considered, the proposed regression model achieves greater flexibility.

4 Bayesian inference

The maximum likelihood estimators of the parameters of the model given in (8) can be obtained in a similar way to Ferrari and Cribari-Neto (2004). However, in some empirical situations, the sample size may be small and in such cases Bayesian inference seems to be more suitable if previous information about the parameters is available, see Congdon (2003) and Gelman et al. (2003). In this section we propose a Bayesian approach.

With independent data, the likelihood function for the Beta rectangular regression model is given by (9) so the posterior distribution is defined as:

$$p(\eta \mid \mathbf{Y}) \propto L(\eta \mid \mathbf{Y})p(\eta). \tag{10}$$

To complete the Bayesian specification of the model, we assume a prior distribution for all the unknown quantities $\eta = (\beta, \delta, \alpha)$. We assume that the elements of the parameter vector are independent and the following condition is verified:

$$p(\eta) = p(\beta)p(\delta)p(\alpha).$$

We can consider the usual priors to coefficients of regression to the mean and dispersion given by $\beta \sim N_k(\mathbf{a}, \mathbf{B})$ and $\delta \sim N_l(\mathbf{c}, \mathbf{D})$. In addition, we can consider a noninformative uniform prior $\alpha \sim Uniform(0, 1)$. So we have that the posterior distribution is given by

$$\begin{aligned} p(\eta \mid \mathbf{Y}) &\propto \prod_{i=1}^{n} \alpha \left(1 - \left|2F_1(\mathbf{x}_i^T\beta) - 1\right|\right) + \left(1 - \alpha\left(1 - \left|2F_1(\mathbf{x}_i^T\beta) - 1\right|\right)\right) \\ &\times\ b\left(y \,\middle|\, \frac{F_1(\mathbf{x}_i^T\beta) - 0.5\alpha\left(1 - \left|2F_1(\mathbf{x}_i^T\beta) - 1\right|\right)}{1 - \alpha\left(1 - \left|2F_1(x_i^T\beta) - 1\right|\right)}, F_2(\mathbf{w}_i^T\delta)\right) \\ &\times\ \phi_k(\beta \mid \mathbf{a}, \mathbf{B}) \times \phi_l(\delta \mid \mathbf{c}, \mathbf{D}) \times 1 \end{aligned} \tag{11}$$

where $\phi_k(.)$ and $\phi_l(.)$ denote the density function of a multivariate normal distribution of order k and l respectively.

In the particular case when we have the mean-regression model without modeling the dispersion parameter, the prior distributions considered for $\eta = (\beta, \phi, \alpha)$ are

$$p(\eta) = p(\beta)p(\phi)p(\alpha)$$

where $\phi \sim Gamma(c, c)$ as in Branscum et al. (2007). In preliminary explorations, we have observed that the inference on the coefficients of regression to the mean which are the main focus of interest are similar under different choices of hyperparameter $c = 0.1, 0.01, 0.001$.

The normalizing constant of the posterior distribution of the parameters in (11) cannot be obtained analytically. So, a possible approximation is obtained though MCMC to draw samples from the posterior density. A simple way is to implement a Metropolis sampling in OpenBUGS software (Lunn, Spiegelhalter, Thomas, and Best 2009). In addition, it is possible to implement a Metropolis-within Gibbs and slice sampling in WinBUGS software (Lunn, Thomas, Best, and Spiegelhalter 2000) which can be more stable. For this, we use an augmented version considering the usual mixture model representation similar to the procedure proposed by Bouguila et al. (2006) introducing latent variables for the mixture parameter, see for example the BUGS code in the appendix.

For all estimations in Sections 5 and 6, we simulate a large number of iterations, discarding the first 50 % of them as a burn-in period. To avoid correlation problems, we considered a spacing of size or thin equal to 10. Efficiency as measured by Effective Sample Sizes above 0.9 was obtained. In addition, convergence criteria given by Geweke (1992) as implemented in CODA (Plummer et al. 2006) were achieved.

4.1 Model Comparison Criteria

There are several methodologies for comparing competing models for a given dataset. We consider three approaches to Bayesian model selection: the Deviance Information Criterion (DIC, Spiegelhalter et al. 2002), the Expected Akaike Information Criterion (EAIC) and the Expected Bayesian Information Criterion (EBIC), which were proposed by Brooks (2002) and used for example in Bolfarine and Bazán (2010). These model criteria are simple to compute as the relevant quantities can be calculated directly from the MCMC output.

In order to obtain this, the following criteria are considered:

$$E\left[D(\eta)\right], \quad D(E[\eta]) \text{ and } \rho_D = E\left[D(\eta)\right] - D(E[\eta])$$

which represent the posterior mean of the deviance; deviance of posterior mean, which is obtained by considering the mean values of the posteriori of the parameters in the deviance of the model; and the effective number of parameters as given in Spiegelhalter et al. (2002).

These quantities can be estimated by using the MCMC output, considering the value of

$$\bar{D} = \frac{1}{B}\sum_{b=1}^{B} D\left(\eta^b\right), \quad \widehat{D} = D\left(\frac{1}{B}\sum_{b=1}^{B}\eta^b\right) \text{ and } \widehat{\rho_D} = \bar{D} - \widehat{D},$$

respectively, where B represents the number of iterations, and $D(\eta^b) = -2\log L(\eta^b \mid \mathbf{Y})$ is the value of the deviance in the iteration b.

The criteria EAIC, EBIC and DIC can be estimated using MCMC output by considering

$$\widehat{EAIC} = \bar{D} + 2p, \quad \widehat{EBIC} = \bar{D} + p\log(n) \text{ and } \widehat{DIC} = \bar{D} + \widehat{\rho_D} = 2\bar{D} - \widehat{D},$$

respectively, where p is the number of parameters in the model and n is the total number of observations.

Smaller values of $\bar{D}$, DIC, EBIC and EAIC imply better model fit.

5 Robustness Study

In this section, we compare the sensitivity of the Beta and Beta rectangular regression models in the presence of outliers by considering two simulation studies, one without covariates and the other considering covariates.

5.1 Simulation Study 1

In this subsection, an extensive simulation study is carried out to evaluate the relative performance of the procedure for estimating the Beta and Beta rectangular models without covariates for Beta data by considering (a) two values of $\phi = \{10, 30\}$, (b) three sample sizes $n = \{50, 100, 200\}$ and (c) percent of cases in outliers $r = \{2\%, 5\%, 8\%\}$.

First, a dataset with sample size n was simulated from the Beta distribution with parameters $\mu = 0.2$ and parameter of dispersion ϕ. Next, a sample of size $r \times n/100$ was taken without replacement from n. Later, these cases were replaced by values generated from the Uniform distribution (q,1), where q corresponds to the 0.999 quantile of the simulated Beta distribution — that is, outliers in the right tail of the distribution by considering values generated from another distribution. This data can be considered contaminated Beta data.

The combination of values of ϕ, n and r produce $2 \times 3 \times 3 = 18$ scenarios of contaminated simulated Beta data (see Table 1). For each scenario, 100 replications of contaminated Beta data were generated. For example, for scenario 2 (with $\phi = 10$, $n = 100$ and $r = 2\%$) 100 simulations of a sample of size 100 of a Beta$(\mu \times \phi = 2, (1 - \mu) \times \phi = 8)$ were initially simulated. Later, 2 of the 100 cases selected without replacement were replaced by values of a Uniform $(q, 1)$ where $q = 0.999, 2, 8$.

For each dataset generated, we fit the Beta rectangular and the Beta model, by considering the estimation procedure described in section (4). We burned-in $10,000$ of the $20,000$ values of the chain. The effective sample is $10,000$. An efficiency of over 0.9 was obtained by using the above Effective Sample Sizes.

Bias and MSE were obtained for each model by considering the replications in each scenario. The percentage of cases in which the Beta rectangular model was selected in relation to Beta was obtained. This was defined in terms of the percent of time that the Beta rectangular model achieved a lower DIC. The results are shown in Table 1. From the simulation study, we found that for any scenario with outliers, there was an improvement in the accuracy (bias and MSE decrease) for the estimation of the model's parameters when using a Beta rectangular rather than the Beta model for a contaminated dataset. We confirm the Beta rectangular model rather than the Beta model as the best choice in a high percentage of cases by considering the DIC. This became more evident as sample size increased and the percentage of outliers is incremented.

Table 1: Comparison of Bias, MSE and percentage of selection of the model Beta Rectangular versus Beta considering DIC for different scenarios of contaminated Beta data (two values of ϕ, three % of outliers and three sample sizes) by considering 100 dataset replications in each scenario.

	Scenarios			Beta		Beta Rectangular		DIC %
	ϕ	% of outliers	n	Bias	MSE	Bias	MSE	
1	10	2	50	0.0253	0.0010	0.0234	0.0008	0.66
2	10	2	100	0.0242	0.0007	0.0186	0.0005	0.81
3	10	2	200	0.0246	0.0007	0.0157	0.0003	0.93
4	10	5	50	0.0655	0.0047	0.0434	0.0022	0.9
5	10	5	100	0.0512	0.0028	0.0325	0.0012	0.95
6	10	5	200	0.0520	0.0028	0.0318	0.0011	1
7	10	8	50	0.0812	0.0070	0.0540	0.0032	0.94
8	10	8	100	0.0805	0.0067	0.0492	0.0026	1
9	10	8	200	0.0797	0.0065	0.0468	0.0023	1
10	30	2	50	0.0204	0.0006	0.0179	0.0004	0.85
11	30	2	100	0.0224	0.0006	0.0159	0.0003	0.97
12	30	2	200	0.0186	0.0004	0.0118	0.0002	0.99
13	30	5	50	0.0547	0.0034	0.0353	0.0014	0.99
14	30	5	100	0.0444	0.0021	0.0291	0.0009	0.99
15	30	5	200	0.0439	0.0020	0.0257	0.0007	1
16	30	8	50	0.0678	0.0050	0.0434	0.0020	0.96
17	30	8	100	0.0689	0.0049	0.0421	0.0018	1
18	30	8	200	0.0671	0.0046	0.0399	0.0016	1

5.2 Simulation Study 2: Sensitivity Study

In this subsection, we conduct a second specific study to compare the sensitivity in the presence of outliers in the Beta and Beta rectangular models, taking covariates into account when a Beta regression dataset is considered. In order to conduct this study, a dataset was first taken from a Beta regression model with parameters $\beta_0 = 0.5$, $\beta_1 = 1$, $\phi = 30$ and a logit link to consider the model regression structure. It was generated from $y_i \sim Beta(\mu_i, \phi)$ and $logit(\mu_i) = \beta_0 + \beta_1 x_i$ and $i = 1, .., n$, where $\mathbf{x} = (x_1, ..., x_n)^T$ was generated from a Uniform distribution on $(-3, 3)$ and $n = 200$. The data was an uncontaminated Beta dataset.

In Table 2, we present the posterior mean and 95% Bayesian confidence interval for the parameters of the model for both Beta and Beta Rectangular models. We note that the Beta and Beta Rectangular fits for the regression parameters were identical for this uncontaminated Beta dataset, but as expected the DIC, EAIC and EBIC for the Beta model (−532.0,−529.0,−519.1) are better than that corresponding to the Beta Rectangular model (−530.4,−525.4,−512.2).

Table 2: Posterior mean and 95% Bayesian credible interval for the parameters for the second simulated dataset.

Model	Beta			Beta Rectangular		
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%
β_0	0.54	0.47	0.61	0.53	0.47	0.60
β_1	0.98	0.93	1.03	0.98	0.92	1.03
ϕ	27.34	22.13	32.97	27.58	22.67	33.86
α				0.02	0.00	0.07

Next, for this uncontaminated Beta dataset we defined four contamination strategies by considering outlier patterns described in the next subsection.

Contaminated data under four perturbation patterns to generate outliers

To evaluate the influence of outlying observations on estimates of regression parameters, we consider contamination of 2% of the observations for the simulated dataset, that is to say, we replaced these data points y_i^* by their contaminated values $y_i^*(\Delta) = y_i^* \pm \Delta$. We consider four perturbation patterns:

(i) A decrease of Δ units of the response values to higher values of x,

(ii) An increase of Δ units of the response values to lower values of x,

(iii) A decrease and increase of Δ units of the response values for higher and lower values of x, respectively

(iv) A decrease of Δ units of the response values for central values of x.

See Figure 3. For perturbation patterns (i), (ii) and (iii) Δ varies from 0 to 0.8 with increments of 0.05 (17 cases) and for perturbation pattern (iv) Δ varies from 0 to 0.5 with increments of 0.05 (11 cases).

For each type of contaminated dataset, we studied three aspects in making a comparison between Beta and Beta Rectangular models: (a) sensitivity in the estimation of the regression coefficients; (b) sensitivity on the posterior distribution of all parameters by considering a measure of global influence and (c) effects on model comparison criteria.

(a) Sensitivity on the estimation of regression coefficients
For each perturbation pattern and each value of Δ we re-estimate the Beta and Beta Rectangular models. To re-estimate the models, we burned-in 10,000 of the 20,000 values of the chain considering thin equal to 10. Thus, we have a total of 1,000 samples upon which the posterior inference is based on. Figures 4 and 5 show the posterior mean and the 95%-credible interval for the regression parameters, β_0 and β_1 respectively, in

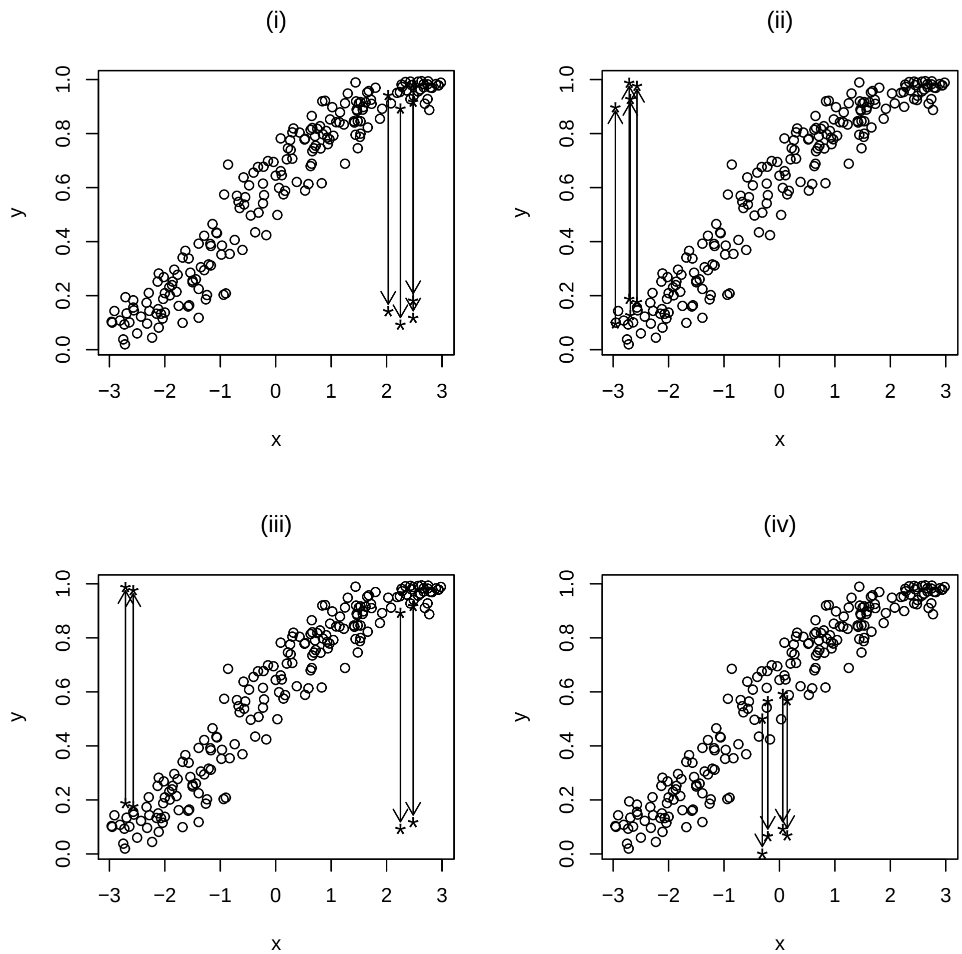

Figure 3: Generation of contaminated datasets under four different outlier patterns in the uncontaminated Beta dataset: (i) decrease Δ units of the response values to higher values of x, (ii) increase Δ units of the response values to lower values of x, (iii) decrease and increase Δ units of the response values for higher and lower values of x and (iv) decrease Δ units of the response values for central values of x.

each perturbation pattern and each case for the Beta and Beta Rectangular models. The posterior mean is depicted by a dashed line and 95%-credible interval by a solid line for both coefficients.

We find that the Beta Rectangular regression models are less sensitive to variations in Δ than the Beta regression models. The variations in Δ have a considerable impact for the Beta model on both the posterior mean and the credible interval for all the outlier patterns considered in this study. Note also that the size of the interval increases as $|\Delta|$ increases in particular for β_0.

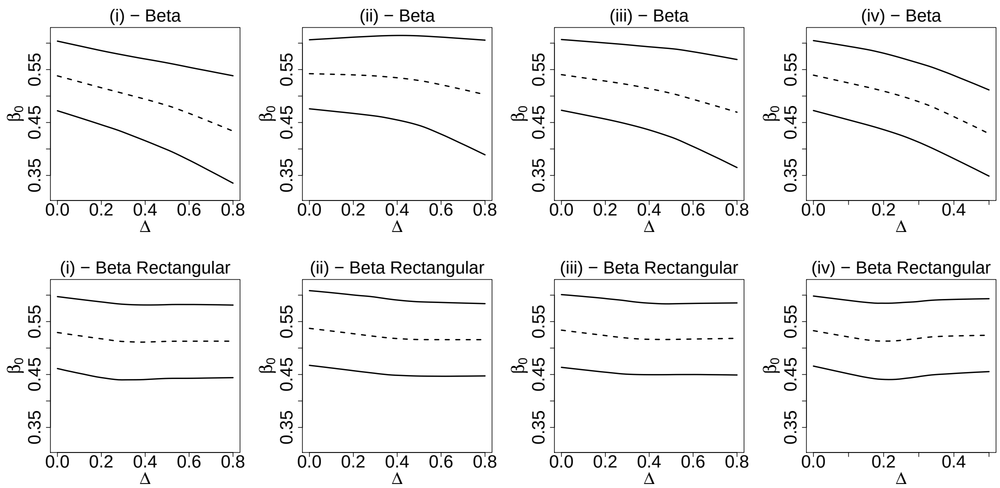

Figure 4: Posterior mean (dashed line) and 95%-credible interval (solid line) for β_0 under Beta and Beta Rectangular regression models for different values of perturbation Δ under four different outlier patterns in the contaminated Beta dataset.

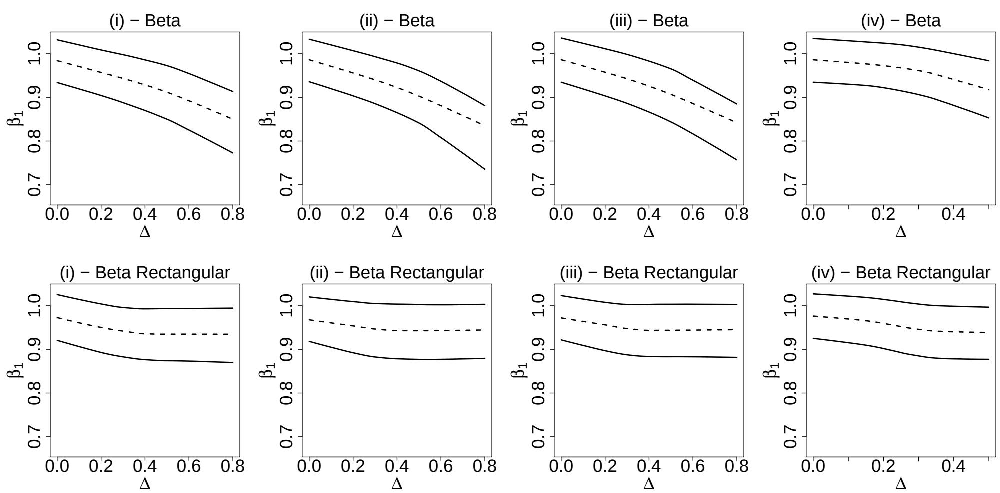

Figure 5: Posterior mean (dashed line) and 95%-credible interval (solid line) for β_1 under Beta and Beta Rectangular regression models for different values of perturbation Δ under four different outlier patterns in the contaminated Beta dataset.

(b) Sensitivity on the posterior distribution of all parameters

Now we evaluate the influence of the outlier perturbations for each Δ by considering a divergence measure between the posterior distribution with the uncontaminated data $p(\eta \mid \mathbf{Y})$ and the posterior distribution with the contaminated dataset $p(\eta \mid \mathbf{Y}(\Delta))$ for Beta and Beta Rectangular models, where $\mathbf{Y}(\Delta)$ represents the contaminated dataset. As a measure of divergence, we consider the Kullback−Leibler divergence measure $KL(\Delta)$

$$KL(\Delta) = \int \log\left(\frac{p(\eta \mid \mathbf{Y})}{p(\eta \mid \mathbf{Y}(\Delta))}\right) p(\eta \mid \mathbf{Y}) d\eta.$$

In order to estimate $KL(\Delta)$ we consider a Monte Carlo estimate of these divergences by using the MCMC output following the ideas of Peng and Dey (1995). In this case the KL estimate is given by:

$$\widehat{KL}(\Delta) = \log \frac{\{\prod_{b=1}^{B} 1/\vartheta(\eta^b)\}^{1/B}}{\{B^{-1}\sum_{b=1}^{B} \vartheta(\eta^b)\}^{-1}}$$

where $\vartheta(\eta^b) = f(y_i + \Delta \mid \eta^s)/f(y_i \mid \eta^s)$, $\{\eta^b\}_{b=1}^{B}$ are MCMC samples from the posterior distribution $p(\eta \mid \mathbf{Y})$ and B is the sample size of the MCMC procedure.

$KL(\Delta)$ can be considered a measure of influence. We estimate these measures for each Beta and Beta Rectangular model. We expected that the robust model where there were outliers would yield a lower estimated value of $KL(\Delta)$.

Figure 6 shows the $\widehat{KL}(\Delta)$ for both models with Beta shown as a solid line and Beta Rectangular shown as a dashed line under the four outlier patterns described above and for each Δ. We note that the $\widehat{KL}(\Delta)$ for the Beta regression model is an increasing function of Δ while for the Beta Rectangular regression model, it is more stable. That is to say, the influence of the outlying observations is unbounded for the Beta model, but clearly bounded for the Beta Rectangular model. That can be considered evidence that the Beta Rectangular model performs better compared with the Beta model.

In particular for closer contamination ($\Delta < 0.2$), all models display the same influence curve. This occurs because the contaminated observation is not distant enough from the typical data to be considered as an outlying observation. Therefore, all models have the same efficiency for close contamination cases.

(c) Sensitivity on model comparison criteria

In Figure 7 we study the behavior of the DIC under the presence of outlying observations. As in Figure 6 using KL, for closer contamination ($\Delta < 0.25$), the DIC favors the Beta regression model. Again, the reason is that the contaminated observation is not considered to be an outlying observation. Also the DIC, as has been shown in Figures 4, 5 and 6, has the same efficiency for close contamination cases. However for greater contamination $\Delta > 0.25$, the DIC favors the Beta Rectangular regression model, indicating the model's robustness where there are outlying observations.

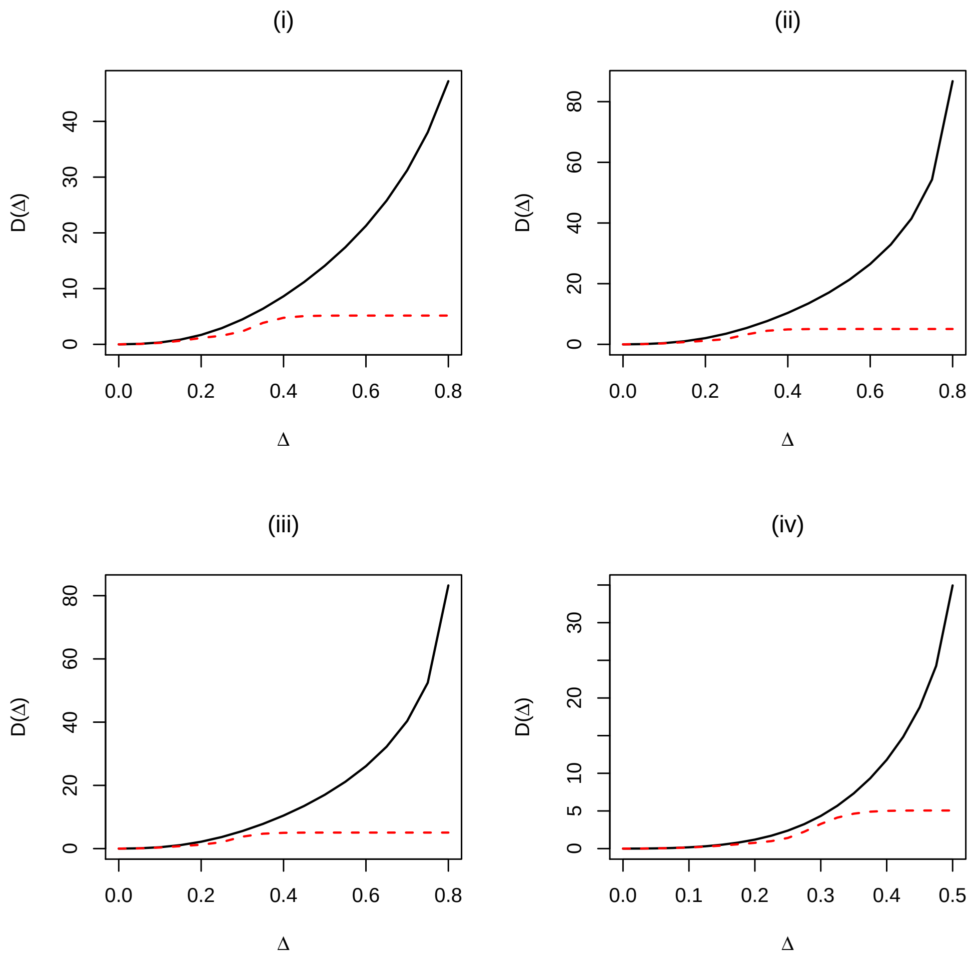

Figure 6: Kullback−Leibler divergence measure between $p(\eta \mid \mathbf{Y})$ and $p(\eta \mid \mathbf{Y}(\Delta))$ under Beta (solid line) and Beta Rectangular (dashed line) regression models for different values of perturbation Δ under four different outlier patterns in the contaminated Beta dataset.

6 Applications

6.1 Application 1

In this section, we consider the Bayesian analysis of the Australian Institute of Sport (AIS) Dataset included in the library `sn` for `R` (available for download at http://azzalini.stat.unipd.it/SN/index.html). We consider only the data of 37 rowing

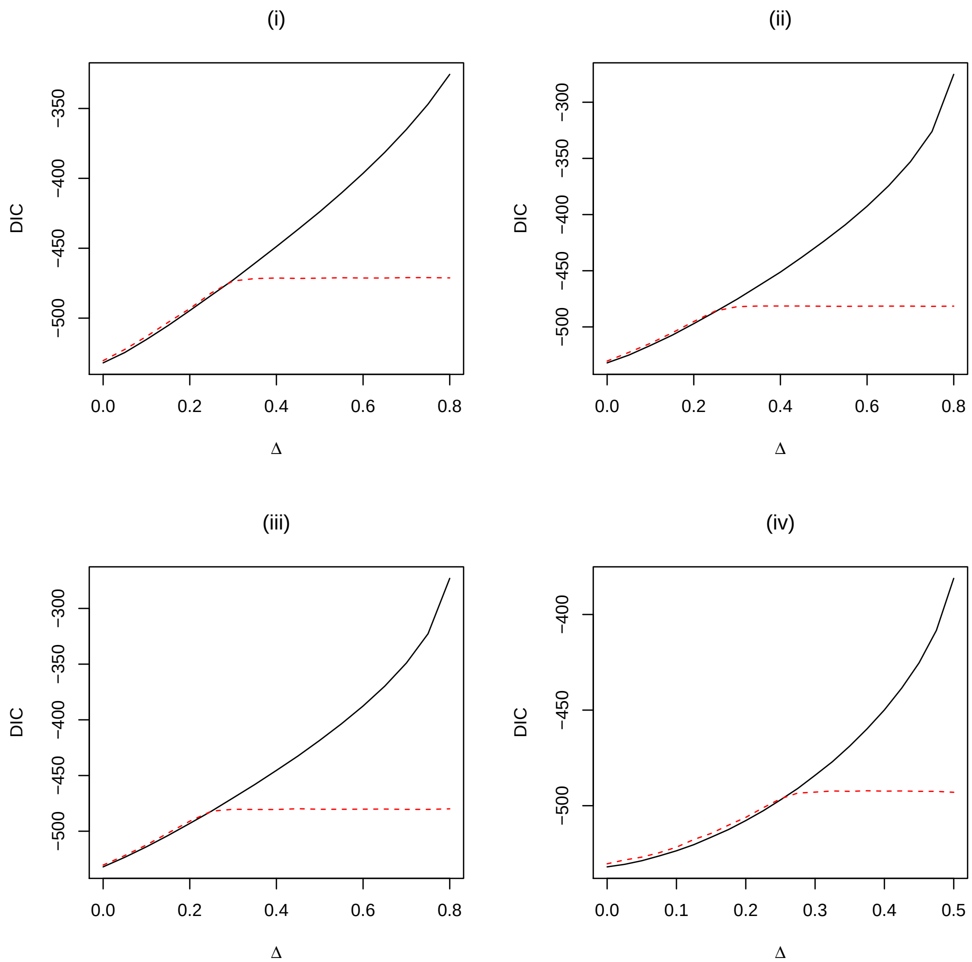

Figure 7: Deviance Information Criterion under Beta (solid line) and Beta Rectangular (dashed line) regression models for different values of perturbation Δ under four different outlier patterns in the contaminated Beta dataset.

athletes in the AIS dataset. We are interested in the prediction of the body fat percentage ($Bfat$) of each athlete by considering their lean body mass (lbm).

We define the Beta Rectangular regression model relating $Bfat_i$ and lbm_i as follows:

$$Bfat_i \sim BRr(\gamma_i, \phi, \alpha), \ \ logit(\gamma_i) = \beta_0 + \beta_1 lbm_i, \ \ i = 1, 2. \dots, n. \tag{12}$$

We fitted two models: a Beta Rectangular regression as defined in (12) and a Beta Regression (when $\alpha = 0$) considering all observations. To fit these models, we burned-

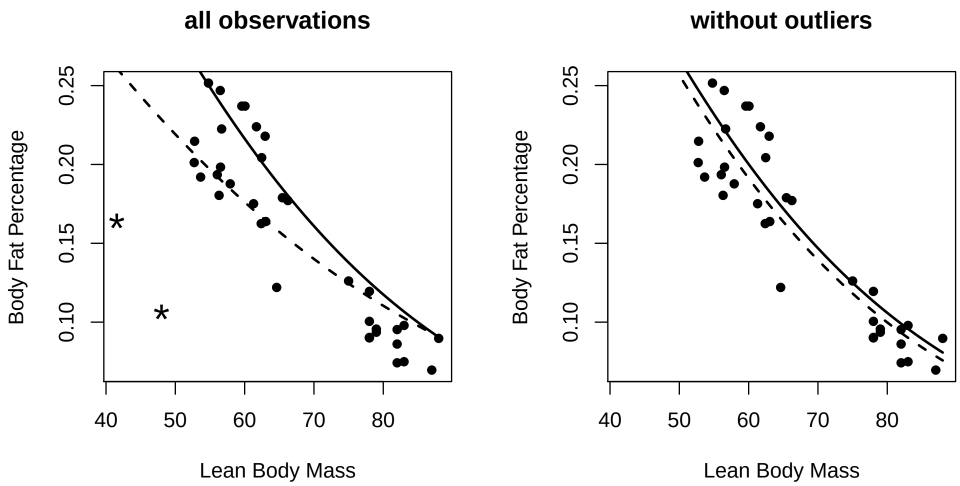

Figure 8: Scatter plot of the AIS dataset with fitted regression lines for the mean of $Bfat$ under the Beta distribution (dashed line) and Beta rectangular distribution (solid line) using all observations (left panel) and without outliers (right panel).

in 50,000 of the 100,000 values of a chain considering thin equal to 10 (the runtime for Beta Rectangular regression model is 206.28 seconds and the runtime for Beta regression model is 39.96 seconds on an Intel Core i-5 processor with 2.30 GHz and 4.00GB RAM). Thus, we have a total of 5,000 samples upon which the posterior inference is based. To carry out a model comparison, we computed DIC, EAIC and EBIC for each model. These criteria and the estimates of the parameters of the models are shown in Table 3. After considering various criteria, the Beta Rectangular regression model yielded the best fit.

As suggested by the Associate Editor, we identified two outlier observations marked as $*$ (see left panel in Figure 8) and then refitted both models after removing these two outliers. The results are shown in Table 3. As expected from Simulation Study 2, the estimates of the parameters are similar but the model comparison criteria correctly choose the Beta regression as the better model. This suggests that there is not much difference between the models in estimating the mean regression model when there are no outliers. However, outlying observations may have a considerable impact on estimates in the Beta regression model. Since this behavior was not observed under the Beta Rectangular regression model, it might be safer to use it rather than the Beta model.

Table 3: Posterior mean and a 95%−credible interval for the parameters under the Beta and Beta rectangular regression models for the AIS dataset using all observations and without outliers.

		Beta			Beta Rectangular		
Observations	Parameter	mean	2.5%	97.5%	mean	2.5%	97.5%
All	β_0	0.089	-0.447	0.611	0.900	0.436	1.310
	β_1	-0.027	-0.035	-0.019	-0.037	-0.043	-0.029
	ϕ	86.879	50.980	132.900	203.143	106.098	316.800
	α				0.188	0.043	0.405
	DIC		-136.13			-145.51	
	EAIC		-133.24			-141.55	
	EBIC		-128.41			-135.11	
Without outliers	β_0	0.830	0.412	1.244	0.859	0.443	1.264
	β_1	-0.038	-0.044	-0.031	-0.037	-0.044	-0.031
	ϕ	204.435	118.985	314.800	203.628	118.098	317.703
	α				0.072	0.002	0.235
	DIC		-160.41			-158.61	
	EAIC		-157.51			-153.81	
	EBIC		-152.84			-147.59	

6.2 Application 2

As a second application, we study the influence of the Human Development Index (HDI) of an electoral district on the proportion of blank votes in the 2006 Peruvian general election. We obtained the voting data from www.onpe.gob.pe, which provided data on past elections, and the HDI data from www.pnud.org.pe.

For these datasets, we consider the following general model:

$$y_i \sim BRr(\gamma_i, \phi, \alpha),\ logit(\gamma_i) = \beta_0 + \beta_1 HDI_i,\ log(\phi_i) = \delta_0 + \delta_1 HDI_i,\ i = 1, 2 \ldots, n \tag{13}$$

where y_i is the proportion of blank votes and HDI_i is the Human Development Index of electoral district i. Peru has $n = 194$ electoral districts. We fitted four models: (1) a Beta rectangular regression with variable dispersion as defined in (13); (2) a Beta regression with variable dispersion ($\alpha = 0$); (3) a Beta rectangular regression ($\delta_1 = 0$ and $\phi = \exp(\delta_0)$) and (4) a Beta regression ($\alpha = 0$, $\delta_1 = 0$ and $\phi = \exp(\delta_0)$). To fit these models, we burned-in 50,000 of the 100,000 values of a chain considering thin equal to 10. Thus, we have a total of 5,000 samples upon which the posterior inference is based.

Table 4 compares the four models using the model selection criteria presented in section 4.1. The runtimes for each model are also included. We note that the models that assume a Beta Rectangular distribution improved over the corresponding Beta models, in particular the Beta Rectangular model without variable dispersion gave the best fit. Posterior mean and a 95%−credible interval for the parameters under both models without variable dispersion are shown in Table 5.

Table 4: Model comparison measures for the 2006 Peruvian electoral dataset.

Model	Runtime*	DIC	EAIC	EBIC
Beta	195.11	-675.56	-672.61	-662.81
Beta Rectangular	1070.27	-695.71	-691.51	-678.44
Beta with variable dispersion	248.03	-674.28	-670.27	-657.20
Beta Rectangular with variable dispersion	1046.02	-693.81	-688.58	-672.24

(*) Runtime in seconds for 100000 iterations in an Intel Core i-5 processor with 2.30 GHz and 4.00GB RAM

Table 5: Posterior mean and a 95%−credible interval for the parameters under Beta and Beta Rectangular regression models for the 2006 Peruvian electoral dataset.

	Beta			Beta Rectangular		
Parameter	mean	2.5%	97.5%	mean	2.5%	97.5%
β_0	2.0472	1.6020	2.5060	2.2603	1.8640	2.6590
β_1	-6.2869	-7.1110	-5.4730	-6.6042	-7.3310	-5.8820
ϕ	80.6561	65.7890	97.9302	107.9722	86.0595	134.0000
α				0.0553	0.0119	0.1255

7 Final Comments

In this paper, we proposed a new regression model for responses measured in the unit interval. The main advantage of this model is its flexibility for working with data where there are outlying observations. Indeed, the Beta Rectangular regression includes the Beta regression model proposed by Ferrari and Cribari-Neto (2004) and the variable dispersion Beta regression model introduced by Smithson and Verkuilen (2006) as particular cases. However, other models can be formulated by considering other values of the mixture parameter between Beta and Uniform distributions.

The Beta Rectangular regression model is easily formulated as the generalization of the Beta regression model. Furthermore, the implementations of the Bayesian approach can easily be obtained in WinBUGS or PROC MCMC of SAS.

Simulation studies on the influence of outliers confirm that the Beta rectangular regression model seems to be a new robust alternative for modeling proportion data and that the Beta regression model offers sensitivity in the estimation of regression coefficients, sensitivity on the posterior distribution of all parameters by considering the Kullback-Liebler divergence as a measure of global influence, and effects on model comparison criteria (DIC) when data is contaminated under the four outlier patterns studied. In addition, the two applications presented here show the potential use of the

Beta Rectangular regression model for practitioners.

We note that WinBUGS may not be that efficient for a large data set, but it does make Bayesian inference with Beta Rectangular regression models easily accessible for applied researchers and its generic structure allows for a lot of flexibility in model specification. From the perspective of a practitioner, we expect that the WinBUGS code can help to popularize the model. Additional studies including the use of new Bayesian algorithms (see for example Gamerman and Lopes 2006) are needed to study how to improve the performance of the MCMC procedure for the Beta Rectangular regression model when the number of observations is extremely large. Also, a general Metropolis-Hastings algorithm could run much faster in a programming language, for example using C++ with R (Eubank and Kupresanin 2012). As pointed out by an anonymous referee, a possible extension of this work may improve the performance of the Beta Rectangular regression model when the number of observations is extremely large.

Additionally, we suggest future work should explore new choices of prior distributions for ϕ and α and their corresponding sensitivity studies for these choices, given that some sensitivity is observed for these parameters. For example we suggest to study the sensitivity of the choice of the hyperparameter c for the prior distribution $\phi \sim Gamma(c, c)$ as in Ghosh et al. (2009). By contrast, sensitivity was not observed in the estimates of the coefficients of regression parameters, which were the main focus of our analysis in this paper.

Also, diagnostic tools in Beta rectangular regression can be considered in future work, for instance in using the Kullback-Liebler divergence as a diagnostic measure for each observation as in Peng and Dey (1995) and Cho et al. (2009), and local influence and residual analysis as proposed by Ferrari et al. (2011) could be formulated from a Bayesian perspective.

Appendix

The WinBUGS code for a Beta Rectangular regression model with logit link for the mean parameter, a log link for the precision parameter and a single continuous covariate is presented below. For example, in the electoral dataset, x denotes HDI and y denotes the proportion of blank votes of an electoral district. The hyperparameters in the priors for all the parameters need to be specified by the user.

```
model{
for(i in 1 : n) {
y[i] ~ dbeta(a[i,u[i]], b[i,u[i]])

v[i] ~ dbern(theta[i])
u[i]<-v[i]+1

a[i,1] <- mu[i]*phi[i]
b[i,1] <- (1-mu[i])*phi[i]
a[i,2] <- 1
```

```
b[i,2] <- 1


logit(gamma[i]) <-  beta0.star+beta1 * (x[i] -mean(x[]))
log(phi[i]) <- -delta0.star-delta1 * (x[i] -mean(x[]))

theta[i]<-alpha*(1-2*abs(gamma[i]-.5))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}

beta1  ~dnorm (,)
beta0.star  ~dnorm (,)
beta0<-beta0.star-beta1 * mean(x[])

delta1  ~dnorm (,)
delta0.star  ~dnorm (,)
delta0<-delta0.star-delta1 * mean(x[])

alpha.star ~ dunif(0,1)
}
```

References

Barnett, V. and Lewis, T. (1995). *Outliers in statistical data*. Wiley. 842

Bolfarine, H. and Bazán, J. L. (2010). "Bayesian Estimation of the Logistic Positive Exponent IRT Model." *Journal of Educational and Behavioral Statistics*, 35: 693–713. 849

Bouguila, N., Djemel, Z., and Monga, E. (2006). "Practical Bayesian estimation of a finite beta mixture through Gibbs sampling and its applications." *Statistical Computing*, 16: 215–225. 843, 848

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). "Bayesian Beta regression; application to household data and genetic distance between foot-and-mouth disease viruses." *Australian & New Zealand Journal of Statistics*, 49(3): 287–301. 841, 848

Brooks, S. P. (2002). "Discussion on the paper by Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A." *Journal of the Royal Statistical Society, Series B*, 64: 616–618. 849

Buckley, J. (2002). "Estimation of models with beta-distributed dependent variables: A replication and extension of Paolino (2001)." *Political Analysis*, 11: 1–12. 842

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009). "Bayesian Case Influence Diagnostics for Survival Models." *Biometrics*, 65: 116–124. 861

Congdon, P. (2003). *Applied Bayesian Modelling*. John Wiley & Sons. 847

Cribari-Neto, F. and Zeileis, A. (2010). "Beta Regression in R." *Journal of Statistical Software*, 34(2): 1–24.
URL http://www.jstatsoft.org/v34/i02/ 842

Espinheira, P., Ferrari, S., and Cribari-Neto, F. (2008a). "Influence diagnostics in beta regression." *Computational Statistics & Data Analysis*, 52(9): 4417–4431. 842

— (2008b). "On Beta Regression Residuals." *Journal of Applied Statistics*, 35(4): 407–419. 842

Eubank, R. L. and Kupresanin, A. (2012). *Statistical Computing in C++ and R* . Chapman & Hall/CRC. 861

Ferrari, S. and Cribari-Neto, F. (2004). "Beta regression for modelling rates and proportions." *Journal of Applied Statistics*, 31: 799–815. 841, 842, 843, 845, 847, 860

Ferrari, S., Espinheira, P. L., and Cribari, F. (2011). "Diagnostic Tools in Beta Regression with Varying Dispersion." *Statistica Neerlandica*, 65: 337–351. 842, 861

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC. 861

García, C., García, J., and van Dorp, J. R. (2011). "Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support." *Statistical Methods and Applications*, 20(4): 463–486. 842, 844

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman and Hall. 847

Geweke, J. (1992). "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 169–193. Oxford: Oxford University Press. 849

Ghosh, P., Bayes, C. L., and Lachos, V. H. (2009). "A Robust Bayesian Approach to Null Intercept Measurement Error Model with Application to Dental Data." *Computational Statistics and Data Analysis*, 53: 1066–1079. 861

Hahn, E. D. (2008). "Mixture densities for project management activity times: A robust approach to PERT." *European Journal of Operational Research*, 188: 450–459. 841, 842, 844

Kelly, G., Garabed, R., Branscum, A., Perez, A., and Thurmond, M. (2007). "Prediction model for sequence variation in the glycoprotein gene of infectious hematopoietic necrosis virus in California, USA." *Diseases of Aquatic Organisms*, 78: 97–104. 842

Kieschnick, R. and McCullough, B. D. (2003). "Regression analysis of variates observed on (0,1): percentages, proportions, and fractions." *Statistical Modeling*, 3: 193–213. 842

Kotz, S. and van Dorp, J. R. (2004). *Beyond Beta: other continuous families of distributions with bounded support and applications* . World Scientific Press, Singapore. 842

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). "The BUGS project: Evolution, critique and future directions (with discussion)." *Statistics in Medicine*, 3049–3082. 848

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and Computing*, 10: 325–337. 848

Markatou, M. (2000). "Mixture Models, Robustness, and the Weighted Likelihood Methodology." *Biometrics*, 56: 483–486. 843

Morris, C. N. (1982). "Natural exponential families with quadratic variance functions." *The Annals of Statistics*, 10: 65–80. 844

Paolino, P. (2001). "Maximum likelihood estimation of models with beta-distributed dependent variables." *Political Analysis*, 9: 325–346. 842, 847

Peng, F. and Dey, D. K. (1995). "Bayesian analysis of outlier problems using divergence measures." *The Canadian Journal of Statistics*, 23: 199–213. 855, 861

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). "Efficient Algorithms for Robust estimation in Linear Mixed-Effects Models Using the Multivariate t Distribution." *Journal of Computational and Graphical Statistics*, 10: 249–276. 842

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, 6(1): 7–11. 849

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). "Improved estimators for a general class of beta regression models." *Computational Statistics and Data Analysis*, 54(2): 348–366. 842, 847

Smithson, M. and Verkuilen, J. (2006). "A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables." *Psychological Methods*, 11(1): 54–71. 841, 842, 847, 860

Spiegelhalter, D. J., Best, N., Carlin, B., and Van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)." *Journal of the Royal Statistical Society, Series B*, 64: 583–640. 849

Wallis, E., Mac Nally, R., and Lake, S. (2009). "Do tributaries affect loads and fluxes of particulate organic matter, inorganic sediment and wood? Patterns in an upland river basin in south-eastern Australia." *Hydrobiologia*, 636: 307–317. 842

Acknowledgments

The authors are grateful to the Editor-in-Chief, an anonymous Referee and an Associate Editor, whose comments have led to an improved version of our manuscript. We would like to thank the Andalusian Regional Governments Ministry of Innovation, Science and Enterprise as sponsor of the Scholarship Program of Academic Mobility AUIP. This work was funded by the Dirección de Gestión de la Investigación at PUCP through grants DGI-2011-0222, DGI-2011-0173 and DGI-2010-0072.