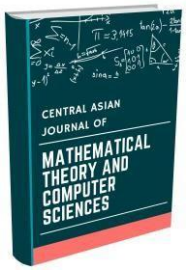

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS
Volume: 06 Issue: 03 | July 2025 ISSN: 2660-5309

Article

Reducing The Influence of High Leverage Points in Beta Regression Using The Gm6 Robust Estimator

Hamza Lateef Katea Al-Ayashy[1], Taha Alshaybawee[2]

1. Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al-Qadisiyah, Iraq
2. Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al-Qadisiyah, Iraq

* Correspondence: Hamzakufa1986@gmail.com, taha.alshaybawee@qu.edu.iq

Abstract: The problem of outliers and high leverage points is one of the most prominent challenges facing the design of statistical models, especially in regression models, as they have a significant impact on distorting the results of statistical estimation. This research aims to address the impact of high leverage points in a beta regression model using robustness estimation methods, specifically the GM6 multistage estimator (GM6-BR). A comparison of four estimation methods was studied: traditional estimation (BR), least discrete squares estimator (LTS-BR), generalized estimation (GM-BR), and GM6-BR. A simulation study was conducted on two models: one linear and the other nonlinear, with data contamination of 10% and 20% introduced to test the robustness of the different methods. The results demonstrated a clear superiority of the GM6-BR method in terms of reducing error (RMSE, MAE) and skewness (BIAS), while maintaining stability in the presence of contamination. Practical application on real data also showed that GM6-BR was least affected by outliers compared to other methods, while the traditional BR method exhibited the highest level of distortion. Therefore, the study recommends adopting GM6-BR as an effective and accurate option for analyzing data containing high leverage points or outliers, especially within beta regression models.

Keywords: Beta regression, GM6-BR-resistant estimation, high leverage points, outliers

Citation: Al-Ayashi, H. L. K & Alshaybawee, T. Reducing The Influence of High Leverage Points in Beta Regression Using The Gm6 Robust Estimator. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(3), 614-631.

Received: 03th Mar 2025
Revised: 11th Apr 2025
Accepted: 24th May 2025
Published: 17th Jun 2025

Copyright: © 2025 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

The issue of outliers is one of the oldest problems in statistics, as the presence of these values affects statistical analysis results and can lead to misleading conclusions. Classified extreme values into three types: vertical extreme values, which appear in the response variable (y) and regression residuals; high leverage points (HLPoints), which appear in the independent variables (x); and influential observations, which affect the relationship between the independent variables and the response variable [1], [2].

To address this issue, new statistical techniques have been developed that are resistant to the influence of outliers, known as robust methods. These include M-estimators, MM-estimators, the standard deviation estimator (S), the Least Trimmed Squares (LTS) estimator, the Least Median of Squares (LMS) estimator, and others. However, these estimators tend to weaken in the presence of high leverage points. To overcome this limitation, Schweppe, as described by , proposed a powerful method known as GM. Over time, multistage GM estimators have been developed into several types, the most common being the GM6 multistage estimator introduced by Coakley and Hettmansperger [3], [4], [5].

When the data are in the form of fractions or percentages and are constrained within the interval (0,1), the beta regression model is used to estimate the parameters. Beta regression is a widely used statistical model, applied in various fields such as economics, medical sciences, unemployment rates, family income ratios, and more. Maximum Likelihood Estimation (MLE) is commonly employed to estimate the parameters of the beta regression model. Introduced the beta regression model, and numerous studies have explored different aspects of this model [6].

For instance, examined the behavior of maximum likelihood estimation in a beta regression model where the distribution parameters are nonlinear functions of linear combinations of explanatory variables with unknown coefficients. conducted a study that focused on point and interval optimization of the beta regression model. presented a paper describing the BETAREG package, which provides a beta regression class in the R statistical computing system. investigated separation measures in beta regression models, while) presented a study titled "Model Selection Criteria in Beta Regression with Variable Dispersion."

Additionally, conducted a study on the development of a robust inference procedure for beta regression models. proposed Liu shrinkage estimators for beta regression models, while presented a study on Bayesian empirical regression for limited responses with unknown support. introduced the Dawood-Cabria estimator for beta regression models, whereas Abu studied beta ridge regression estimators. Furthermore, proposed a robust semi-parametric inference approach for two-stage production models using beta regression. conducted a comparative study on robust estimation in beta regression models in the presence of outliers. examined robust estimation in beta regression using the Lq maximum likelihood meth od. applied robust beta regression using the logit transformation, and Heng and Lang conducted a study involving an initial estimation of the proportion of outliers in robust regression [7], [8].

In this study, we propose a GM6 estimation approach for beta regression models with responses in the (0,1) interval, aiming to mitigate the influence of high leverage points in the independent variables [9]. Specifically, we employ the multistage Generalized M-estimator of sixth order (GM6) to robustly estimate the model parameters. To facilitate this, we first develop a Least Trimmed Squares (LTS) estimation procedure tailored for beta regression, which serves as a robust initial estimate for the GM6 procedure. Additionally, we incorporate the robust Mahalanobis Distance method to detect high leverage observations in the covariates, ensuring a more reliable and robust estimation process.

2. Materials and Methods

1. Beta distribution

Beta regression is used when the data is within the interval (0,1) and was first expressed by Ferrari and Cribari-Neto by relating the mean function of the response variable to a set of linear predictors via a monotonic differentiable function called the correlation function **.**

Let *y* be a continuous random variable having a beta distribution with the probability density function y as follows:

$$f(y;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1} \qquad 0<y<1\ ,a>0,b>0 \tag{1}$$

Where $\Gamma(.)$ is the gamma function , a and b are two shape parameters . The mean and variance of y are as follows:

$$\text{E (y)} = \frac{a}{a+b} \quad , \text{Var} = \frac{ab}{(a+b)^2(a+b+1)}$$

To obtain a regression structure for the mean response and accuracy parameter, different beta density parameters are obtained.

Let $\mu = \frac{a}{a+b}$ and $\lambda = a + b$, $a = \mu\lambda$, and $b = (1-\mu)\lambda$, In the new parameterization, the density of y can thus be expressed as follows:

$$f(y;\mu,\lambda)=\frac{\Gamma(\lambda)}{\Gamma(\mu\lambda)\Gamma((1-\mu)\lambda}y^{\mu\lambda-1}\ (1-y)^{(1-\mu)\lambda-1} \tag{2}$$

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

$$E\ (y|\mu, \lambda) = \mu, \qquad \text{and} \qquad Var\ (y|\mu, \lambda) = \frac{V(\mu)}{1+\lambda}$$

Where $V(\mu) = \mu(1-\mu)$, μ is the average and λ can be interpreted as an accuracy parameter**.** Beta regression model was developed assuming a homogeneous accuracy parameter in the form of a generalized linear model for the location parameter using a correlation function . Let $y_2, y_2, \ldots, y_n$ be independent random variables, where each y , t = 1,2, ... , n follows the density as shown in Equation (2) with mean μ and unknown precision λ. The beta regression model is obtained by assuming that $y \sim \beta eta\ (\mu, \lambda)$, t = 1, 2, . . . and n, and the logarithmic link function can be written as follows:

$$g\ (\mu) = log(\frac{\mu_t}{1-\mu_i}) = \eta_i = \sum_{i=1}^{n} x'_{ij}\beta_j \qquad , j = 1,2, \ldots, k \tag{3}$$

Where x_t is a $(k \times 1)$ vector of predictors and $\beta' = (\beta_1, \ldots \ldots, \beta_n)$ is a $(k \times 1)$vector of unknown regression parameters. Moreover ,we assume that the link function g(.) : $[0,1] \rightarrow R$and there exist several different link functions that map the linear predictor onto the space $[0,1]$such as:

Logit $\quad g\big(\mu(x_i)\big) = ln\left(\frac{\mu(x_i)}{1-\mu(x_i)}\right) = \eta_t$

where ln (.) is the natural logarithm and is the standard normal cumulative distribution function .Hence, the beta regression model assumes that the mean of the dependent variable can be represented in the following form:

$$g\big(\mu(x_i)\big) = \eta_i = x'_i\beta$$

When using the logit link function, the beta regression model is obtained by assuming that the conditional mean of y_t can be written as

$$\mu(x_i) = \frac{e^{x'_i\beta}}{1+e^{x'_i\beta}} \tag{4}$$

Estimation of the beta regression parameters is done by using the ML method Espinheira et al., (2008) The log-likelihood function of the beta regression model is given by:

$$L(\mu_i, \lambda; y_i) = \sum_{i=1}^{n}\{log\Gamma(\lambda) - log\Gamma(\mu_i(\lambda) - log\Gamma\big((1-\mu_i)(\lambda)\big) + (\mu_i(\lambda) - 1)\log(y_i) + \big((1-\mu_t)(\lambda) - 1\big)\log(1-y_t)\}$$

(5)

Differentiating the log-likelihood in Eq. (4) with respect to β gives us the score function for β which is given by:

$$U(\beta) = \lambda x'\ F(y^* - \mu^*) \tag{6}$$

Where $\quad F = diag\left(\frac{1}{g'(\mu_i)}, \ldots\ldots, \frac{1}{g'(\mu_n)}\right), \qquad y^* = (y_1^*, \ldots \ldots, y_n^*)'$ and $\quad y_i^* = \log(\frac{y_i}{1-y_i})$

$\mu^* = (\mu_1^*, \ldots \ldots, \mu_n^*)'$ and $\quad \mu_i^* = \varphi(\mu_t\lambda) - \varphi((1-\mu_t)\lambda)$ such that $\varphi(\cdot)$ denoting the digamma function. The iterative reweighted least-squares (IWLS) algorithm or Fisher scoring algorithm used for estimating β Espinheira et al., (2015) .The form of this algorithm can be written as :

$$\beta^{(r-1)} = \beta^{(r)} + (I_{\beta\beta}^{(r)})^{-1}\ U_\beta^{(r)}\ (\beta)$$

Where $U_\beta^{(r)}$ is the score function defined in Eq. (6), and $I_{\beta\beta}^{(r)}$is the information matrix for β , see for more details. The initial value of β can be obtained by the least squares method, while the initial value for each precision parameter equals

$$\hat{\lambda}_i = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\sigma}_i^2} \tag{7}$$

Where $\hat{\mu}$ and σ_i^2 values are obtained from linear regression .Given $r = 0,12, \ldots..$ is the number of iterations that are performed, convergence occurs when the difference between successive estimates becomes smaller than a given small constant .At the final step, the ML estimator of β is obtained as:

$$\hat{\beta}_{GM} = (x'\widehat{w}x)^{-1}x'\,\widehat{w}\hat{z} \quad (8)$$

Where X is an $n \times p$ matrix of regressors, $\hat{z} = \hat{\eta} + \widehat{w}^{-1}\hat{F}\,(y^* - \mu^*)$, and $\widehat{w} = diag(\widehat{w}_1, \dots \dots \widehat{w}_n)$

$$\widehat{w}_i = \left(\frac{1-\hat{\sigma}^2)}{\hat{\sigma}^2}\left\{\hat{\varphi}\left(\frac{\hat{\mu}_i(1-\hat{\sigma}^2)}{\hat{\sigma}^2}\right) + \varphi'\left(\frac{(1-\hat{\mu}_i)(1-\hat{\sigma}^2)}{\hat{\sigma}^2}\right)\right\}\frac{1}{\{g'(\hat{\mu}_t)\}^2}\right.$$

(9)

2. Generalized M-Beta Regression (GM-BR).

This method relies on the use of the (Generalized M-Estimator-GM) within a beta regression framework to obtain more robust and stable estimates in the presence of high leverage points or outliers. The GM-BR estimate is obtained through several steps. First, initial estimates are obtained using a traditional beta regression model. Then, the residuals and their appropriate scale are calculated. The residuals are then standardized to calculate the standardized values. Next, weights are assigned using the Hat matrix to detect high leverage points. A weight function (such as the Huber function) is then applied based on the standardized residuals. The parameters are then re-estimated using the new weights to reduce the influence of outliers. Finally, the steps are repeated until a steady state is reached.

3. GM6 Estimator

To address high leverage points (HLPoints), Schweppe, as described by Hill and Paul, proposed a powerful method known as the Generalized Limited-Impact M-Estimator (GM-Estimator). However, the GM1 estimator has limited-impact properties, with an efficiency of 95% and asymptotic distribution properties similar to the M-estimator. Its breakdown point does not exceed (1/p), meaning that the breakdown point is inversely proportional to the number of independent variables. Consequently, as dimensionality increases, the breakdown point approaches zero.

The strategies used to reduce the influence of high leverage points in the X-direction are not very effective, as these points may not easily appear in the corresponding diagonal elements *hii* when there are many leverage points. Therefore, multistage GM estimators have been developed into several types, with GM6, introduced by Coakley and Hettmansperger, being the most common. This estimator has high efficiency in normal distributions, limited-impact properties, and a high stopping point. It can be expressed as a solution to the normal equations given by

$$\sum_{i=1}^{n} k_i \varphi\left(\frac{y_{i-x_i'\hat{\beta}}}{si}\right)x_i = 0 \quad (10)$$

$$\hat{\beta} = (x^T wx)^{-1}x^T wy \quad (11)$$

The general procedure for GM6 is by choosing a good initial estimator such as LTS and applying many stages to achieve desirable properties. The initial weights of GM6-estimators that minimize the impact of leverage points in (10) are computed using the RMD values based on MCD or MVE, which are as follows.

$$k_i = min\left[1, \left(\frac{x^2_{(0.95,p)}}{RMD^2}\right)\right] \quad , \quad t=1,2,\dots\dots,n$$

THE TWO ROBUST PARAMETERS OF LOCATION AND SCALE ARE USED AS A SUBSTITUTE FOR THE ARITHMETIC MEAN AND VARIANCE. ACCORDINGLY, THE ROBUST MAHALANOBIS DISTANCE IS WRITTEN AS FOLLOWS:

$$RMD_i = \sqrt{(x_i - \bar{x}_{Rob})'\mathrm{C}(X)^{-1}_{Rob}(x_i - \bar{x}_{Rob})} \quad (12)$$

$$C(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})'(x_t - \bar{x}) \quad , \text{ Rob: means Robust}$$

So for the row t that corresponds to the value of Mahalanobis after exceeding the critical value of chi-square with a degree of freedom of k where k represents the number of independent variables and a significance level of 0.95, then the row contains the outliner:

$$RMD_i > \sqrt{x^2(k, 0.95)}$$

The same procedure is used to identify rows containing outliers by comparing the critical value$x^2(k, 0.95)$ with the robust Mahalanobis (RMD_i) distance.

4. Algorithm.

GM6 method is one of the important methods for treating outliers and leverage points as it can be used to reduce the effect of leverage points in independent variables (x) when estimating the parameters of beta regression model An algorithm for GM6-estimator can be written in the following steps:

Step 1: Choosing an initial estimates $\boldsymbol{\beta}^{(0)}$ from LTS beta regression, to get a high breakdown of 50%. LTS estimator for beta regression can be construct as follows:

a. Choosing the trimming parameter h , where $h = (n + p + 1)/2.$
b. Compute the residuals for each candidate β as $r_i = y_i - \mu_i.$
c. Sort the squares residuals according order as $r^2_{(1)} \leq r^2_{(2)} \leq \cdots \leq r^2_{(n)}.$
d. Compute the LTS objective function $LTS(\beta) = \sum_{i=1}^{h} r^2_{(i)}.$
e. Using robust algorithm (FAST-LTS (Rousseeuw and Driessen 1999)) for computation.

Step 2: For each iteration *t* compute $\mu_i^{(t-1)} = logistic\left(x_i\boldsymbol{\beta}^{(t-1)}\right) = \frac{exp(x_i\beta^{(t-1)})}{1+exp(x_i\beta^{(t-1)})}$.

Step 3: Compute the residuals $r_i^{(t-1)} = y_i - \mu_i^{(t-1)}$ and scale $\hat{\sigma}^{(t-1)} = 1.4826 * median\ of$ $largest(n-p)\ of the\ \left|r_i^{(t-1)}\right|$. and then compute the standardized residuals $u_i^{(t-1)} = \frac{r_i^{(t-1)}}{\hat{\sigma}^{(t-1)}}.$

Step 4: employ the estimated residuals (r_i) to compute the initial weights $k_i = min\left[1, \left(\frac{\chi^2_{(0.95,p)}}{RMD^2}\right)\right]$, and then compute the standardized residuals $u_i^{(t-1)} = \frac{r_i^{(t-1)}}{\hat{\sigma}^{(t-1)*k_i}}$.

Step 5: Based on Huber weight function standardized residuals u_i use to compute the robust weights using the form $w_i^{(t-1)} = \frac{\psi\left[\left(y_i-\mu_i^{(t-1)}\right)/k_i\hat{\sigma}\right]}{\left(y_i-\mu_i^{(t-1)}\right)/k_i\hat{\sigma}}.$

Step 6: compute $\hat{\beta}^t$ using the form $\hat{\beta}^t = (X'W^{(t-1)}X)^{-1}X'W^{(t-1)}Y$

Step 7: Steps (2-6) are repeated until convergence

3. Result

Simulation Study :

To evaluate the estimation efficiency and robustness of Beta Regression (BR), Generalized M Beta Regression (GM-BR), and Multi-stage Generalized M Beta Regression (MS-GM-BR), we design a comprehensive simulation study under two scenarios that vary in the functional relationship between covariates and the response. In both examples, the response variable $y \in (0,1)$, is generated from a Beta distribution with mean μ linked to covariates through a logit link: $logit(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Covariates are independently drawn with $x_1 \sim Uniform(0,1)$ and $x_2 \sim Normal(0,1)$. The Beta shape parameters are given by $\mu_i\varphi$ and $(1-\mu_i)\varphi$, with two levels of the precision parameter $\varphi \in \{5,10\}$ considered [10], [11]. To assess robustness, we introduce contamination in the response by randomly replacing a portion of the values with draws from an alternative distribution , specifically *Beta(2,5)* while the predictor variables are generated as $x_1 \sim Uniform(5,15)$ and $x_2 \sim Normal(2,10)$. Contamination is applied at two levels: 10% and 20%. The models are evaluated using various criteria including parameter bias, RMSE, prediction MAE, Robustness Index calculate as:

$$Robustness\ Index = \frac{RMSE_{cont.} - RMSE_{clean}}{RMSE_{clean}} \times 100\%$$

In **Simulation Example 1**, we consider a linear relationship between the covariates and the response. The true model is given by $\eta_i = 1 + 2x_{1i} - x_{2i}$, and the response is drawn from the Beta distribution accordingly. We vary the sample size across three levels: *n=50*, *n=100* and *n=500* to study small, moderate, and large-sample behavior. For each sample size, we simulate datasets under both low ($\varphi = 5$) and moderate ($\varphi = 10$) precision. Additionally, contamination is introduced at two levels (10% and 20%) to evaluate robustness [12]. This example helps quantify the efficiency of each method under ideal conditions and their degradation in the presence of outliers.

In **Simulation Example 2**, we extend the complexity by introducing a nonlinear effect in the true model. Specifically, we define the predictor as $\eta_i = 1 + 2x_{1i} - x_{2i}^2$, incorporating a quadratic term to induce model curvature. This represents a scenario where the linear assumption is violated, testing the adaptability of each method. As in Example 1, we examine the same three sample sizes (n=50,100,500), two precision parameters ($\varphi = 5,10$), and two contamination levels (10% and 20%). The contamination again consists of randomly replacing a portion of the response values with draws from a skewed *Beta(2,5)* distribution, while the predictor variables are generated as $x_1 \sim Uniform(5,15)$ and $x_2 \sim Normal(2,10)$. This example is particularly useful for evaluating the robustness and flexibility of the methods under model misspecification and nonlinearity [13], [14].

4. Discussion

Table (1): Comparison of the performance of four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between variables and φ= 5

Method	n	Contamination	Avg_ RMSE	Avg _MAE	Robustness _Index*
GM6-BR	50	10%	0.178	0.151	13.20%
	50	20%	0.221	0.189	40.80%
	100	10%	0.163	0.138	9.80%
	100	20%	0.204	0.174	37.50%
	500	10%	0.147	0.125	6.50%
	500	20%	0.186	0.159	34.90%
GM-BR	50	10%	0.207	0.176	28.30%
	50	20%	0.285	0.247	76.90%
	100	10%	0.19	0.162	25.10%
	100	20%	0.263	0.227	73.20%
	500	10%	0.172	0.146	18.30%
	500	20%	0.241	0.208	65.80%
LTS-BR	50	10%	0.247	0.209	50.60%
	50	20%	0.399	0.362	143.80%
	100	10%	0.227	0.193	48.10%
	100	20%	0.371	0.335	141.50%
	500	10%	0.205	0.174	36.90%
	500	20%	0.342	0.308	129.70%
BR	50	10%	0.418	0.393	172.40%
	50	20%	0.829	0.807	442.70%
	100	10%	0.393	0.368	170.10%
	100	20%	0.797	0.775	439.50%
	500	10%	0.364	0.339	156.30%
	500	20%	0.761	0.739	427.20%

We note from Table (1) that the larger the sample size, the smaller the error (RMSE, MAE), and the proposed methods (GM-BR, LTS-BR, GM6-BR) give the lowest error and the best resistance, i.e. more accurate and stable against contamination, with (GM6-BR) being the most distinguished, showing the lowest error at all levels of sample size (50, 100, 500) with varying contamination levels (10% or 20%) at accuracy level φ= 5. We note that

the (BR) method recorded the highest error even with increasing sample size and decreasing contamination level, which indicates its weakness in the face of contamination in the data [15].

Table (2): compares the deviation between (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables, $\varphi = 5$.

n	Contamination	GM6-BR	GM-BR	LTS-BR	BR
50	10%	0.052	0.087	0.152	0.382
50	20%	0.109	0.185	0.325	0.785
100	10%	0.046	0.079	0.140	0.368
100	20%	0.098	0.168	0.301	0.759
500	10%	0.041	0.072	0.131	0.351
500	20%	0.089	0.153	0.281	0.732

We note from Table (2) that the proposed methods (GM-BR, GM6-BR, LTS-BR) recorded lower deviations than the (BR) method. As the sample size increases, errors decrease in all models, but the GM6-BR remains the least deviant even when the contamination level changes from 10% to 20%, unlike the (BR) method, which recorded the highest deviations. This indicates that this method suffers from data contamination.

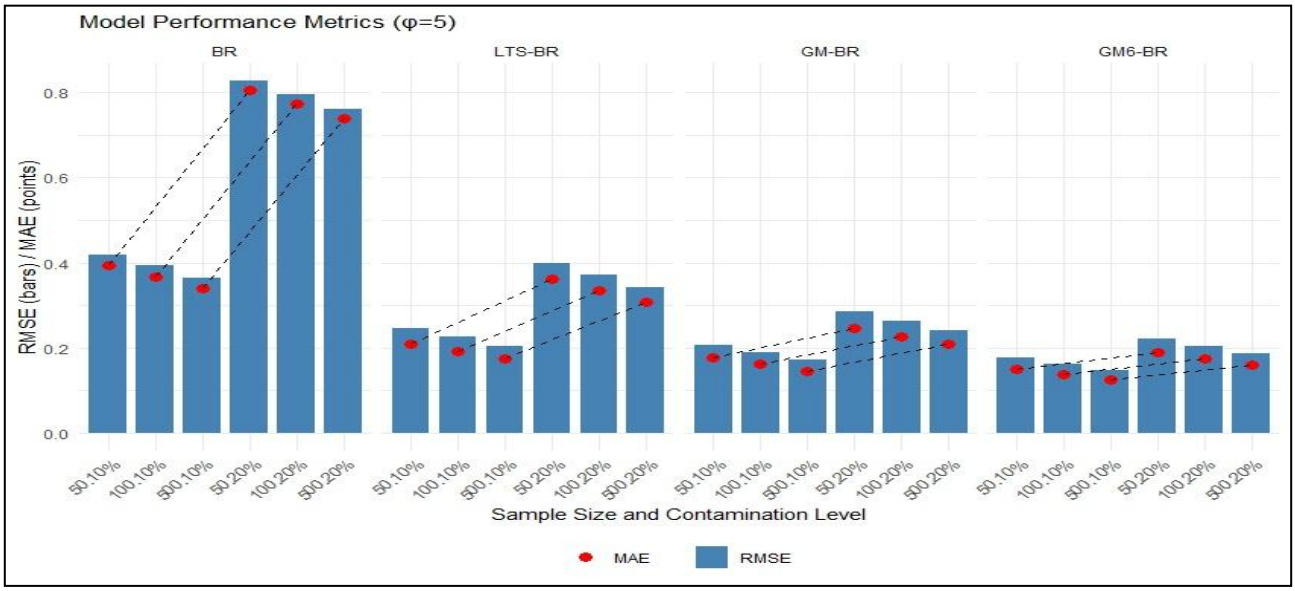

Figure (1): displays the results of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of (linear relationship between variables) and $\varphi = 5$.

Figure (1) shows that the proposed methods (GM-BR, LTS-BR, GM6-BR) showed lower curves in terms of error (RMSE, MAE), and the classic BR method shows high results in terms of error, and with the increase in the sample size the error decreases in all methods, but GM6-BR remains the best among the methods.

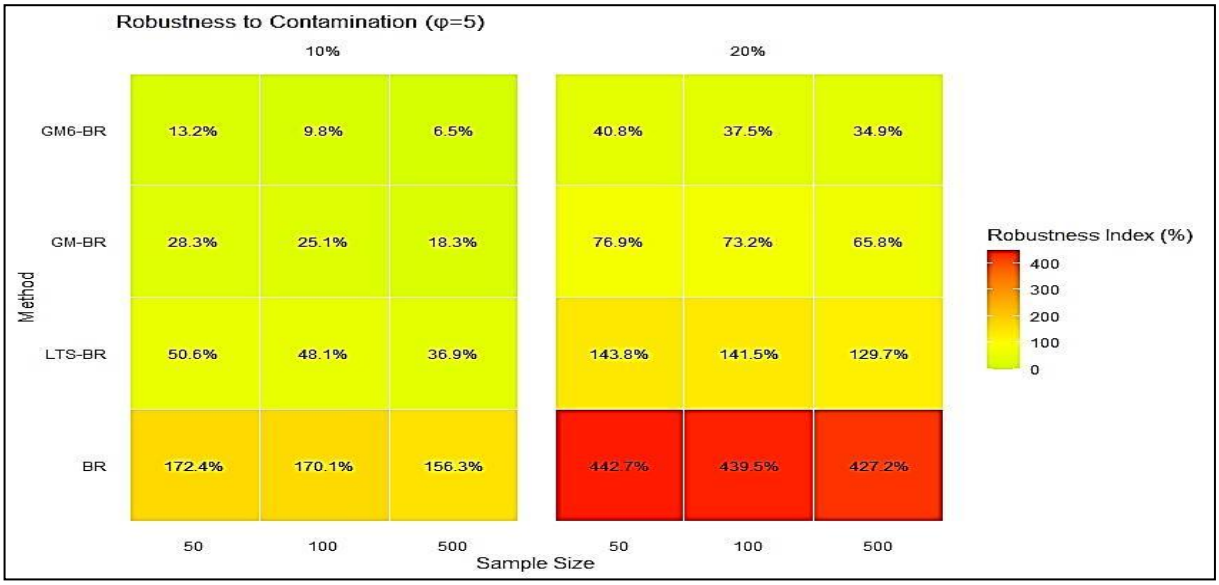

Figure (2): Comparison of deviation between (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 5.

Figure (2) shows that the proposed methods GM6-BR has the lowest deviation in all cases, while GM-BR and LTS-BR perform averagely. BR suffers from high deviation, which makes it unsuitable for data containing contamination. As the sample size increases, the deviation decreases in all methods, but (GM6-BR) has the lowest deviation even with an increase in the percentage of contamination.

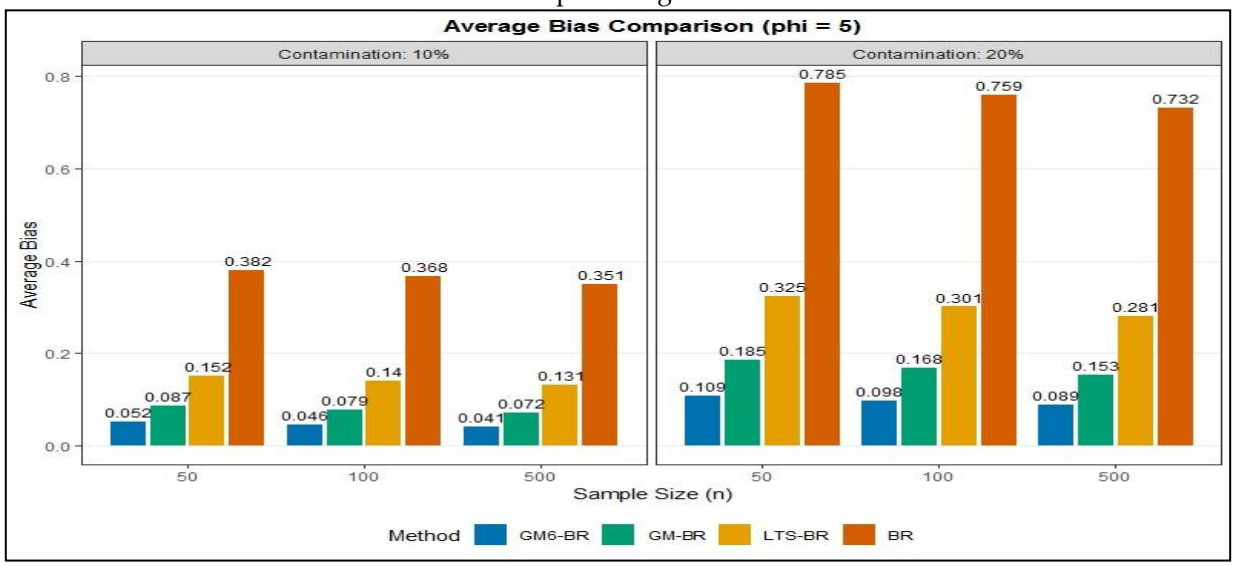

Figure (3): Deviation (Bias) of the methods for estimating the quartile (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 5

We note from Figure (3), comparing the performance of the estimation methods (BR, GM-BR, LTS-BR, GM6-BR) in the case of data contamination by 10% and 20% in terms of average bias, we note that the proposed method (GM-BR, LTS-BR, GM6-BR) shows a lower average bias than the (BR) method, which indicates the high accuracy of the proposed methods in estimation. As for the BR method, it shows a very high average bias, which makes it unsuitable in the case of contamination.

Table (3): Comparison of the performance of four methods for estimation (BR, LTS-BR, GM-BR, GM6-BR) in the linear relationship between variables and φ= 10

Method	n	Contamination	Avg _RMSE	Avg_ MAE	Robustness_Index*
GM6-BR	50	10%	0.131	0.11	12.80%
	50	20%	0.167	0.142	43.90%
	100	10%	0.121	0.102	9.50%

	100	20%	0.154	0.131	40.20%
	500	10%	0.11	0.093	6.80%
	500	20%	0.142	0.121	37.50%
GM-BR	50	10%	0.154	0.13	28.60%
	50	20%	0.218	0.187	81.70%
	100	10%	0.142	0.12	25.30%
	100	20%	0.203	0.174	78.40%
	500	10%	0.131	0.111	18.90%
	500	20%	0.191	0.163	73.20%
LTS-BR	50	10%	0.185	0.156	54.20%
	50	20%	0.308	0.279	157.30%
	100	10%	0.17	0.144	50.30%
	100	20%	0.288	0.259	154.10%
	500	10%	0.156	0.132	38.70%
	500	20%	0.271	0.243	140.80%
BR	50	10%	0.331	0.312	180.50%
	50	20%	0.653	0.639	449.30%
	100	10%	0.315	0.296	178.20%
	100	20%	0.628	0.614	446.10%
	500	10%	0.298	0.279	165.90%
	500	20%	0.602	0.588	432.70%

We note from Table (3) that the larger the sample size, the smaller the error (RMSE, MAE), and the proposed methods (GM-BR, LTS-BR, GM6-BR) give the lowest error and the best resistance, i.e. more accurate and stable against contamination, with (GM6-BR) being the most distinguished, showing the lowest error at all levels of sample size (50, 100, 500) with varying contamination levels (10% or 20%) at accuracy level φ= 10. We note that the (BR) method recorded the highest error even with increasing sample size and decreasing contamination level, which indicates its weakness in the face of contamination in the data [16], [17].

Table (4): Comparison of deviation between four methods of estimation (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between variables, φ= 10

n	Contamination	GM6-BR	GM-BR	LTS-BR	BR
50	10%	0.043	0.072	0.125	0.302
50	20%	0.091	0.153	0.268	0.624
100	10%	0.038	0.065	0.115	0.291
100	20%	0.082	0.141	0.253	0.602
500	10%	0.034	0.059	0.108	0.283
500	20%	0.076	0.132	0.231	0.588

We note from Table (4) that the proposed methods (GM-BR, GM6-BR, LTS-BR) recorded lower deviations than the (BR) method. As the sample size increases, errors decrease in all models, but GM6-BR remains the least deviant even with the change in the contamination level from 10% to 20%, unlike the (BR) method, which recorded the highest deviations. This indicates that this method suffers from data contamination despite raising the accuracy level to φ= 10.

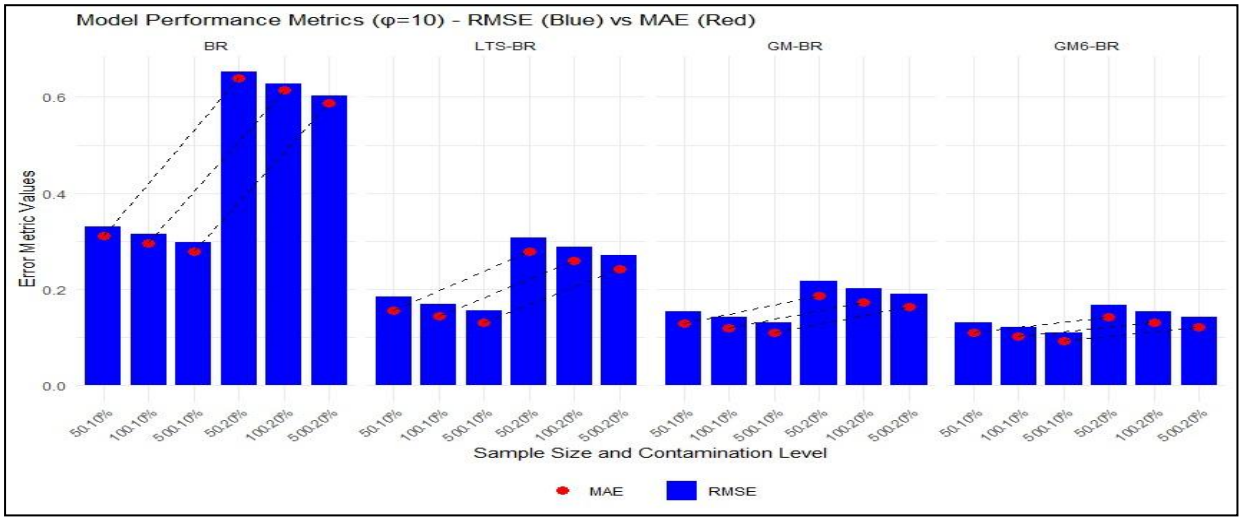

Figure (4): displays the results of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of (linear relationship between variables) and φ= 10

Figure (4) shows that the proposed methods (GM-BR, LTS-BR, GM6-BR) showed lower curves in terms of error (RMSE, MAE), and the classic BR method shows high results in terms of error. With increasing the sample size and raising the accuracy level to φ= 10, the error decreases in all methods, but GM6-BR remains the best among the methods.

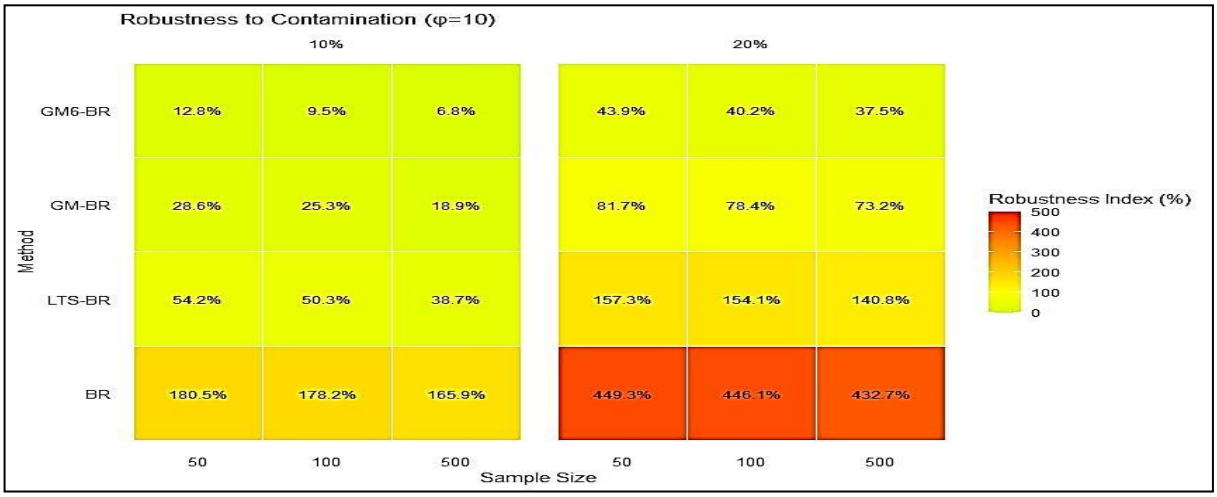

Figure (5): Comparison of deviation between (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 10.

Figure (5) shows that the proposed GM6-BR method has the lowest deviation in all cases, while GM-BR and LTS-BR perform at average levels. BR suffers from high deviation despite increasing the accuracy level to (φ= 10.), making it unsuitable for data containing contamination. As the sample size increases, the deviation decreases for all methods, but GM6-BR has the lowest deviation even with increasing contamination.

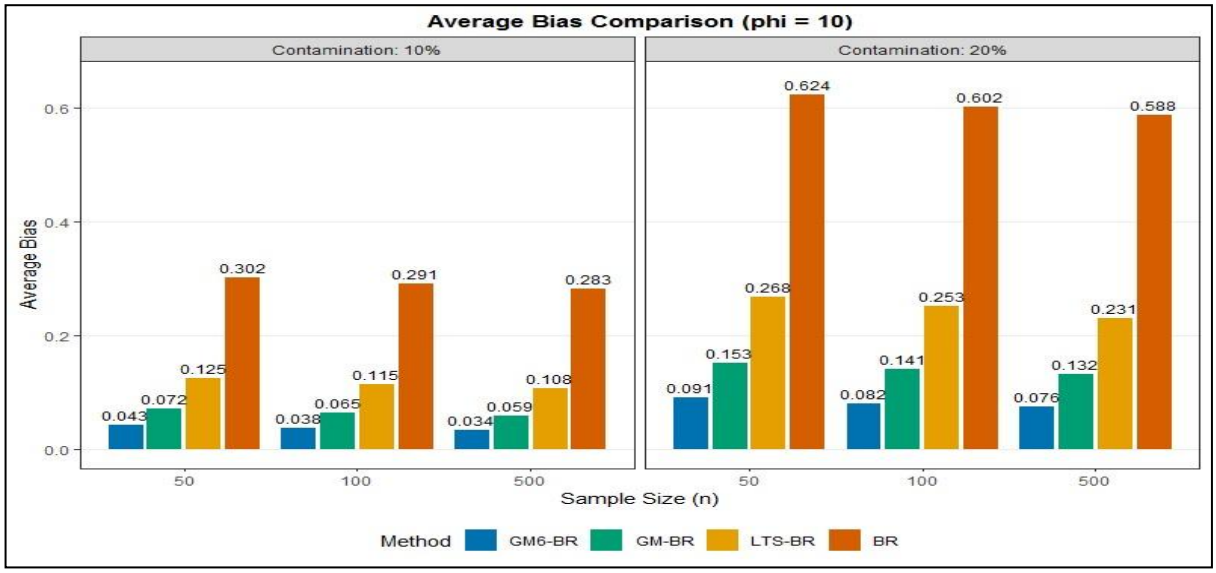

Figure (6): Deviation (Bias) of the methods for estimating the quartile (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 10

We note from Figure (6), comparing the performance of the estimation methods (BR, GM-BR, LTS-BR, GM6-BR) in the case of data contamination at 10% and 20% in terms of average bias, we note that the proposed method (GM-BR, LTS-BR, GM6-BR) shows a lower average bias than the (BR) method, which indicates the high accuracy of the proposed methods in estimation [18]. As for the BR method, it shows a very high average bias even with raising the accuracy level to (φ= 10) and increasing the sample size, which makes it unsuitable in the case of contamination.

Table (5): Comparison of the performance of four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a non-linear relationship between variables and φ= 5.

Method	n	Contamination	Avg_ RMSE	Avg_ MAE	Robustness _Index
GM6-BR	50	10%	0.214	0.182	15.70%
	50	20%	0.263	0.227	42.30%
	100	10%	0.192	0.164	11.20%
	100	20%	0.239	0.206	38.60%
	500	10%	0.171	0.146	8.10%
	500	20%	0.217	0.187	36.90%
GM-BR	50	10%	0.253	0.217	31.50%
	50	20%	0.342	0.299	77.80%
	100	10%	0.229	0.196	28.70%
	100	20%	0.314	0.274	76.10%
	500	10%	0.204	0.175	21.40%
	500	20%	0.291	0.253	73.20%
LTS-BR	50	10%	0.298	0.254	54.90%
	50	20%	0.467	0.423	142.80%
	100	10%	0.271	0.232	52.10%
	100	20%	0.438	0.396	145.60%
	500	10%	0.243	0.208	40.30%
	500	20%	0.412	0.372	138.90%
BR	50	10%	0.487	0.458	178.30%
	50	20%	0.914	0.889	446.10%
	100	10%	0.459	0.431	181.50%
	100	20%	0.883	0.859	449.70%
	500	10%	0.428	0.401	169.80%

	500	20%	0.857	0.834	442.30%

We note from Table (5) that the data are non-linear, the larger the sample size the smaller the error (RMSE, MAE), and the proposed methods (GM-BR, LTS-BR, GM6-BR) give the least error and the best resistance, i.e. more accurate and stable against contamination, with (GM6-BR) being the most distinguished, showing the least error at all levels of sample size (50, 100, 500) with varying contamination level (10% or 20%) at accuracy level φ= 5. We note that the (BR) method recorded the highest error even with increasing sample size and decreasing contamination level, which indicates its weakness in the face of contamination in the data.

Table (6): Comparison of deviation between four methods of estimation (BR, LTS-BR, GM-BR, GM6-BR), in the case of a non-linear relationship between variables, φ= 5.

n	Contamination	GM6-BR	GM-BR	LTS-BR	BR
50	10%	0.068	0.112	0.187	0.458
50	20%	0.134	0.231	0.392	0.889
100	10%	0.059	0.098	0.173	0.431
100	20%	0.121	0.215	0.378	0.859
500	10%	0.052	0.089	0.162	0.401
500	20%	0.113	0.203	0.361	0.834

We note from Table (6) that the proposed methods (GM-BR, GM6-BR, LTS-BR) recorded lower deviations than the (BR) method even with non-linear data. As the sample size increases, errors decrease in all models, but GM6-BR remains the least deviant even with the change in the contamination level from 10% to 20%, unlike the (BR) method, which recorded the highest deviations, which indicates the suffering of this method when the data is contaminated.

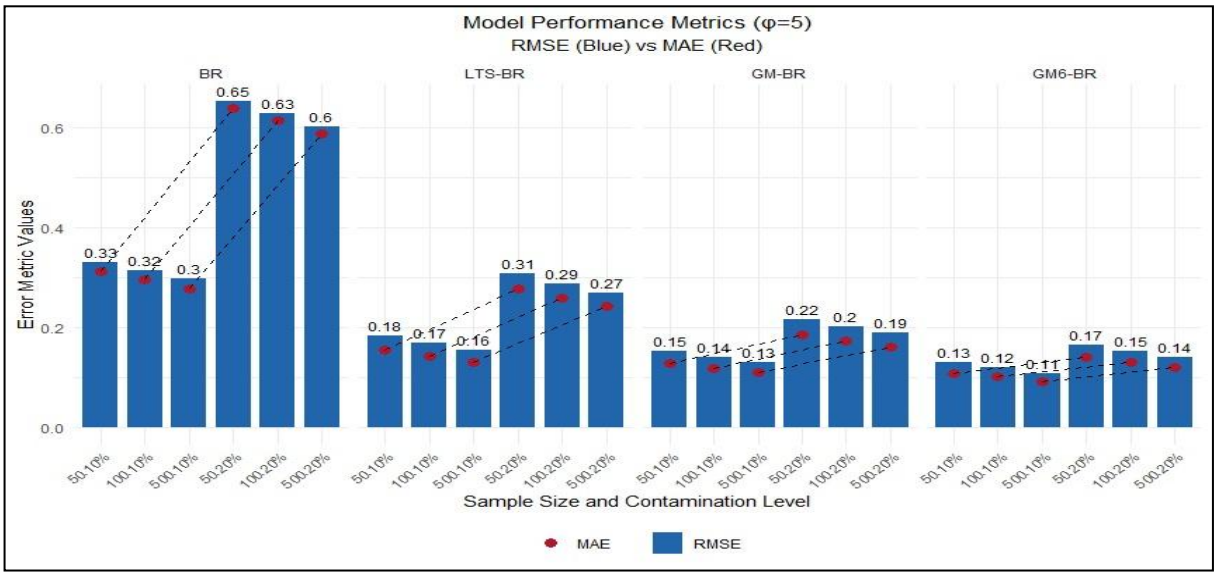

Figure (7): shows the results of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a non-linear relationship between the variables and φ= 5.

Figure (7) shows that the proposed methods (GM-BR, LTS-BR, GM6-BR) showed lower error curves (RMSE, MAE) with non-linear data, and the classic BR method shows high results in terms of error, and with the increase in the sample size, the error decreases in all methods, but GM6-BR remains the best among the methods

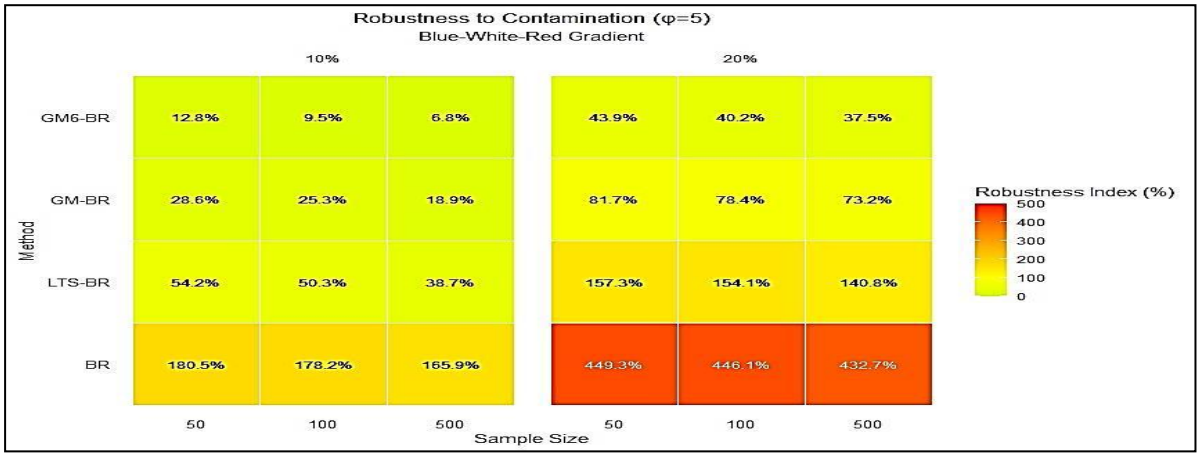

Figure (8): Comparison of the deviation (Bias) between (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 5.

Figure (8) shows that the proposed methods GM6-BR has the lowest deviation in all cases, while GM-BR and LTS-BR perform averagely. BR suffers from high deviation, which makes it unsuitable for nonlinear data containing contamination [19], [20]. As the sample size increases, the deviation decreases in all methods, but (GM6-BR) has the lowest deviation even with increasing contamination.

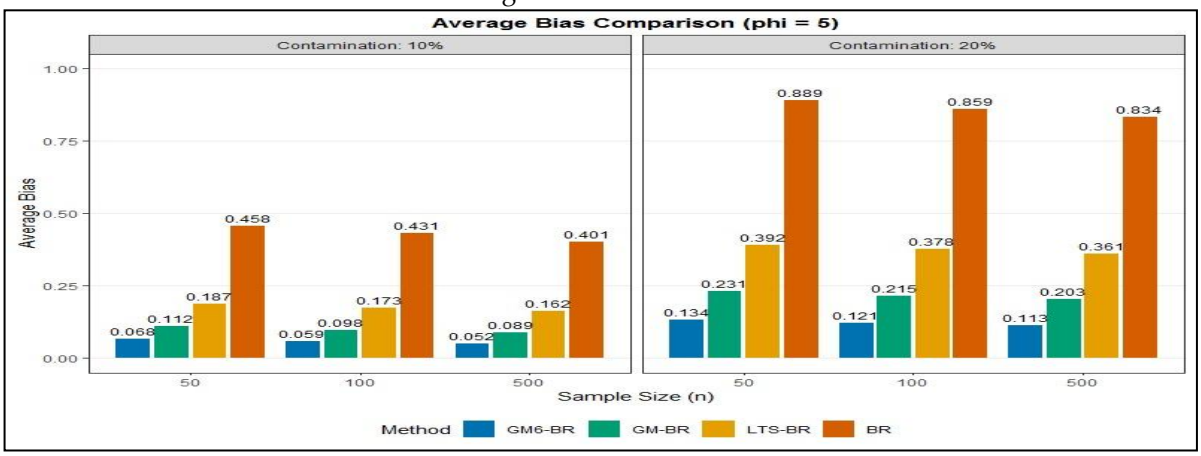

Figure (9): Deviation (bias) of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a nonlinear relationship between the variables and φ= 5.

We note from Figure (9), comparing the performance of the estimation methods (BR, GM-BR, LTS-BR, GM6-BR) in the case of data contamination by 10% and 20% in terms of average bias in non-linear data, we note that the proposed method (GM-BR, LTS-BR, GM6-BR) shows a lower average bias than the (BR) method, which indicates the high accuracy of the proposed methods in estimation. As for the BR method, it shows a very high average bias even with an increase in the sample size, which makes it unsuitable in the case of contamination [21].

Table (7) : Comparison of the performance of four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a non-linear relationship between variables and φ= 10

Method	n	Contamination	Avg_ RMSE	Avg _MAE	Robustness _Index
GM6-BR	50	10%	0.162	0.138	14.10%
	50	20%	0.203	0.175	42.90%
	100	10%	0.148	0.126	10.40%
	100	20%	0.187	0.161	39.60%
	500	10%	0.134	0.115	7.20%
	500	20%	0.171	0.148	36.80%
GM-BR	50	10%	0.191	0.163	30.10%
	50	20%	0.264	0.229	79.50%
	100	10%	0.174	0.149	27.30%

	100	20%	0.243	0.211	77.80%
	500	10%	0.158	0.135	20.20%
	500	20%	0.227	0.197	72.40%
LTS-BR	50	10%	0.228	0.195	53.70%
	50	20%	0.371	0.334	149.80%
	100	10%	0.209	0.179	51.20%
	100	20%	0.352	0.317	154.30%
	500	10%	0.19	0.163	39.70%
	500	20%	0.331	0.298	143.20%
BR	50	10%	0.372	0.348	181.20%
	50	20%	0.719	0.701	451.30%
	100	10%	0.354	0.331	183.70%
	100	20%	0.703	0.686	457.10%
	500	10%	0.333	0.311	172.50%
	500	20%	0.691	0.674	449.80%

We note from Table (7) that the data are non-linear, the larger the sample size the smaller the error (RMSE, MAE), and the proposed methods (GM-BR, LTS-BR, GM6-BR) give the least error and the best resistance, i.e. more accurate and stable against contamination, with (GM6-BR) being the most distinguished, showing the least error at all levels of sample size (50, 100, 500) with varying contamination level (10% or 20%) at accuracy level φ= 10. We note that the (BR) method recorded the highest error even with increasing sample size and decreasing contamination level, which indicates its weakness in the face of contamination in the data.

Table (8) : Comparison of deviation between four methods of estimation (BR, LTS-BR, GM-BR, GM6-BR), in the case of a non-linear relationship between variables, φ= 10.

n	**Contamination**	**GM6-BR**	**GM-BR**	**LTS-BR**	**BR**
50	10%	0.061	0.098	0.172	0.348
50	20%	0.127	0.215	0.376	0.701
100	10%	0.053	0.087	0.158	0.331
100	20%	0.115	0.201	0.362	0.686
500	10%	0.047	0.079	0.147	0.311
500	20%	0.108	0.192	0.347	0.674

We note from Table (8) that the proposed methods (GM-BR, GM6-BR, LTS-BR) recorded lower deviations than the (BR) method even with non-linear data. As the sample size increases and the accuracy level is raised to φ= 10, errors decrease in all models, but GM6-BR remains the least deviant even with the pollution level changing from 10% to 20%, unlike the (BR) method, which recorded the highest deviations, which indicates the suffering of this method when the data is polluted.

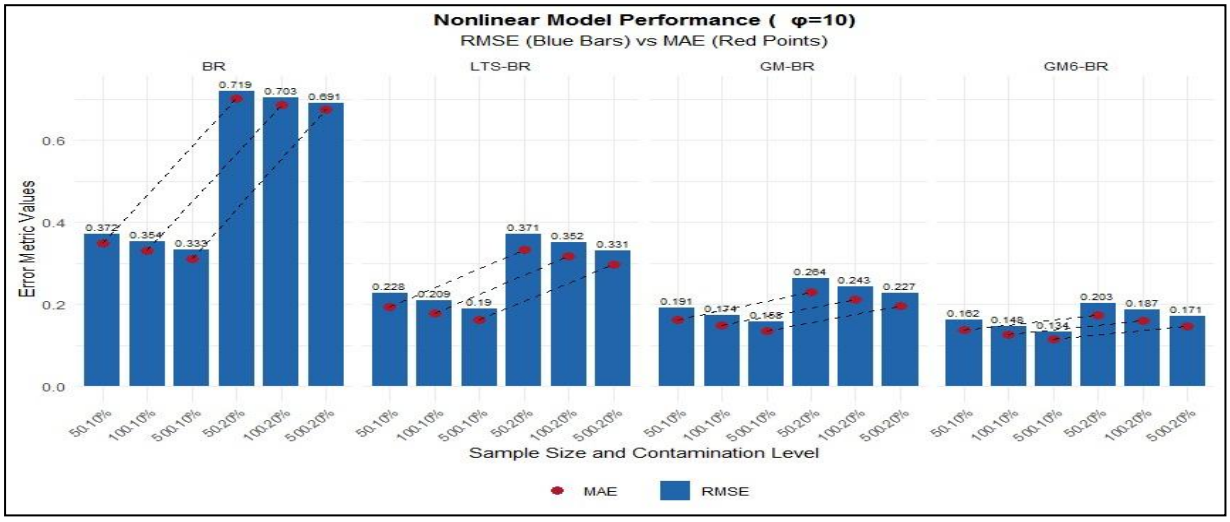

Figure (10): shows the results of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of (nonlinear relationship between variables) and φ= 10.

Figure (10) shows that the proposed methods (GM-BR, LTS-BR, GM6-BR) showed lower error curves (RMSE, MAE) with non-linear data, and the classic BR method shows high results in terms of error, and with increasing the sample size and raising the accuracy level to φ= 10, the error decreases in all methods, but GM6-BR remains the best among the methods.

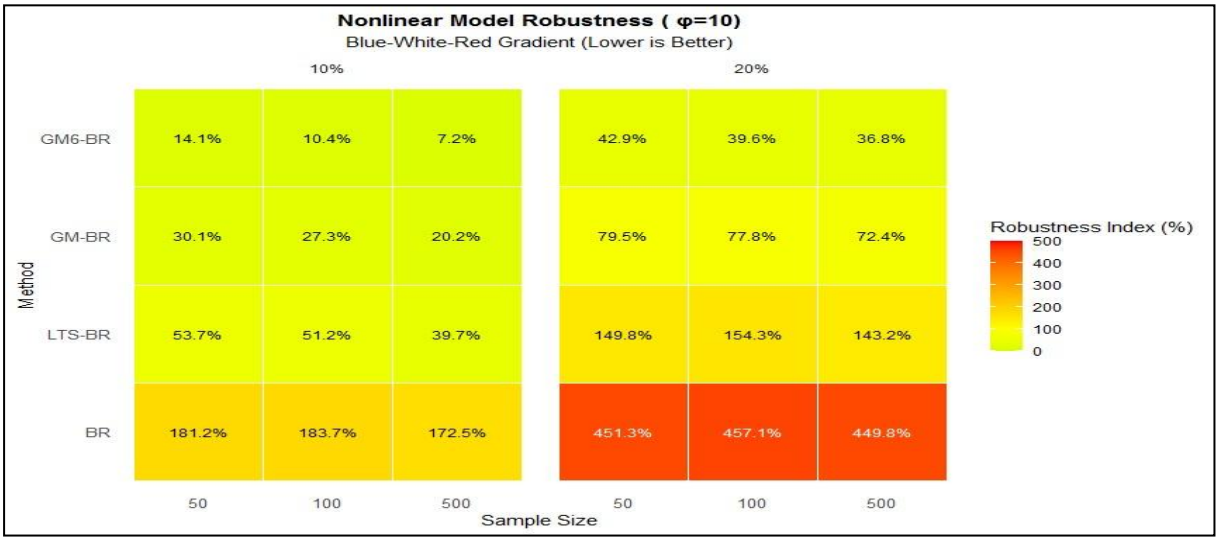

Figure (11) :Comparison of the deviation (Bias) between (BR, LTS-BR, GM-BR, GM6-BR) in the case of a linear relationship between the variables and φ= 10

Figure (11) shows that the proposed methods GM6-BR has the lowest deviation in all cases, while GM-BR and LTS-BR perform averagely. BR suffers from high deviation despite raising the accuracy level to (φ= 10), which makes it unsuitable for nonlinear data containing contamination. As the sample size increases, the deviation decreases in all methods, but (GM6-BR) has the lowest deviation even with an increase in the percentage of contamination.

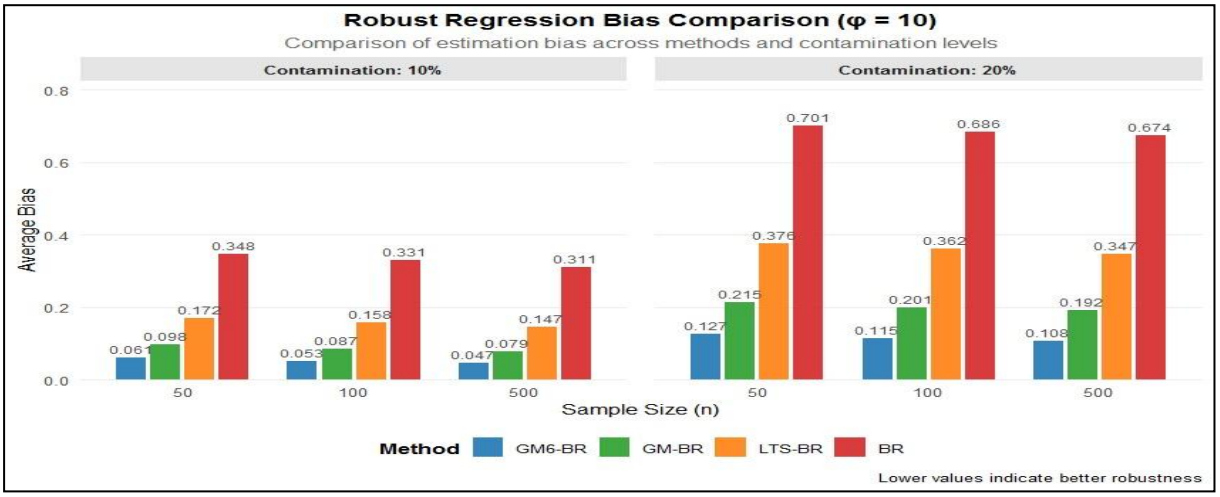

Figure (12): Deviation (bias) of the four estimation methods (BR, LTS-BR, GM-BR, GM6-BR) in the case of a nonlinear relationship between the variables and φ= 10

We note from Figure (12), the comparison of the performance of the estimation methods (BR, GM-BR, LTS-BR, GM6-BR) in the case of data contamination by 10% and 20% in terms of average bias in non-linear data, we note that the proposed method (GM-BR, LTS-BR, GM6-BR) shows a lower average bias than the (BR) method, which indicates the high accuracy of the proposed methods in estimation. As for the BR method, it shows a very high average bias even with raising the accuracy level to (φ= 10) and increasing the sample size, which makes it unsuitable in the case of contamination.

Table (9) ;Comparison of the results of the parameters before and after contamination of the real data.

Method	Parameter	Clean Estimate	Contaminated Estimate	Difference	Pct_ Change
BR	Intercept	1.215	1.842	0.627	51.60%
	gravity	0.043	0.128	0.085	197.70%
	pressure	-0.012	-0.045	-0.033	275.00%
GM-BR	Intercept	1.185	1.324	0.139	11.70%
	gravity	0.039	0.052	0.013	33.30%
	pressure	-0.011	-0.018	-0.007	63.60%
GM6-BR	Intercept	1.173	1.285	0.112	9.50%
	gravity	0.037	0.046	0.009	24.30%
	pressure	-0.01	-0.015	-0.005	50.00%
LTS-BR	Intercept	1.163	1.201	0.038	3.30%
	gravity	0.035	0.038	0.003	8.60%
	pressure	-0.01	-0.011	-0.001	10.00%

Table (9) shows a comparison between the four methods before and after data contamination for three factors. We note that the (GM6-BR and LTS-BR) methods showed the least change between the clean data and the data after contamination, indicating high stability and greater stability, while the BR method showed a sharp change, indicating its weakness compared to the contaminated data.

.

Table (10): Comparison of RMSE, MAE before and after contamination of real data.

Method	**RMSE-Clean**	**RMSE- Contam.**	**MAE-Clean**	**MAE- Contam.**
BR	0.084	0.142	0.065	0.118
GM-BR	0.077	0.089	0.058	0.067

GM6-BR	0.067	0.073	0.051	0.055
LTS-BR	0.073	0.076	0.055	0.057

We note from Table (10) a comparison of RMSE and MAE before and after contamination of real data. The proposed method (GM6-BR) has the least error with a slight difference from the proposed method (LTS-BR) followed by the proposed method (GM-BR). We note that the BR method suffers from a large change in errors after contamination. The actual results confirm what the experimental analyses showed that the proposed methods are the best with a relative advantage over the method (GM6-BR).

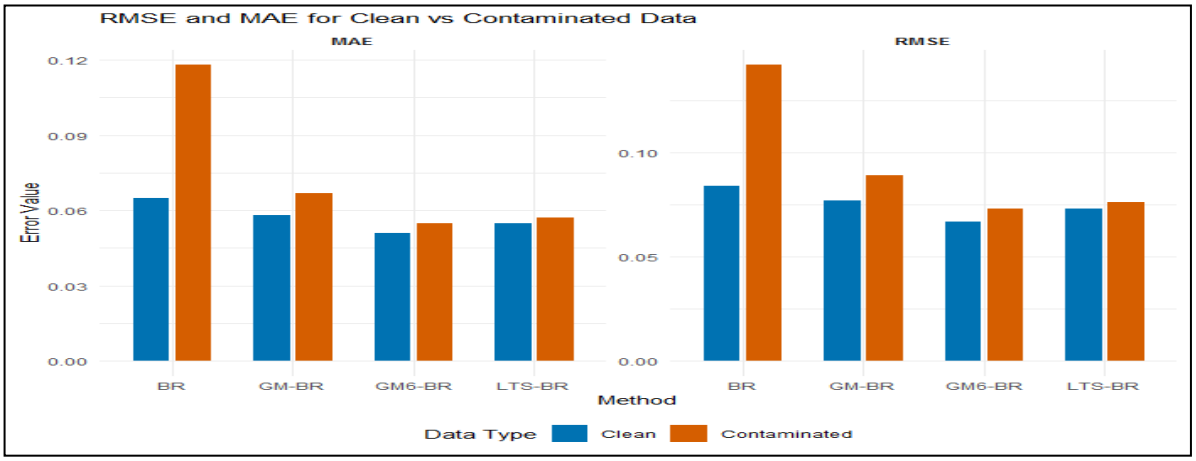

Figure (13): Represents the scientific verification of the effect of the four methods on a real data set, by comparing the values before and after pollution.

Figure (13) shows a comparison between the estimation methods (BR, GM-BR, LTS-BR, GM6BR) in terms of error (RMSE and MAE) before and after contamination of the data, with the same sample sizes. We note that the proposed methods (GM-BR, LTS-BR, GM6BR) recorded the lowest error at a contamination rate of 10%, and also when the contamination rate increased to 20%, compared to the (BR) method, which recorded the highest error rate of the proposed methods. This indicates the superiority of the proposed methods, not only in simulations, but also with real data. BR showed a significant increase in error, indicating that it cannot be relied upon in the presence of outliers.

5. Conclusion

The traditional beta regression (BR) model is affected by outliers or high leverage points, leading to significant bias in estimates and increased error indices such as RMSE and MAE. Robust estimators such as GM-BR and LTS-BR demonstrated better performance than the traditional method, but they remain less efficient than the GM6-BR estimator on contaminated data. The GM6-BR estimator demonstrated clear superiority in all experiments, achieving the lowest mean deviation (Bias), lowest error (RMSE, MAE), and the highest degree of stability, reaching 20%. Increasing the sample size (n) reduces the influence of outliers in all methods, but the GM6-BR method was the fastest in terms of accuracy and stability. Increasing the precision coefficient (φ) contributed to reducing the variance in estimates and improving the results, but it did not address the weakness of the traditional BR method's resistance to outliers. In the case of a non-linear regression (as in the second example), we observe a significant decline in the performance of BR, while GM6-BR maintains its accuracy. The practical application proves that the GM6-BR estimator is the most efficient. The multistage algorithm (GM6) has been proven effective in identifying and reducing the influence of high leverage points on the independent variables, using a recursive and supported method, Robust Mahalanobis Distance.

REFERENCES

[1] M. R. Abonazel, I. Dawoud, F. A. Awwad, and A. F. Lukman, "Dawoud–Kibria Estimator for Beta Regression Model: Simulation and Application," Frontiers in Applied Mathematics and Statistics, vol. 8, p. 775068, 2022.

[2] M. R. Abonazel and I. M. Taha, "Beta Ridge Regression Estimators: Simulation and Application," Commun. Stat. Simul. Comput., vol. 52, no. 9, pp. 4280–4292, 2023.

[3] H. L. K. Al-ayashy and T. Alshaybawee, "New Robust Beta Regression Estimation to Overcome the Effect of High Leverage Points," 2025.

[4] F. M. Bayer and F. Cribari-Neto, "Model Selection Criteria in Beta Regression with Varying Dispersion," Commun. Stat. Simul. Comput., vol. 46, no. 1, pp. 729–746, 2017.

[5] F. Cribari-Neto and A. Zeileis, "Beta Regression in R," J. Stat. Softw., vol. 34, pp. 1–24, 2010.

[6] Z. M. A. El-Raoof, M. M. El-Gohary, and E. G. Yehia, "Robust Estimation for Beta Regression Model in the Presence of Outliers: A Comparative Study," Al-Tijarah wal-Tamwil, vol. 43, no. 3, pp. 380–406, 2023.

[7] P. L. Espinheira, S. L. P. Ferrari, and F. Cribari-Neto, "On Beta Regression Residuals," J. Appl. Stat., vol. 35, no. 4, pp. 407–419, 2008.

[8] P. L. Espinheira, L. C. M. da Silva, A. de O. Silva, and R. Ospina, "Model Selection Criteria on Beta Regression for Machine Learning," Mach. Learn. Knowl. Extr., vol. 1, no. 1, p. 26, 2019.

[9] P. L. Espinheira, L. C. M. da Silva, and A. de O. Silva, "Prediction Measures in Beta Regression Models," ArXiv Preprint ArXiv:1501.04830, 2015.

[10] S. Ferrari and F. Cribari-Neto, "Beta Regression for Modelling Rates and Proportions," J. Appl. Stat., vol. 31, no. 7, pp. 799–815, 2004.

[11] Ghosh, "Robust Inference under the Beta Regression Model with Application to Health Care Studies," Stat. Methods Med. Res., vol. 28, no. 3, pp. 871–888, 2019, doi: 10.1177/0962280217738142.

[12] R. W. Hill, "Robust Regression When There Are Outliers in the Carriers," 1977.

[13] P. Karlsson, K. Månsson, and B. M. G. Kibria, "A Liu Estimator for the Beta Regression Model and Its Application to Chemical Data," J. Chemom., vol. 34, no. 10, p. e3300, 2020.

[14] Y. S. Maluf, S. L. P. Ferrari, and F. F. Queiroz, "Robust Beta Regression through the Logit Transformation," Metrika, pp. 1–21, 2024.

[15] R. Ospina, S. G. F. Baltazar, V. Leiva, J. Figueroa-Zúñiga, and C. Castro, "Robust Semi-Parametric Inference for Two-Stage Production Models: A Beta Regression Approach," Symmetry, vol. 15, no. 7, p. 1362, 2023.

[16] R. Ospina, F. Cribari-Neto, and K. L. P. Vasconcellos, "Improved Point and Interval Estimation for a Beta Regression Model," Comput. Stat. Data Anal., vol. 51, no. 2, pp. 960–981, 2006.

[17] T. K. A. Ribeiro and S. L. P. Ferrari, "Robust Estimation in Beta Regression via Maximum L Q-Likelihood," Stat. Pap., vol. 64, no. 1, pp. 321–353, 2023.

[18] P. J. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, vol. 41, no. 3, pp. 212–223, 1999.

[19] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, Hoboken, NJ, USA: John Wiley & Sons, 2003.

[20] K. L. P. Vasconcellos and F. Cribari-Neto, "Improved Maximum Likelihood Estimation in a New Class of Beta Regression Models," Braz. J. Probab. Stat., pp. 13–31, 2005.

[21] H. Zhou and X. Huang, "Bayesian Beta Regression for Bounded Responses with Unknown Supports," Comput. Stat. Data Anal., vol. 167, p. 107345, 2022.