Beta Regression in the Presence of Outliers - a wieldy Bayesian solution

Janet van Niekerk[1], Andriette Bekker[1] and Mohammad Arashi[1,2]
[1]University of Pretoria, South Africa
[2]Shahrood University of Technology, Iran
janet.vanniekerk@up.ac.za

February 5, 2019

Abstract

Real phenomena often leads to challenges in data. One of these is outliers or influential values. Especially in a small sample, these values can have a major influence on the modeling process. In the beta regression framework, this issue has been addressed mainly in two ways: the assumption of a different response model and the application of a minimum density power divergence estimation (MDPDE) procedure. In this paper, however, we propose a simple hierarchical Bayesian methodology in the context of a varying dispersion beta response model that is robust to outliers, as shown through an extensive simulation study and analysis of two real data sets. To robustify Bayesian modeling a heavy-tailed Student's t prior with uniform degrees of freedom is adopted for the regression coefficients. This proposal results in a wieldy implementation procedure which avails practical use of the approach.

1 Introduction

Data within the unit interval is prevalent in many experiments within the applied sciences like medical research and sociology, amongst others. Financial data is also often reported in terms of rates or percentages. Data like this can be modelled using the beta distribution since it provides support for the unit interval $(0, 1)$. The beta distribution is characterized by two parameters α and β with density function

$$f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} \quad 0 < y < 1, \tag{1}$$

where $B(\alpha, \beta)$ is the beta function. The expected value is $\alpha\beta$ and the variance is $\alpha\beta^2$. In many real life applications there is a variable of interest i.e. a response variable as well as predictors which could be used to explain the underlying data-generating mechanism of the response for improved inference. In this case the density function (1) proves difficult to implement and interpret as the response model in a regression setup. For this reason, a reparameterized version of the beta distribution is used with the density function as follows:

$$f(y) = \frac{y^{\mu\phi-1}(1-y)^{\phi(1-\mu)-1}}{B(\mu\phi, \phi(1-\mu))} \quad 0 < y < 1, \tag{2}$$

with $\mathrm{E}(Y) = \mu, 0 < \mu < 1, \mathrm{Var}(Y) = \frac{\mu(1-\mu)}{\phi+1}, \phi > 0$. We use the notation $y \sim \mathrm{Beta}(\mu, \phi)$, where μ is the location parameter and ϕ can be interpreted as a precision parameter since large values of ϕ results in a smaller variance. Comparing (1) and (2) we note that $\alpha = \mu\phi, \beta = \phi(1-\mu)$. This framework was adopted by Paolino,(Paolino, 2001) Kieschnick and McCullough (2001) and Ferrari and Cribari-Neto (2004) amongst others to develop a beta regression model. The beta regression model and its variants/compositions have become popular in medical studies because of their practical use. (Albert et al., 2014; Liu and Li, 2016; Wang and Luo, 2017; Kim and Lee, 2017; Liu and Eugenio, 2018)

Initially, the beta regression model was developed assuming a homogeneous precision parameter and a location sub-model in the form of a generalized linear regression model (GLM) for the location parameter, using a link function

$$\boldsymbol{g}(\mu) = \boldsymbol{X}^\top \boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{x}_i \in \mathbb{R}^m$ is the i-th covariate, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ and $\boldsymbol{g}(.)$ is an appropriate link function. Later, this framework was developed to accommodate heterogeneous precision using a precision sub-model as follows,

$$\boldsymbol{h}(\phi) = \boldsymbol{Z}^\top \boldsymbol{\gamma}, \tag{4}$$

$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is the i-th covariate, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ and $\boldsymbol{h}(.)$ is an appropriate link function. Since the initial development based on maximum likelihood estimates, a Bayesian framework for the estimation of the parameters has been proposed by Cepeda-Cuervo et al. (2016) Bayesian analysis has been shown to perform competitively well when compared to the MLE and often better in the case of small samples. (McNeish, 2016) Currently there are two R packages available to practitioners for fitting a beta regression model, *betareg* and *Bayesianbetareg*. The first package uses maximum likelihood estimation to estimate the regression coefficients for a location and precision model, based on the work by Ferrari and Cribari-Neto (2004). A Bayesian framework using normal priors are used in the second package, based on the work by Cepeda-Cuervo et al. (2016) Both of these packages are very efficient in the estimation process. Although this framework is for responses within the unit interval, it can be applied to data within any finite interval using the well-known transformation $\frac{y-a}{b-a} \in [0,1]$ for $a < y < b$. Also, data consisting of datum inside a certain interval as well as on the boundaries can be transformed to be within the unit interval using the transformation $\frac{\frac{y-a}{b-a}(n-1)+0.5}{n} \in (0,1)$, with no observations on the boundaries.

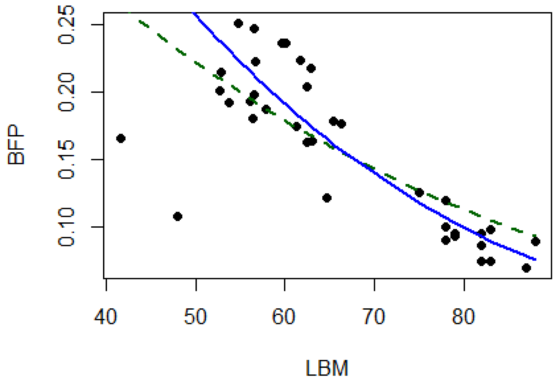


Figure 1: AIS data set: Body fat percentage versus lean body mass

Real data, however, often presents challenges such as outliers or influential values such as the *AIS* data set as presented in Bayes et al. (2012) and illustrated in Figure 1, where the solid line is the regression model obtained using the data without the outlying values and the broken line is the regression model based on the full data. The aim is to model the body fat percentage (BFP) using lean body mass (LBM) as a predictor. In Figure 1 the effect of the outliers on the regression model is evident. The question arises whether the outliers should be included or not.

Espinheira et al. (2008) developed influential analysis to discover values with high leverage that are also influential in parameter estimation under various perturbation schemes. Once the influential values are determined, the choice of manual removal is left to the practitioner. In smaller data sets it is simple to investigate the data and remove outlying or influential values if the practitioner chooses to do so. The process of removing datapoints is not to be encouraged on the basis of contradicting model information, and is a monumentous task for larger data sets or a high dimensional covariate space. This complex decision is further exacerbated by the effect on the significance test of the intercept, as summarized in Table 1 for this data set. It is clear from Table 1 that the intercept is deemed significant in the case where the outliers are not included but insignificant if the outliers are included in the modeling. This is especially of concern since anatomically, the intercept should not be zero. The decision of removing outliers clearly has profound consequences. Also, the decision is made on a subjective conjecture which is troublesome in any scientific study. Rather, a modeling framework that can effectively handle outliers should be developed

Table 1: 95% Confidence intervals from the *betareg* procedure

	Full data	Outliers deleted
β_0	$(-0.408; 0.6043)$	$(0.463; 1.2127)$
β_1	$(-0.035; -0.019)$	$(-0.044; -0.032)$

to circumvent the necessity of such a decision. Ghosh (2017) successfully developed the MDPDE framework for the beta regression model which is robust with respect to outliers. In this method however, there are still some future research needed to determine the optimal tuning parameter and a basic computational framework for ease of implementation by a practitioner.

The aim of this paper is to develop a modeling framework and estimation procedure that is robust to outliers or influential values and computationally inexpensive, with a comprehensible implementation. The assumption of homogeneity of dispersion is not necessary for our proposal. We present the framework as well as a simple efficient implementation procedure using R. Our method is developed for regularization and stabilization of inference, using a Bayesian procedure. The proposed framework with details on implementation is presented in Section 2. A simulation study founding the proof of concept is presented in Section 3 and real data sets are analyzed in Section 4. The paper is concluded with a discussion and some recommendations.

2 Robust Bayesian beta regression model (RBBRM)

The framework we propose will be general in the sense that we do not assume homogeneity of the precision parameter but instead model it using a proper link function and predictors, if needed. We use the reparameterized beta distribution (2) as the response model and employ a location and dispersion (not precision) regression model, similar to (3) and (4). The location sub-model is defined as:

$$\boldsymbol{g}(\mu) = \boldsymbol{X}^\top \boldsymbol{\beta}, \tag{5}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{x}_i \in \mathbb{R}^m$ is the i-th covariate with corresponding regression parameter β_i, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$, and $\boldsymbol{g}(.)$ is an appropriate link function like the *logit* or *probit* function. The *dispersion* sub-model is defined as:

$$\boldsymbol{h}(\phi) = -\boldsymbol{Z}^\top \boldsymbol{\gamma}, \tag{6}$$

$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is the i-th covariate with corresponding regression parameter γ_i, with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$, and $\boldsymbol{h}(.)$ is an appropriate link function like the *exp* function. Note the difference between (4) and (6). The negative sign is incorporated to model the dispersion rather than the precision. Using (6) allows the interpretation of γ_i in the same direction as the dispersion, rather than the precision. A positive γ_i implies that an increase in the i-th predictor results in a higher dispersion parameter. This setup facilitates a more logical interpretation than the model in (4).

2.1 Bayesian specification

In our proposal the estimation space is of dimension $\mathbb{R}^{m+p}$, which might be troublesome in the case of a small sample size. A Bayesian approach is used for the estimation of the parameters to facilitate smaller sample sizes. Additionally, the robustness needed in the case of outliers, is established by the Bayesian viewpoint as proposed in this paper. The Bayesian procedure established by Cepeda-Cuervo et al. (2016) assumes a Gaussian prior for the regression coefficients. The Gaussian distribution is light tailed and very sensitive to outliers. To develop a regularized robust Bayesian procedure we employ a Student's t prior on the regression coefficients for both the location and dispersion submodels, (5) and (6) (see Lange et al. (1989) and Gelman et al. (2003) for more details). Let $t(\mu, \tau, \nu)$ denote the Student's t-distribution with mean μ, scale τ, and degrees of freedom (DF) ν, such that the expected value is μ and the variance is $\frac{\tau\nu}{\nu-2}, \nu > 2$. Then, we assume

$$\beta_j \sim t(\mu_{\beta_j}, \tau_{\beta_j}, \nu_{\beta_j}), \quad j = 0, 1..., m, \tag{7}$$

and

$$\gamma_k \sim t(\mu_{\gamma_k}, \tau_{\gamma_k}, \nu_{\gamma_k}), \quad k = 0, 1..., p. \tag{8}$$

To avoid the issue of overfitting due to prior specification and surrendering to Occam's razor, (Simpson et al., 2015) we will set $\mu_{\beta_j} = 0$ for $j = 0, 1, ..., m$ and $\mu_{\gamma_j} = 0$ for $k = 0, 1, ..., p$, so that the modes of the prior distributions are situated at values that will result in a simpler model. Using this approach ensures that a more complicated (high-dimensional regression) model is chosen above a simpler (constant) model, only if the data necessitates it to be so.

It is well-known that as $\nu_{\beta_j}; \nu_{\gamma_k} \to \infty$, the Student's t distribution approaches the Gaussian distribution. If the Student's t prior is appropriate for the data then the estimated DF should be small. For this purpose we employ a hierarchical framework by assuming a uniform prior (designate $U(2, 100)$) for the DF. If the DF is less than two then the variance is infinite. In practice, there are only marginal differences between the Student's t and Gaussian distributions as the DF increases beyond 30. The scale hyperparameters $\tau_{\beta_j}, j = 0, 1, ..., m$ and $\tau_{\gamma_k}, k = 0, 1, ..., p$ are set to large values (which results in large variances) as to construct weakly informative Student's t priors. Alternatively, a prior could also be envoked on the scale hyperparameters for a full hierarchical specification, which is not within the scope of the current research.

Suppose $y_1, \ldots, y_n$ are n independent responses each taking a value in $(0, 1)$, and are associated with respective m- and p-dimensional covariate sets $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. The full model specification is then

$$\begin{aligned}
& y_i|(\mu_i, \phi_i) \sim \text{Beta}(\mu_i, \phi_i), \quad i = 1, ..., n \\
& g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_m X_{i,m} \\
& \beta_j|\nu_{\beta_j} \sim t(0, \tau_{\beta_j}, \nu_{\beta_j}) \\
& \nu_{\beta_j} \sim U(2, 100) \\
& h(\phi_i) = -\boldsymbol{z}_i^\top \boldsymbol{\gamma} = -\gamma_0 - \gamma_1 Z_{i,1} - \gamma_2 Z_{i,2} - ... - \gamma_p Z_{i,p} \\
& \gamma_k|\nu_{\gamma_k} \sim t(0, \tau_{\gamma_k}, \nu_{\gamma_k}) \\
& \nu_{\gamma_k} \sim U(2, 100)
\end{aligned} \tag{9}$$

The hyperparameters in this specification are $\tau_{\beta_j}, 1, ..., m$ and $\tau_{\gamma_k}, k = 1, ..., p$. These hyperparameters govern the dispersion of the priors of the regression coefficients and will be analysed for their effect on the posterior results in a sensitivity study.

The Laplace and Cauchy distributions can also be used as weakly informative priors but increases the computational cost of the estimation procedure. Since our aim is to provide a robust framework for inference within the beta regression model that is simple in the implementation and use for practitioners, we will focus on the prior specification as set out above.

2.2 Posterior validation

The propriety of the posterior distribution should be established in any Bayesian model since if the posterior is not proper (not a valid probability distribution) the posterior inference based on sampling schemes are not valid. If the priors are all proper and non-degenerate, then the posterior is proper everywhere except on the Lebesgue set (null set). In this paper we use a bounded uniform distribution as a hyperprior for the DF of the Student's t prior on the regression coefficients. This prior specification is proper everywhere ensuring a proper posterior. The propriety of the posterior is easily established as follows.

Let $\boldsymbol{D}_{\text{obs}} = (n, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ denote the observed data. From the full model specification (9), the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ has the form

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{D}_{\text{obs}}) = L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{D}_{\text{obs}})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}), \tag{10}$$

where the likelihood has form

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{D}_{\text{obs}}) \quad = \quad \prod_{i=1}^{n} \frac{y_i^{\mu_i\phi_i - 1}(1 - y_i)^{\phi_i(1-\mu_i)-1}}{B(\mu_i\phi_i, \phi_i(1 - \mu_i))}$$

with $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, $\phi_i = h^{-1}(-\boldsymbol{z}_i^\top \boldsymbol{\gamma})$, $\pi(\boldsymbol{\beta}) = \int \pi(\boldsymbol{\beta}|\boldsymbol{\nu_\beta})\pi(\boldsymbol{\nu_\beta})\mathrm{d}\boldsymbol{\nu_\beta}$ and $\pi(\boldsymbol{\gamma}) = \int \pi(\boldsymbol{\gamma}|\boldsymbol{\nu_\gamma})\pi(\boldsymbol{\nu_\gamma})\mathrm{d}\boldsymbol{\nu_\gamma}$.

As an example, if the logit and log link functions are used for the location and dispersion sub-models, respectively, the likelihood has the form

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{D}_{\text{obs}}) \quad = \quad \prod_{i=1}^{n} \frac{y_i^{\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\exp\left(-\boldsymbol{z}_i^\top\boldsymbol{\gamma}\right)-1}(1-y_i)^{\exp\left(-\boldsymbol{z}_i^\top\boldsymbol{\gamma}\right)\left(1-\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\right)-1}}{B\left(\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\exp(-\boldsymbol{z}_i^\top\boldsymbol{\gamma}),\exp(-\boldsymbol{z}_i^\top\boldsymbol{\gamma})\left(1-\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\right)\right)}.$$

The posterior distribution of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ is proper if and only if

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^m} L(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{D}_{\text{obs}})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\gamma} < \infty.$$

Since the priors are proper and due to the fact that the Student's t distribution can be represented as a scale mixture of normal's with the weight function as the inverse gamma distribution, the propriety of the posterior distribution follows from Figueroa-Zuniga et al. (2013)

2.3 Software implementation

The Bayesian analysis is done using simulated samples from the stationary posterior distribution of the parameters (10), using MCMC sampling schemes. Establishing convergence and posterior inference is done based on these simulated samples. The *coda* package in R(Plummer et al., 2006) provides trace plots, convergence plots and diagnostics like the Gelman(Gelman et al., 2003) or Geweke diagnostics, highest posterior density intervals, credible intervals and posterior summaries. The code provided in the supplementary material uses *coda* to investigate the convergence(Gelman et al., 2003) and autocorrelation of the simulated samples, as well as to calculate the estimates and credible intervals for the parameters. The data analysis in this paper has been done using JAGS(Plummer, 2003) and R.

3 Simulation study

In this section simulated data is used to illustrate the proposed framework and as proof of concept. The sensitivity of the posterior distribution to the hyperparameter values is also investigated in this section to establish the robustness of the posterior results.

We start by simulating a data set with **no outliers** based on a predictor set of dimension 1 for the location model with a fixed $\phi = 155$.
As an example the logit link for $g(.)$ is used and constant dispersion (identity link for $h(.)$ with an intercept only) is assumed in this section. In general, any appropriate link function will suffice.

3.1 No outliers

A predictor set was simulated from a Gaussian sample and the regression coefficients were set to $\beta_0 = 0.5$ and $\beta_1 = 1$ respectively.

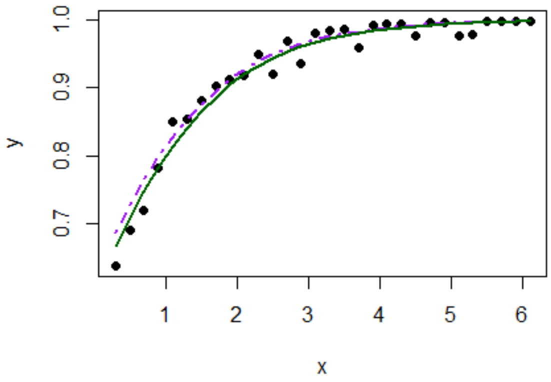


Figure 2: Simulated data: MLE (solid line) and RB-BRM (dashed line)

The *betareg* package was used to fit a regression model to the data (broken line) as well as the new method (solid line) with the lines presented together with the data in Figure 2. The estimated regression equations are

$$\begin{aligned}&\text{logit}_{\text{RB}-\text{BRM}}(\mu) = 0.4947(0.4308; 0.5742) + 0.9796(0.9073; 1.0582)x\\&\phi = 150.08(86.2367; 228.4974)\end{aligned} \tag{11}$$

and

$$\begin{aligned}&\text{logit}_{\text{MLE}}(\mu) = 0.4687(0.2878; 0.6496) + 0.97116(0.8735; 1.0689)x\\&\phi = 176.07(83.15; 268.99).\end{aligned} \tag{12}$$

From (11) and (12) it is clear that both methods elicit the true regression structure, $logit(\mu) = 0.5+x, \phi = 155$. The pseudo R^2,(Efron, 1978) for the MLE is 99.2% and for the RB-BRM is 99.35 %. The estimated DF is $\nu_{\beta_j} = 28, j = 0, 1$. These large values indicate the good performance of the MLE's or a Bayesian set-up with a Gaussian prior. It is evident that in the case of a well-behaved data set, as in this case, the two methods perform competitively well. In the next section the performance in the presence of gross outliers are investigated.

3.2 Contaminated data under different perturbation schemes

Now the simulated data set is manually contaminated with outliers according to four perturbation schemes as in Bayes et al. (Bayes et al., 2012) The perturbation schemes are described as:

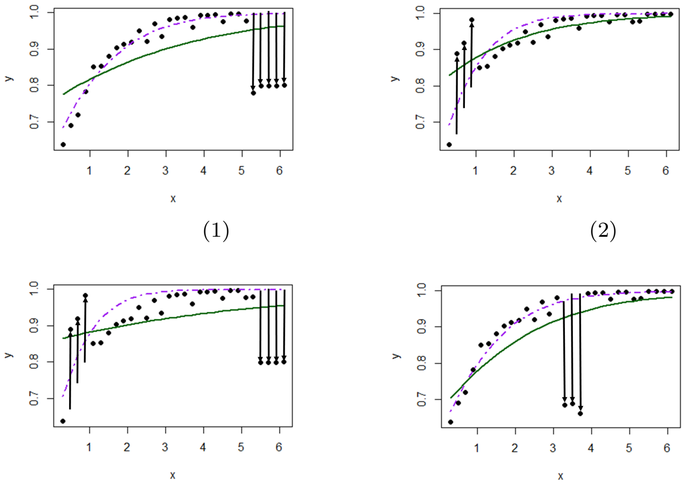
1. Decreasing the response by δ units for larger values of the predictor.

2. Increasing the response by δ units for smaller values of the predictor.

3. Combining schemes 1 and 2 i.e. decreasing the response for larger values of the predictor and increasing the response for smaller values of the predictor by δ units.

4. Decreasing the response by δ units for middle values of the predictor.

The results are illustrated in Figure 3. It is clear that the MLE is very much influenced by outliers and the outliers influence the estimation method in such a way that the regression curve is not representative of the data anymore. The robust estimation, on the other hand, is not drastically influenced by the outliers, as the MLE's are, and provide more robust results. Keeping in mind that the true regression structure is $\beta_0 = 0.5, \beta_1 = 1$ we investigate the resulting credible and confidence intervals under the two estimation procedures, the RB-BRM and MLE. This is illustrated in Figure 4. Note that perturbation scheme 0 represents the simulate data with no outliers. It is evident that the estimates resulting from MLE does not contain the true underlying regression structure for most of the perturbed schemes. This is not true for the estimates under the RB-BRM procedure. The resulting estimates are stable and the intervals contain the true value in each case. It is important to note that the same hyperparameters, to ensure vagueness, were used for all scenario's as $\tau_{\beta_j} = 1000, j = 0, 1$. In the case of the perturbed data the estimated DF is $\nu_{\beta_j} = 2.578, 3.12, 7.68, 7.12, j = 0, 1$ for perturbation schemes 1,2,3 and 4, respectively. These values motivate the use of a Student's t prior since the distribution of the regression parameters is heavy-tailed due to the outliers in each case.



(1)

(2)

(3)

(4)

Figure 3: Perturbed data: MLE (solid line) and RB-BRM (dashed line)

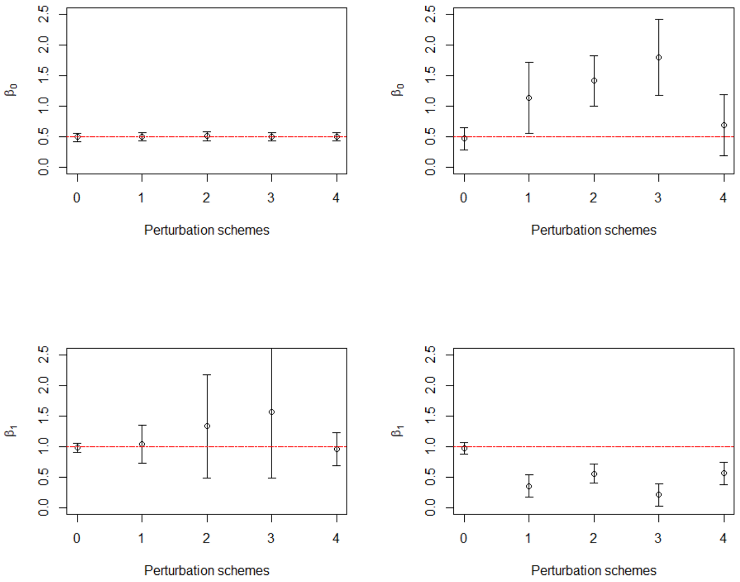


Figure 4: Simulated (0) and Perturbed (1-4) data: Credible (left) and confidence (right) intervals for the parameters under the RB-BRM (left) and MLE (right)

3.3 Sensitivity analysis

In this section the sensitivity of the posterior results are investigated when the prior parameters assume different values. (Robert, 2007) Assessment of the robustness of the prior distribution, can be done trough the fact that the full model specification has a hierarchical structure as in (9). As pointed out by Robert, (Robert, 2007) the main purpose of the hierarchical extension was actually to avoid the restrictive framework of conjugate priors. The sensitivity analysis for the proposed framework is done by investigating the posterior results based on changes in the the hyperparameter values. The simulated data set presented in Section 3.1, is used as the data. Recall that the true underlying values of the regression coefficients are $\beta_0 = 0.5$ and $\beta_1 = 1$.

Firstly, the scale hyperparameters $\tau_{\beta_j}, j = 0, 1$ are fixed at one of eight values (1-3000) and the resulting credible intervals for β_0 and β_1 are displayed in Figure 5. The credible intervals presented in Figure 5 contains the true underlying value for both the parameters β_0 and β_1, in most cases. This illustrates the robust inference based in the RB-BRM framework proposed in this paper. The null hypothesis of the true values is rejected for very small values of the scale hyperparameters, since the prior is centered at zero. This result is expected,

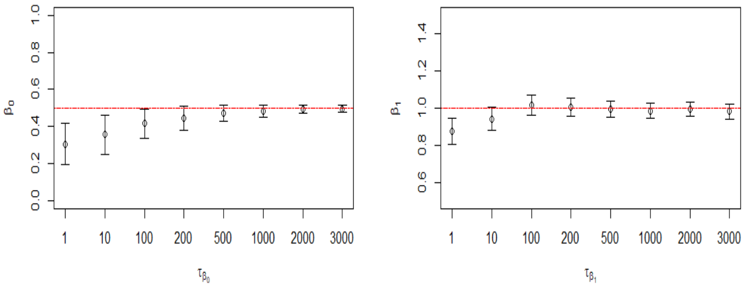


Figure 5: Credible intervals for the parameters under the RB-BRM procedure for different values of $\tau_{\beta_j}, j = 0, 1$

since the resulting prior is extremely informative since it is centered at zero with very small variance. For larger values of $\tau_{\beta_j}, j = 0, 1$ the null hypothesis is not rejected, preserving the true underlying regression model. Although, from a practical viewpoint, large values of $\tau_{\beta_j}, j = 0, 1$ is feasible to construct weakly informative priors.

Secondly, various upper bounds of the uniform prior for the degrees of freedom were investigated and the procedure is very robust to this deviation. The upper bound has very little to no effect on the regression coefficients. To ensure propriety of the posterior a finite upper bound is needed. We advise that the upper bound should exceed 30 as to deter from a forced heavy-tailed prior on the regression coefficients, since it might not be justified in all cases.

4 Applications

4.1 AIS data set

In the first application we return our attention to the *AIS* data set introduced in Section 1. The *AIS* data set available in the R package *sn* contains health measurements of several athletes collected at the Australian Institute of Sport (AIS). A subset of this data set, the data on rowing athletes, has been studied by Bayes et al. (Bayes et al., 2012) since there are outliers present in this subgroup. The aim is to use lean body mass (LBM) as a predictor of body fat percentage (BFP), which is a value between 0% and 100%, theoretically. There are two outlying values present in the subset of rowers. In the introduction of this paper, the impact of the outliers have been highlighted. Now we reanalyze the data set using the proposed RB-BRM framework. For comparability we assume constant dispersion and a location sub-model based on the predictor

LBM as follows:

$$\begin{aligned}&\mathrm{BFP_i} \sim \mathrm{Beta}(\mu_\mathrm{i}, \phi)\\&\mathrm{logit}(\mu_i) = \beta_0 + \beta_1 \mathrm{LBM_i}\end{aligned}$$

The model was fitted using the new proposed RB-BRM framework and also the MLE's using the *betareg* package on the full data set and also on the data set where the outliers were removed. The values of the hyperparameters were assumed to be $\tau_{\beta_0} = \tau_{\beta_1} = \tau_{\gamma_0} = 1000$. The estimated equations (with credible and confidence intervals) are as follows for the full data set:

$$\begin{aligned}&\mathrm{BFP_{i,MLE}} \sim \mathrm{Beta}(\mu_\mathrm{i}, 96.62(51.76; 141.48))\\&\mathrm{logit_{MLE}}(\mu_i) = 0.098(-0.408; 0.6043) - 0.027(-0.035; -0.019)\mathrm{LBM_i}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{BFP_{i,RB-BRM}} \sim \mathrm{Beta}(\mu_\mathrm{i}, 78.623(46.84; 118.77))\\&\mathrm{logit_{RB-BRM}}(\mu_i) = 0.828(0.7601; 0.8932) - 0.03848(-0.0405; -0.036)\mathrm{LBM_i},\\&\widehat{\nu_{\beta_0}} = 12.78, \widehat{\nu_{\beta_1}} = 24.05,\end{aligned}$$

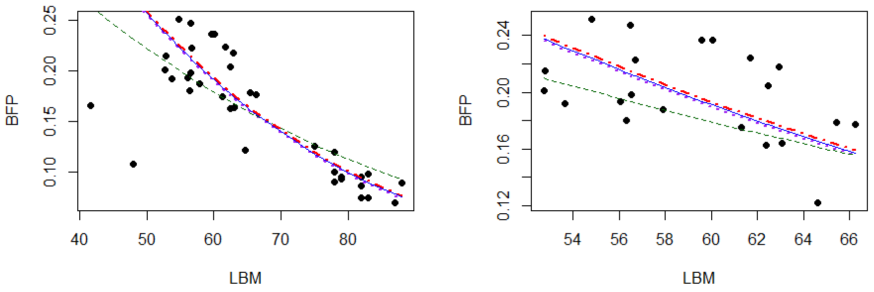
and for the data set without outliers:

$$\begin{aligned}&\mathrm{BFP_{i,MLE}} \sim \mathrm{Beta}(\mu_\mathrm{i}, 246.31(128.61; 363.81))\\&\mathrm{logit_{MLE}}(\mu_i) = 0.8377(0.463; 1.2127) - 0.038(-0.044; -0.032)\mathrm{LBM_i}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{BFP_{i,RB-BRM}} \sim \mathrm{Beta}(\mu_\mathrm{i}, 79.147(47.832; 118.644))\\&\mathrm{logit_{RB-BRM}}(\mu_i) = 0.828(0.763; 0.889) - 0.0385(-0.0404; -0.0365)\mathrm{LBM_i}\\&\widehat{\nu_{\beta_0}} = 51.27, \widehat{\nu_{\beta_1}} = 50.56.\end{aligned}$$

These results compare well with available results.(Bayes et al., 2012; Smithson and Verkuilen, 2006) The difference in the RB-BRM with regards to the inclusion or exclusion of the outliers, is negligible. This is not the case for the model based on the MLE's which emphasize the necessity of our proposal. The estimated DF of the priors indicate the ability of the model to effectively handle the outliers. When the outliers are included, the estimated DF is small, indicate the heavy tail behaviour of the regression coefficients. In the case where the outliers are removed, the RB-BRM proposal that utilizes Student's t priors, reduces to the Gaussian priors due to the large estimated DF. The fitted results are illustrated in Figure 6. It is evident from Figure 6 that the regression based on the MLE's are very sensitive to the inclusion of the outliers. This is especially clear in Figure 6 (b) where the close-up reveals how similar the results are for the RB-BRM irrespective of the inclusion of the outliers. This illustrates the robustness of the new RB-BRM proposal to outliers, in contrast to the sensitivity of the MLE's.



(a) Full data set

(b) Close-up of $50 < LBM < 70$

Figure 6: AIS data: MLE (full data set - dashed line and without outliers - solid line) and RB-BRM (full data set - dotted line and without outliers - dot-dashed line).

4.2 Psychology data set

A psychological study based on 166 nonclinical women was conducted in Australia to investigate the effects of stress, anxiety and depression. More details can be found in Smithson and Verkuilen(Smithson and Verkuilen, 2006) where it is pointed out that this data set presents with heteroscedasticity. However, Ghosh (2017) analyzed this data set under the assumption of homogeneous dispersion. Smithson and Verkuilen(Smithson and Verkuilen, 2006) noted that there are a number of inluential observations which might influence the estimates. In this section we will analyze the data by assuming firstly homogeneous dispersion (see Figure 7 (a)), and then secondly heterogeneous dispersion (see Figure 7 (b)). Also, the analysis was conducted using the full data set and then by discarding some influential values.

The response is the anxiety score, and the stress score is used as the predictor for the location as well as the dispersion model. It is clear from Figure 7 that the model that assumes heterogeneous dispersion is more appropriate. A beta regression model is fitted to the data using the MLE and the RB-BRM, respectively, which yielded the following results for the full data set,

$$\begin{aligned}
\text{logit}_{\text{RB-BRM}}(\mu_i) &= -3.854(-4.161; -3.526) + 4.473(3.589; 5.389)\text{Stress}_\text{i}, \\
\widehat{\nu_{\beta_0}} &= 2.472, \quad \widehat{\nu_{\beta_1}} = 2.449, \\
\log_{\text{RB-BRM}}(\phi_i) &= 3.637(3.058; 4.157) - 3.396(-4.93; -2.851)\text{Stress}_\text{i}, \\
\widehat{\nu_{\gamma_0}} &= 2.506, \quad \widehat{\nu_{\gamma_1}} = 2.398
\end{aligned}$$

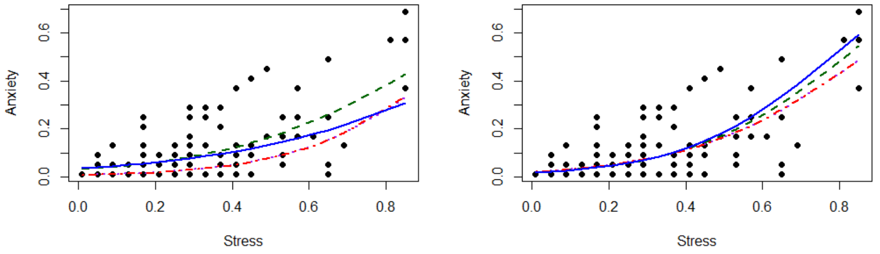
and

$$\begin{aligned}
\text{logit}_{\text{MLE}}(\mu_i) &= -4.0237(-4.2881;-3.7593) + 4.9414(4.1388;5.744)\text{Stress}_\text{i}, \\
\log_{\text{MLE}}(\phi_i) &= 3.9608(3.5698;4.3518) - 4.2733(-5.3389;-3.1861)\text{Stress}_\text{i},
\end{aligned}$$

and for the data set without outliers,

$$\begin{aligned}
\text{logit}_{\text{MLE}}(\mu) &= -4.111(-4.3804;-3.8416) + 5.2799(4.3651;6.1947)\text{Stress}_\text{i}, \\
\log_{\text{MLE}}(\phi) &= 4.2012(3.7894;4.613) - 5.2515(-6.4959;-4.0071)\text{Stress}_\text{i}
\end{aligned}$$

The estimated DF again indicated the appropriateness of the heavy tailed prior when the data includes outliers. The model based on the RB-BRM is not presented for the data set without outliers since the difference to the model based on the full data set is marginal. From the estimates, $\widehat{\gamma_1} = 3.396$ which implies that increased stress leads to increased dispersion in the anxiety for a specific individual due to our dispersion model specification in (6). In Figure 7, however, all four models are illustrated, for both the homogeneous (a) and heterogeneous (b) dispersion assumption. It is clear from Figure 7 that the assumption of homogeneous dispersion is not a valid assumption as noted by Smithson and Verkuilen (2006). In Figure 7 the unstable behaviour of the MLE is again clear. The robustness of the new framework is evident under the assumption of homogenous as well as heterogeneous dispersion, since the estimated regression models fitted with and without outliers, virtually overlap.



(a) Homogenous dispersion

(b) Heterogeneous dispersion

Figure 7: Anxiety and Stress data: MLE (full data set - dashed line and without outliers - solid line) and RB-BRM (full data set - dotted line and without outliers - dot-dashed line)

5 Discussion

In this paper we addressed the need for an easily implementable estimation procedure within the context of varying dispersion beta regression. The approach comprises of a hierarchical Bayesian set-up using weakly informative heavy-tailed Student's t priors for the regression coefficients and a proper uniform prior for the degrees of freedom. This proposal facilitates a comprehensible implementation that supports the model definition and practical application of the framework. The restrictive assumption of homogeneity of the dispersion or precision is not necessary for out proposed method.

The use of proper priors ensure the posterior propriety which makes for an attractive alternative solution to the challenge of outliers within the beta regression framework. A simulation study using different perturbation patterns thoroughly investigated the robustness of the proposed model in the presence of various types of outliers. The robustness of the posterior to the choice of the hyperparameter values, were illustrated in a sensitivity analysis.

The robust Bayesian variable dispersion beta regression model (RB-BRM), proposed in this paper, was applied to the *AIS* data set and the psychology data set with success. In the *AIS* data set, the outliers are clear to the naked eye since the focus is on one predictor. In a case such as this, visual inspection for outliers is feasible. This is, however, not the case in the second data set we analyzed, the psychology data set, since there are two predictors. A predictor space of non-unit dimension poses a challenge to visual inspection for prominent outliers. For a high dimension, visualization is a near impossible task. The need for the robust model is clear in the psychology data set, as well as the violation of the homogeneous dispersion assumption.

The applicability and necessity for this framework is evident from the simulation study as well as the analysis of the real data sets. This framework is an easily implementable tool for use by practitioners in the field of medical research where the response is measured in any finite interval. The proposed framework can be extended to facilitate random effects and non-linear behaviour in future. This extension can be easily incorporated into the computational procedure presented with this paper.

Acknowledgements

The authors would like to hereby acknowledge the support of the StatDisT group.

Funding

This work is based upon research supported by the National Research Foundation (Re:CPRR3090132066 No 91497 & Re:IFR170227223754 No 109214) and the National Teaching Development Grant No 11017.

References

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

Kieschnick R and McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions, *Statistical Modelling* 2001, 3(3): 193-213.

Ferrari S and Cribari-Neto F. Beta regression for modelling rates and proportions, *Journal of Applied Statistics* 2004, 31(7): 799-815.

Albert JM, Wang W and Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries, *Statistical Methods in Medical Research* 2014, 23(3): 257-278.

Liu F and Li Q. A Bayesian model for joint analysis of multivariate repeated measures and time to event data in crossover trials. *Statistical Methods in Medical Research* 2016, 25(5): 2180-2192.

Wang J and Luo S. Bayesian multivariate augmented Beta rectangular regression models for patient-reported outcomes and survival data. *Statistical Methods in Medical Research* 2017; 26(4): 1684-1699.

Kim G and Lee Y. Marginal versus conditional beta-binomial regression models, *Statistical Methods in Medical Research* 2017, DOI:10.1177/0962280217735703.

Liu F and Eugenio EC. A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research* 2018, 27(4): 1024-1044.

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

Cepeda-Cuervo E, Jaimes D, Marin M and Rojas J. Bayesian beta regression with Bayesianbetareg R-package, *Computational Statistics* 2016, 31(1): 165-187.

McNeish D. On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal* 2016, 23(5): 750-773.

Bayes CL, Bazán JL and García C. A new robust regression model for proportions, *Bayesian Analysis* 2012, 7(4): 841-866.

Espinheira PL, Ferrari SL and Cribari-Neto F. Influence diagnostics in beta regression, *Computational Statistics & Data Analysis* 2008, 52(9): 4417-4431.

Ghosh A. Robust inference under the beta regression model with application to health care studies, *Statistical Methods in Medical Research*. Epub ahead of print 27 November 2017, DOI:10.1177/0962280217738142.

Lange K, Little R and Taylor J. Robust Statistical Modeling Using the t Distribution *Journal of the American Statistical Association* 1989, 84(408): 881.

Gelman A, Carlin J, Stern H and Rubin D. *Bayesian Data Analysis*, Second Edition. London: Chapman & Hall, 2003.

Simpson DP, Rue H, Martins TG, et al. Penalising model component complexity: a principled, practical approach to constructing priors 2015; arXiv:1403.4630v4.

Figueroa-Zuniga JI, Arellano-Valle RB and Ferrari SLP. Mixed beta regression: A Bayesian perspective, *Computational Statistics & Data Analysis* 2013, 61: 137-147.

Plummer M, Best N, Cowles K and Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R news* 2006, 6(1): 7-11.

Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*, 2003, 124(125.10).

Efron B. Regression and ANOVA with zero-one data: Measures of residual variation, *Journal of the American Statistical Association* 1978, 73: 113-121.

Robert CP. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. Springer-USA, 2007.

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