Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Published online: 30.01.2025
DOI : 10.15672/1544513

Research Type

Influence of residuals on Cook's distance for Beta regression model: Simulation and application

Javaria Ahmad Khan*[1], Atif Akbar[2], B. M. Golam Kibria[3]

[1,2] *Department of Statistics, Bahuddin Zakariya University, Multan, PAKISTAN*
[3] *Department of Mathematics and Statistics, Florida International University, Miami, FL 33199, USA*

Abstract

Cook's distance is one of the renowned and classic tools for the detection of influential observations. In this article, we propose to use Cook's distance with different residuals in the Beta regression model, which is appropriate for modeling the response variable that undertakes a proportion data set. The influence of outlying observations on the basis of its estimated parameters and mean squared error is examined, and performance of residuals is compared. Based on the simulation results and the empirical application, it is observed that the performance of the deviation and weighted residuals is better than that of the rest of the residuals for the detection of influential observations. The observations deleted by the deviance residuals have a large impact on the regression coefficients and on the mean squared error for the Beta regression model.

Mathematics Subject Classification (2020). 62J07, 62J12, 62J20

Keywords. Beta regression, Cook's distance, outliers, reading skills data, residuals, crude oil conversion data.

1. Introduction

A class of beta regression model (BRM) was proposed by Ferrari and Cribari-Neto [14], which are similar to generalized linear models (GLM) in many aspects. The beta distribution is a continuous type of distribution and they considered situations where the response is restricted to the interval (0; 1), such as percentages, proportions, rates, and fractions. Outliers or influential observation(s) are not appreciated in the data sets. In practice, this situation is violated in the linear regression model (LRM) and in GLM, affecting the estimation of the parameters and the related inference [15]. Therefore, it is necessary to diagnose and then treat these unusual points before fitting a model.
A vast amount of literature is available that provides different diagnostic techniques for LRM; see [4, 6, 7, 18, 27], among others. With reference to GLM, Pregibon [22] took initiative to study the diagnostic of influential observations in the logistic regression model using Cook's distance, which was followed by many researchers who proposed different diagnostic methods for different regression models. The residuals play a significant role

*Corresponding Author.
Email addresses: jakhan0@yahoo.com (J. A. Khan), atifakbar@bzu.edu.com (A. Akbar), kibriag@fiu.edu. (B. M. G. Kibria)
Received: 06.09.2024; Accepted: 22.01.2025

in regression diagnostics and different modified forms of famous residuals along with new types have been presented, e.g., [1,17,23,26,28–31,41] among others. Specifically, for BRM, Espinheira et al. [10, 11] proposed some residuals and the likelihood distance method for influential diagnostics. Simas et al. [33] generalized the results [10] and constructed some residuals, and a Portmanteau test for serial correlation. Rocha and Simas [24] also worked on the influence diagnostics of the beta regression model. Anholeto et al. [2] studied the adjusted Pearson residuals for the beta regression model. Espinheira et al. [12] proposed a model selection criterion that is directly related to the leverage, residuals, and influence of the observations. Pereira [21] proposed quantile residuals for BRM and Caribari-Naeto et al. [35] developed tests of correct specification for the BRM model. So, according to the literature, researchers focused on proposing new residuals and illustrated their performance, but the impact of different residuals has not been studied, especially with Cook's distance technique.
The objective of this study is to highlight the performance of different residuals in Cook's distance and to highlight how much their detected observations influence them. This means that after the exclusion of suspected observations, how do the model coefficients, p-values, mean squared errors, etc. respond? This would help the researchers identify the suitable residual with the Cook's distance in BRM.

The paper unfolds as follows. Section 2 presents the BRM, Cook's distance, and the associated residuals. A simulation study has been conducted in Section 3. Real data examples of reading skills are presented in Section 4. Finally, concluding remarks are given in Section 5.

2. The Beta regression model, Cook's distance, and residuals

In this section, we summarize the BRM, Cook's distance, influence diagnostics, and associated residuals.

2.1. The Beta regression model

Let y be the dependent variable which comes from the beta distribution type I, with shape parameter μ and scale parameter ϕ, which is denoted as Beta (α,β) and the probability density function of the beta distribution is given as

$$f(y;\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)},\ y \in (0,1), \alpha > 0, \beta > 0 \tag{2.1}$$

where $B(\alpha,\beta)) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. The mean and variance of beta distribution are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively. For the formulation of BRM, consider the following notation.
By following [14], [32] and [36], we re-parametrize Equation (2.1) as $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha+\beta$, so after reparameterization beta density function becomes

$$f(y;\alpha,\beta) = \frac{\Gamma(\phi)y^{\mu\phi-1}(1-y)^{\phi-\mu\phi-1}}{\Gamma(\mu\phi)\Gamma(\phi-\mu\phi)},\ y \in (0,1), 0 < \mu < 1 \tag{2.2}$$

where ϕ is precision parameter and reciprocal of ϕ is dispersion parameter.
Let $\mathbf{y}_i = (y_1, y_2, \ldots, y_n)'$, $\quad i = 1, 2, \ldots, n$, be the vector of the independent response variable, where $Y \sim \beta_e(\mu,\phi)$. Different link functions can be used for BRM, such as logit, probit, loglog, complementary loglog, and Cauchy link functions. Ferrari and Cribari-Neto [14] suggested using the logit link function as $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} = \text{logit}(\mu_i)$, where

$$\mu_i = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}.$$

Here, $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})'$ is the matrix of $(p+1)$ explanatory variables, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients.
The log-likelihood function of the beta distribution (2.2) is given as

$$\sum_{i=1}^{n} l_i(\mu_i, \phi) = \sum_{i=1}^{n} [ln\Gamma(\phi) - ln\Gamma(\mu_i, \phi) - ln\Gamma(\phi - \mu_i\phi) + (\mu_i\phi - 1)lny_i + (\phi - \mu_i\phi - 1)ln(1 - y_i)].$$

The maximum likelihood (ML) estimators for $\hat{\mu}$ and $\hat{\phi}$ can be observed by solving following simultaneously equations

$$\psi(\hat{\mu}) - \psi(\hat{\mu} + \hat{\phi}) = n^{-1} \sum_{i=1}^{n} lny_i.$$

$$\psi(\hat{\phi}) - \psi(\hat{\mu} + \hat{\phi}) = n^{-1} \sum_{i=1}^{n} ln(1 - y_i).$$

where $\psi(.)$ is the digamma function using the iterative weighted least squares (IWLS) method or Fishers scoring algorithm [13]. For parameter estimation, we used **R** package *betareg*, which based on the auxiliary linear regression of the transformed response as initial values for estimation [8].

2.2. Influence diagnostics

Outliers are first noted by [3] as unusual values in regression modeling that affect parameter estimation and statistical inference. Extreme value in the response variable termed an outlier while extreme value in the explanatory variable(s) known as influential observation. Although an influential observation strongly affects the parameter estimates and fitted values, outliers may or may not affect the parameter estimates. It is necessary to address these values while fitting any regression model. Various tools for influence diagnostics are available in the literature, as we mentioned before, for LRM and GLM. Some of them are discussed here for the influence diagnostics of BRM because limited work has been done on this issue. Residuals analysis has a vital role in formulating theories and their validation in regression modeling. In this study, we used the popular measure, i.e. Cook's distance, to detect influential observations for the BRM. The Cook's distance results using eight different residuals are calculated for BRM and they are Pearson, Deviance, Response, Working, Standardized, Weighted, Sweighted and Sweighted2. The details are available in Table 1.

2.3. Cook's distance and considered residuals

Cook's distance was first proposed by [6] for the LRM and Pregibon [22] later applied this technique to GLM, to identify influential observations. It measures the overall change in the fitted model when the i^{th} observation is deleted from the model. The Cook's distance statistic for the BRM is defined as

$$CD_i = \frac{(\hat{\beta}_{ML} - \hat{\beta}_{ML(i)})' X' W X (\hat{\beta}_{ML} - \hat{\beta}_{ML(i)})}{(k+1)\hat{\phi}}. \tag{2.3}$$

where $\hat{\beta}$ is the estimated BRM coefficients vector for full model and $\hat{\beta}_{ML(i)}$ is the estimated BRM coefficients vector after deleting the i^{th} observation. McCullagh and Nelder [19], simplify Equation (2.3) as

$$CD_i = \frac{\pi_i^2}{k+1} \frac{h_{ii}}{1 - h_{ii}}. \tag{2.4}$$

where h_{ii} is the i^{th} value of the hat matrix and π_i is the residual i^{th}, which we explain later. The largest value of Cook's distance indicates that i^{th} is the influential observation. Cut-off point for the detection of influential observation using Cook's distance statistics

Table 1. Summary of residuals.

Type	Mathematical notation	Reference
Pearson Residual (P)	$r_t = \frac{y_t - \hat{f}(x_t)}{\sqrt{\hat{f}(x_t)}}$	[9]
Deviance Residual (D)	$r_t^d = sign(y_t - \hat{\mu}_t)\sqrt{(d_t)}$	
Working Residual (Wor)	$r_t = \frac{y_t - \hat{f}(x_t)}{\hat{f}(x_t)}$	
Response Residual (R)	$r_t = y_t - \hat{f}(x_t)$	
Weighted Residual (W)	$r_t^* = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\phi\nu_t}}$	[10]
Standardized Weighted Residual (SW)	$r_t^\omega = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\nu_t}}$	
Standardized Weighted 2 Residual (SW2)	$r_t^{\omega\omega} = \frac{r_t^\omega}{\sqrt{1-h_{tt}}}$	
Standardized Residuals (St)	$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{var}(y_t)}}$	

in the BRM is 2*mean (Cook's distance) [16]. We consider residuals summarized in Table 1 for Cook's distance.

3. Simulation study

The primary objective of this section is to compare the performance of the Cook's distance on BRM through a simulation study. The following Monte Carlo simulation study is considered with 1000 replications.

The dependent variable of the BRM is generated from the Beta distribution as $y_i \sim B(\mu, \phi)$ for $i = 1, 2, \ldots, n$, where $\mu_i = E(y_i) = 0.5$ is the arbitrary mean, and ϕ is the dispersion parameter that is assumed to take arbitrary values $\phi = 0.5, 1, 3, 10$. These values represent low-, medium-, and high-variance conditions. Two explanatory variables x_1 and x_2 are kept fixed throughout the simulation study. Here, the design matrix $\mathbf{X}$, with no influential points, of the sample sizes $n = 25, 50, 100,$ and 200, is generated as

$$X_{ij} \sim U(0, 0.5), \quad i = 1, 2, \ldots, n;\ j = 1, 2,$$

and then we make the 5th, 10th, 15th, 20th, and 25th points influential in $\mathbf{X}$ as

$$x_{ij} = a_0 + x_{ij}, \quad i = 5, 10, 15, 20, 25;\ j = 1, 2,$$

where $a_0 = \bar{x}_j + 100$. The simulated outlier detection rate (in percentage) of the BRM Cook's distance under different factors such as sample size, dispersion parameter, and different types of residuals is presented in Table 2. For a better picture, we obtained the average detection rate along with the standard deviation for all methods in n and ϕ and presented them in the last two rows of Table 2.

Table 2 reveals that the performance of Cook's distance is very poor with all residuals in detection of influential observations for small sample size. It becomes worse with residual working and for low dispersion, that is, $\phi = 0.5, 1$, but the increase in sample size with highly dispersed data makes the detection percentage very appropriate in all cases. For $\phi = 10$, performance of Cook's distance is remarkable with almost all sample sizes and it performs a tremendous with large sample size. Table 2 indicates that the percentage of detection of an outlier for all residuals increases as the sample size increases. The dispersion parameter also affects directly on the BRM Cook's distance in detecting the influential observations. The detection power of the BRM Cook's distance is increased for diagnosing the influential observations when the value of the dispersion parameter is increased. By comparing the performance of all residuals of BRM, we find that Cook's distance with weighted and deviance residuals performed better than rest of residuals, and working residual performed worse. It is important to mention that Cook's distance with Weighted and Standardized Weighted residual are performing exactly the same in

detection of influential observations for each sample size and dispersion. Espinheira et al. [10] showed that the SWeighted2 residual is the best choice to be used in likelihood displacement (LD) but the SWeighted2 residual fails to perform best with Cook's distance in detection of influential observations, although its performance is acceptable.

Table 2. Estimated outlier detection rate (%) of the BRM influence diagnostics with different residuals.

n	ϕ	P	D	R	W	SW	SW2	Wor
25	**10**	27.5	28.8	28.1	28.4	28.4	26.5	0.8
	3	22.6	24.9	22	23.1	23.1	20.5	0.7
	1	10.6	17.7	12.2	21.1	21.1	19.6	0.1
	0.5	5.7	13.2	9.2	20.2	20.2	19.2	0.2
50	**10**	86.9	87.3	86.9	87.3	87.3	83.5	51.7
	3	88.5	90.6	88.4	89.8	89.8	86.3	48.2
	1	82.3	80.3	80.5	83.1	83.1	78.1	40.5
	0.5	72.9	83.9	71.1	88.1	88.1	86.3	31.2
100	**10**	98.3	98.7	98.3	98.7	98.7	96.8	90
	3	98.1	98.3	98.3	98.8	98.8	96.6	89.4
	1	98.2	98.8	97.9	99.4	99.4	98.9	92
	0.5	97.1	98.2	97.1	99.8	99.8	99.3	92.5
200	**10**	99.8	99.8	99.9	99.8	99.8	99.4	98.5
	3	99.8	99.9	99.7	99.9	99.9	99.8	99.2
	1	99.7	100	99.9	99.9	99.9	99.8	99.3
	0.5	99.5	99.7	99.6	99.9	99.9	99.9	98.3
Mean		74.2	76.3	74.3	77.3	77.3	75.7	58.3
SD		35.5	33.6	34.9	32.8	32.8	33.1	41.2

4. Real data applications

BRM with Cook's distance can be widely used in different fields of life where large values are not appreciated, e.g. medical, engineering, agriculture, etc. To illustrate its use in diverse fields, we examine the performance of Cook's distance on BRM through two different real-life datasets where detecting influential observations is critical.

4.1. Reading skills data

This application is based on the data set given in [20] which was analyzed by [25]. In this data set, the response variable (y) is the scores on a reading accuracy test of 44 children, and the covariates are dyslexia versus non-dyslexia status (x_1), nonverbal IQ converted to z scores (x_2) and an interaction variable (x_3). Participants (19 dyslexics and 25 controls) were recruited from primary schools in the Australian Capital Territory. The ages of the children ranged from eight years and five months to twelve years and three months. The covariate x_1 assumes the value 1 when the child is dyslexic and -1 otherwise. The observed scores y were linearly transformed from their original scale to the open unit interval $(0, 1)$.

Table 3 shows that the mean accuracy score is 0.900 for non-dyslexic readers and 0.606 for the dyslexic group. The scores ranged from 0.459 to 0.990, with the overall mean score of 0.773.

Table 3. Basic statistics of reading skills data.

	Accuracy	**Dyslexia**	**IQ**
Mean	0.7727616	-0.1363636	-2.272727e-05
Standard Deviation	0.1790063	1.002112	1.000064

Table 4. Parameter estimates of reading skills data.

Parameter	β_0	β_1	β_2	β_3
Estimate	1.334	-0.974	0.161	-0.219
p-value	0.0000	0.0000	0.2317	0.1049

Table 4 contains the estimates of β's, where the only covariate that is statistically significant at the usual nominal levels is the dyslexia status. This indicates an unexpected result: IQ makes little or no clear independent contribution. For details, see [25].

Espinheira et al. [10] computed and analyzed different residuals for the constant dispersion BRM using the reading accuracy data described above. They mentioned that the standardized weighted residual 2 was more successful in identifying influential observations. Espinheira et al. [11] used a Cook-like distance, called likelihood displacement. They identified observation 1 as atypical and showed that this observation is only slightly omitted in the index plot of the standardized weighted residuals. Here, we considered all the residuals mentioned above in Cook's distance in Equation (2.4) and display in Figure 1. We used $2 \times$ mean (Cook's distance), which was recommended by [11].

Table 5. Summary of outliers identified by different residuals using Cook's distance in reading skills data.

S. No.	**Residuals**	**Outliers**
1	Pearson	6, 8, 17, 19, 23, 24
2	Deviance	6, 7, 8, 14, 17, 18, 19, 20, 23, 24, 25
3	Response	6, 8, 17, 19, 23, 24
4	Weighted	6, 8, 15, 22
5	SWeighted	6, 8, 15, 22
6	Sweighted2	6, 8, 15, 22
7	Working	6, 8, 17, 19, 23, 24, 28, 38
8	Standardize	32, 33

Figure 1 reveals the high impact of the residual type on Cook distance statistics. Using different residuals, different observations are detected and the summary of all suspected outliers can be seen in Table 5, where the outliers detected by the class of weighted residuals are the same. Deviance residual is most sensitive as it detects many observations as outliers. Similarly, standardized residual detects only two observations as outliers. For this purpose, all alleged are excluded one by one and βs and MSE are calculated and presented in Table 6. The group deletion is made for all residuals and their βs and MSE are presented in Table 7.

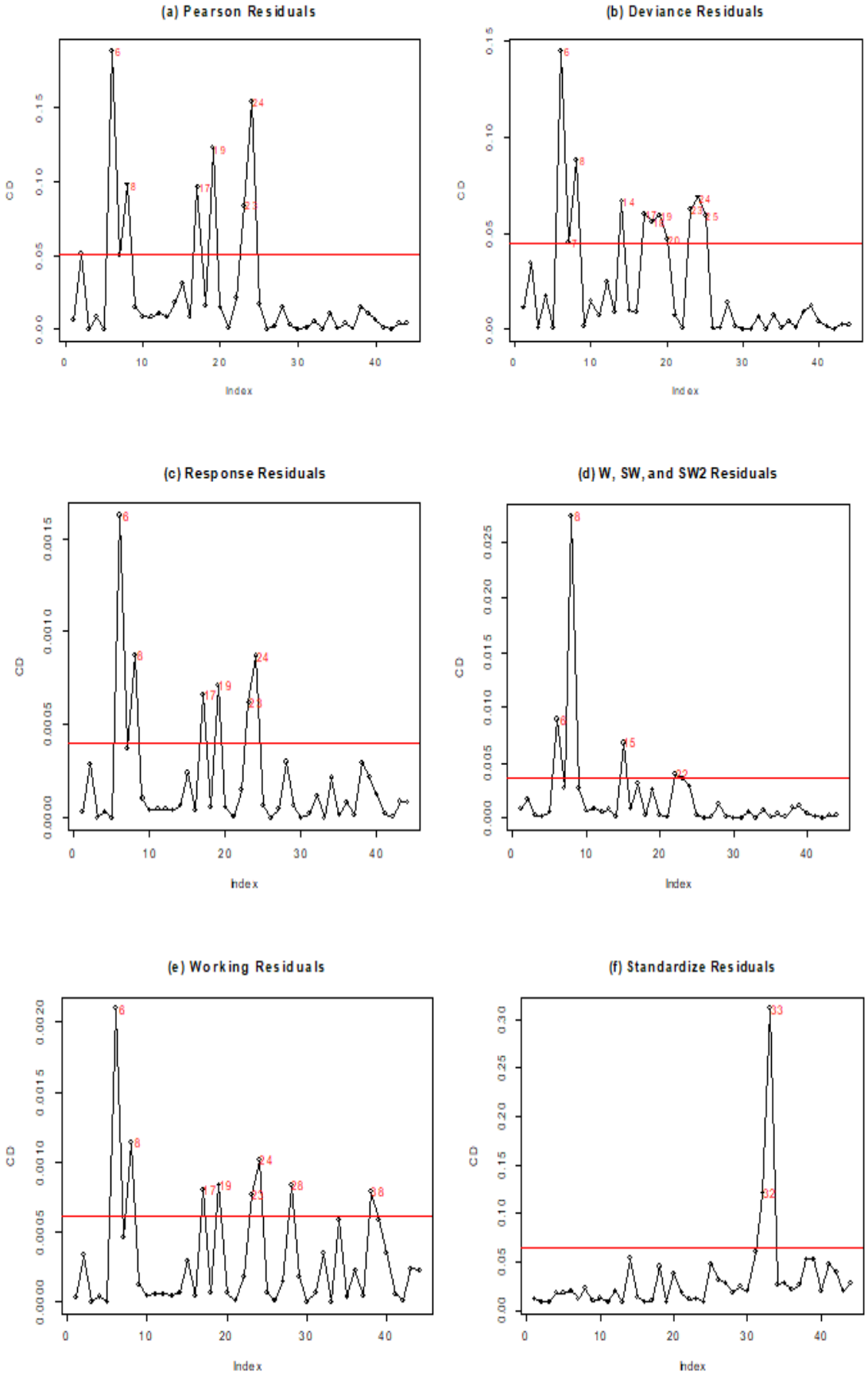

Figure 1. Real reading skills data Cook's distance with different residuals

Table 6. Parameter estimates, standard errors (S.E.), relative changes in estimates due to one-by-one exclusion, and respective p-values with MSE.

Obs		Intercept	Dyslexia	IQ	Dyslexia × IQ	MSE
Full	Estimate	1.3338	-0.9736	0.1608	-0.2186	0.0096
	S.E.	0.1357	0.1335	0.1344	0.1345	
	p-value	0.0000	0.0000	0.2317	0.1049	
6	Estimate	1.3922	-1.0303	0.1185	-0.1766	0.0093
	S.E.	0.1407	0.1385	0.1386	0.1387	
	p-value	0.0000	0.0000	0.3920	0.2030	
7	Estimate	1.3715	-1.0103	0.1412	-0.1991	0.0096
	S.E.	0.1388	0.1365	0.1355	0.1356	
	p-value	0.0000	0.0000	0.2970	0.1420	
8	Estimate	1.2541	-0.8916	0.2663	-0.3245	0.0090
	S.E.	0.1325	0.1306	0.1368	0.1369	
	p-value	0.0000	0.0000	0.0516	0.0178	

Table 7. Parameter estimates, standard errors (S.E.), relative changes in estimates due to group exclusions, and respective p-values with mean squared error (MSE).

Residuals		Intercept	Dyslexia	IQ	Dyslexia × IQ	MSE
Deviance	Estimate	1.7816	-1.4000	-0.0014	-0.0601	0.00394
	S.E.	0.1497	0.1488	0.1857	0.1857	
	p-value	0.0000	0.0000	0.994	0.746	
• Pearson • Response	Estimate	1.6490	-1.2692	0.1573	-0.2185	0.00421
	S.E.	0.1343	0.1333	0.1317	0.1318	
	p-value	0.0000	0.0000	0.2324	0.0974	
• Weighted • SWeighted • SWeighted2	Estimate	1.1386	-0.7707	0.3911	-0.4502	0.00803
	S.E.	0.1331	0.1320	0.1424	0.1425	
	p-value	0.0000	0.0000	0.0060	0.0016	
Working	Estimate	1.6674	-1.2592	0.1675	-0.2111	0.00368
	S.E.	0.1339	0.1328	0.1322	0.1323	
	p-value	0.0000	0.0000	0.2050	0.1110	
Standardized	Estimate	1.3498	-0.9392	0.1829	-0.1922	0.00995
	S.E.	0.1622	0.1598	0.1579	0.1579	
	p-value	0.0000	0.0000	0.2470	0.2240	

From Table 6, it is observed that elimination of single suspected outlier has no significant effect neither on MSE nor on estimates. So, it is better to deal with a group of observations that are identified as outliers. Although some observations are commonly detected by most residuals, Cook's distance detected '11' observations as outliers, that is, [6, 7, 8, 14, 17, 18, 19, 20, 23 24, 25]. The standardized residuals are the least delicate because they identify only two observations, i.e. [32, 33], which are not detected by any other residual, and also exclusion of these observations leads to an increase in MSE. Due to this, we stopped using the standardized residual further.

Cook's distance with Weighted, Sweighted and Sweighted2 residuals proposed by [10], detected same observations, i.e. 6, 8, 15, 22. The omission of these observations has no prominent effect on MSE but makes all covariates significant which are nonsignificant earlier, i.e. IQ and interaction of dyslexia and IQ. The detection of observations by using working residuals provides minimum MSE as compared to other residuals. So, the performance of working residual is best due to least MSE. Performance of Pearson and Response residuals are similar with Cook's distance; both types identified same observations as outliers [6, 8, 17, 19, 23, 24]. Elimination of such an observation has a significant impact on

MSE but has no effect on estimates.
It can be seen that the minimum MSE is obtained using the working residual, which detected '8' observations as outliers and the Deviance residual has the second minimum MSE among all residuals with the highest number of suspected outliers. Moreover, the use of the Deviance residual results in a negative coefficient of the variable 'IQ', which is positive in all other cases after eliminating the alleged outliers. Such a dramatic change in relation turned out that these observations have a strong influence on the data and the addition of such observations is recommended rather than removal [5].

4.2. Crude Oil Conversion Data

This empirical application is based on a data set from [37]. It has four explanatory variables; the first is the gravity of crude oil (x_1), which is measured using the index suggested by the American Petroleum Institute, and these variables measure the density of a liquid. Second, is the vapor pressure of the crude oil (x_2), and this variable is measured using the Reid vapour pressure defined as the pressure needed to keep the liquid from vaporizing at 100 degrees Fahrenheit. Third, the temperature (degrees Fahrenheit) at which 10 percent of crude oil has vaporized (x_3) and the temperature (degrees Fahrenheit) at which all gasoline is vaporized (x_4). The proportion of crude oil converted to gasoline after distillation and fractionation is a dependent variable (y).
Atkinson [38] used LRM to analyze this data set and examined that the error term is not symmetric and transformed the dependent variable. Then, Lemonte et al. [39] used this data set and considered that the dependent variable follows a beta distribution. Ferrari and Cribari-Neto [40] used the data for the detection of outliers and found observation 4 to be influential. The data set is also part of R package *betareg*.
Now, we consider this data set to examine the role of residuals in detection. Table 8 presents the basic statistics of the data considered, and Table 9 provides the estimates of β's, where x_3 and x_4 are statistically significant covariates.

Table 8. Basic statistics of crude oil conversion data

	Fractionation	**Gravity**	**Pressure**	**Temp10**	**Temp**
Mean	0.1965938	39.25	4.18125	241.5	332.0938
Standard Deviation	0.1072242	5.635429	2.61983	37.54138	69.75596

Table 9. Parameter estimates of crude oil conversion data.

Parameter	β_0	β_1	β_2	β_3	β_4
Estimate	-2.694942	0.004541	0.030413	-0.011045	0.010565
p-value	0.0000	0.524871	0.279117	0.0000	0.0000

We also consider all the residuals mentioned above in Cook's distance (4) and display them in Figure 2 with the same cut-off point used in the previous example. A summary of basic statistics and the alleged outliers can be examined in Tables 8 and 9, respectively, which shows a strong influence of residuals on detection. Figure 2 shows the suspected outliers that have been mentioned in Table 9. Following the previous pattern, all suspected are excluded one by one, and s and MSE are calculated. In addition, the group deletion is made for all residuals and their s and MSE are observed. Its related results are presented in Tables 10 and 11.

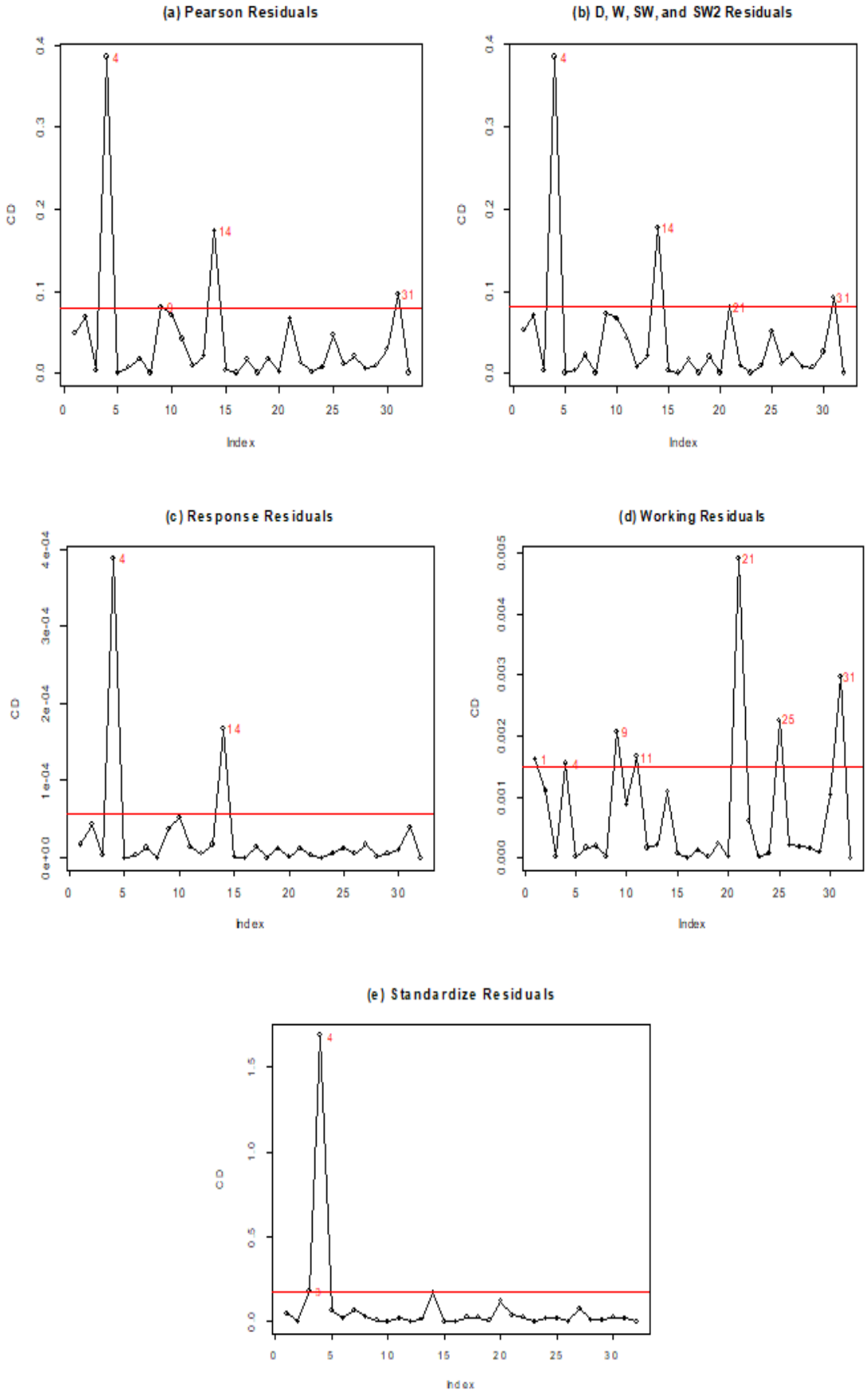

Figure 2. Real crude oil conversion data Cook's distance with different residuals

Table 10. Summary of outliers identified by different residuals using Cook's distance

S. No.	**Residuals**	**Outliers**
1	Pearson	4, 9, 14, 31
2	Deviance	4, 14, 21, 31
3	Response	4, 14
4	Weighted	4, 14, 21, 31
5	SWeighted	4, 14, 21, 31
6	Sweighted2	4, 14, 21, 31
7	Working	1, 4, 9, 11, 21, 25, 31
8	Standardize	3, 4

Table 11. Parameter estimates, standard errors (S.E.), relative changes in estimates due to one-by-one exclusion, and respective p-values with MSE

Obs		**Intercept**	**Gravity**	**Pressure**	**Temp10**	**Temp**	**MSE**
Full	Estimate	-2.694942	0.004541	0.030413	-0.011045	0.010565	0.0005665989
	SE	0.762569	0.007142	0.028101	0.002264	0.000515	
	p-value	0.0000	0.524871	0.279117	0.0000	0.0000	
1	Estimate	-2.549369	0.002959	0.024494	-0.011551	0.010735	0.0005819624
	SE	0.7772915	0.0073450	0.0287146	0.0023251	0.0005491	
	p-value	0.00104	0.68708	0.39365	0.0000	0.0000	
3	Estimate	-2.63853	0.00392	0.02887	-0.01114	0.01056	0.0005801207
	SE	0.808101	0.007697	0.029211	0.002333	0.000524	
	p-value	0.001090	0.610610	0.322930	0.000002	0.0000	
4	Estimate	-3.15597	0.00970	0.04017	-0.01060	0.01091	0.0005256371
	SE	0.780426	0.007420	0.027751	0.002207	0.000538	
	p-value	0.000053	0.191000	0.148000	0.000002	0.0000	
9	Estimate	-2.77136	0.00562	0.02617	-0.01118	0.01078	0.0005195174
	SE	0.723555	0.006787	0.026686	0.002145	0.000500	
	p-value	0.000128	0.407735	0.326684	0.000000	0.00000	
11	Estimate	-2.64444	0.00341	0.03493	-0.01084	0.01037	0.0005568741
	SE	0.750206	0.007054	0.027766	0.002229	0.000523	
	p-value	0.000424	0.628896	0.208412	0.000001	0.00000	
14	Estimate	-2.54809	0.00166	0.03288	-0.01163	0.01087	0.0004894554
	SE	0.723588	0.006872	0.026580	0.002156	0.000509	
	p-value	0.000429	0.808693	0.216150	0.000000	0.0000	
21	Estimate	-2.77954	0.00664	0.02867	-0.01061	0.01031	0.0005400972
	SE	0.722436	0.006839	0.026605	0.002150	0.000502	
	p-value	0.000119	0.331584	0.281205	0.000001	0.0000	
25	Estimate	-2.69799	0.00560	0.02666	-0.01088	0.01041	0.0005401782
	SE	0.735051	0.006926	0.027165	0.002183	0.000505	
	p-value	0.000242	0.418980	0.326315	0.000001	0.0000	
31	Estimate	-2.45230	0.00402	0.02424	-0.01200	0.01064	0.0005628977
	SE	0.758392	0.006983	0.027610	0.002290	0.000506	
	p-value	0.001220	0.564540	0.379980	0.000000	0.0000	

As we suggested earlier its better to deal with a group of observations that are identified as outliers than handle every single observation. Other results also support our previous findings. Firstly, some observations are commonly detected, too, by most of the residuals.

Secondly, the use of standardized residuals identifies only two observations (3, 4), no other residual has detected observation '3' as an outlier, and the detection of these observations has no significant effect either on coefficients or on MSE. Third, Cook's distance with Deviance, Weighted, SWeighted, and SWeighted2 residuals detected the same observations, i.e. [4, 14, 21, 31].
The detection of observations by using the Pearson residual provides a minimum MSE compared to other residuals. So, the performance of Pearson residual is best due to the least MSE. Performance of Response and Working residuals are similar in Cook's distance; surprisingly, both types identified different observations as outliers, but the elimination of such observations has a similar significant impact on MSE. It can be seen that the Working residual detects a maximum number of observations as outliers, but the elimination of such observation does not provide minimum MSE and the Working residual has maximum MSE among all residuals with a minimum number of suspected outliers.

Table 12. Parameter estimates, standard errors (S.E.), relative changes in estimates due to group exclusions and respective p-values with MSE

Residuals		**Intercept**	**Gravity**	**Pressure**	**Temp10**	**Temp**	**MSE**
Pearson	Estimate	-2.94378	0.00845	0.03562	-0.01242	0.01174	0.0003035964
	SE	0.608371	0.005747	0.021529	0.001767	0.000454	
	p-value	0.000001	0.142000	0.098000	0.000000	0.000000	
Response	Estimate	-3.09845	0.00753	0.04576	-0.01119	0.01138	0.0004000963
	SE	0.708165	0.006758	0.025270	0.002015	0.000519	
	p-value	0.000012	0.265500	0.070200	0.000000	0.000000	
• Deviance • Weighted • SWeighted • SWeighted2	Estimate	-2.90106	0.00855	0.03762	-0.01188	0.01126	0.0003626837
	SE	0.645177	0.006130	0.022838	0.001886	0.000489	
	p-value	0.000007	0.163300	0.099500	0.000000	0.000000	
Working	Estimate	-2.88642	0.01052	0.02417	-0.01122	0.01066	0.0004209662
	SE	0.629558	0.006076	0.022469	0.001851	0.000508	
	p-value	0.000005	0.083500	0.282000	0.000000	0.000000	
Standardize	Estimate	-3.40166	0.01244	0.04635	-0.01023	0.01100	0.0005381028
	SE	0.872704	0.008540	0.029576	0.002296	0.000562	
	p-value	0.000097	0.145000	0.117000	0.000008	0.0000	

5. Some concluding remarks

This paper considers Cook's distance for the BRM with different residuals. The BRM is used for a positively skewed continuous dependent variable. Comparisons of residuals with Cook's distance are assessed through a simulation study and by real data sets, which yielded important conclusions. First, Cook's distance for BRM can be helpful in determining the choice of residual related to the motive of the study. If the goal of the researcher is to observe the influence on estimators, then deviation is the best choice, which can change the relationship scenario. The use of this residual also reduces the MSE. Similarly, if purpose is to just reduce the mean squared error, then working and deviance residuals are the best choice with Cook's distance. Secondly, the percentage of the detection rate exhibits the performance for different sample sizes and for different values of the dispersion parameter. It reveals how the sample size and dispersion effect the detection of influential observations. Thirdly, the Cook distance presented in this paper may be useful in determining whether one should dummy the model in order to account for parameter non-constancy. The diagnostic measures presented in this article can help practitioners identify a typical observation and assess the specification of the model. In particular, a systematic relationship between influential observations and covariates is indicative of model misspecification.

6. Future recommendations

As mentioned in the previous section, the performance of Cook's distance is not worth for a small size with all considered residuals, so it is better to use any other diagnostic measure to detect outliers. Similarly, other residuals that might perform better than the existing ones may also be proposed, e.g., a class of adjusted residuals, a quantile residual.

Acknowledgements

We thank our respected reviewers for comments and suggestions that led to a much improved manuscript.

Author contributions. All the co-authors have contributed equally in all aspects of the preparation of this submission.

Conflict of interest statement. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. No author has received any funding.

Data availability. No personal data was used for the research described in the article.

References

[1] M. Amin, M. Amanullah, M. Aslam, and M. Qasim, *Influence diagnostics in gamma ridge regression model*, *J. Stat. Comput. Simul.* **89**, 536-556, 2019.

[2] T. Anholeto, C.M. Sandoval, and D.A. Botter, *Adjusted Pearson residuals in beta regression models*, *J. Stat. Comput. Simul.* **84**, 999-1014, 2014.

[3] A.C. Atkinson, *Two graphical displays for outlying and influential observations in regression*, *Biometrika* **68**, 13-20, 1981.

[4] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, *Stat. Sci.*, 379-393, 1986.

[5] S.W. Choi, *The effect of outliers on regression analysis, pp. regime type and foreign direct investment*, *Q. J. Polit. Sci.* **4**, 153-165, 2009.

[6] R.D. Cook and S. Weisberg, *Residuals and influence in regression*, New York: Chapman and Hall, 1982.

[7] R.D. Cook, *Detection of influential observation in linear regression*, *Technometrics* **42**, 65-68, 2000.

[8] F. Cribari-Neto and A. Zeileis, *Beta Regression in R*, *J. Stat. Softw.* **34**, 1-24, 2010.

[9] A. Davison, *Residuals and diagnostics*, *Stat. Theory. Mod.*, 83, 1991.

[10] P.L. Espinheira, S.L.P. Ferrari, and F. Cribari-Neto, *On beta regression residuals*, *J. Appl. Stat.* **35**, 407-419, 2008.

[11] P.L. Espinheira, S.L.P. Ferrari, and F. Cribari-Neto, *Influence diagnostics in beta regression*, *Comput. Statist. Data Anal.* **52**, 4417-4431, 2008.

[12] P.L. Espinheira, L.C.M. da Silva, A.O. Silva, and R. Ospina, *Model selection criteria on beta regression for machine learning*, *Mach. Learn. Knowl. Extr.* **1**, 26, 2019.

[13] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical distributions*, John Wiley & Sons, 2011.

[14] F. Silvia and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, *J. Appl. Stat.* **3**, 799-815, 2004.

[15] A.M. Garay, E.M. Hashimoto, E.M.M. Ortega, and V.H. Lachos, *On estimation and influence diagnostics for zero-inflated negative binomial regression models*, *Comput. Statist. Data Anal.* **55**, 1304-1318, 2011.

[16] J.W. Hardin, J.M. Hilbe, and J. Hilbe, *Generalized linear models and extensions*, Stata Press, 2007.
[17] K. Venezuela, M.D.A. Botter, and M.C. Sandoval, *Diagnostic techniques in generalized estimating equations*, *J. Stat. Comput. Simul.* **77**, 879-888, 2007.
[18] S. Liu, S.E. Ahmed, and L.Y. Ma, *Influence diagnostics in the linear regression model with stochastic linear restrictions*, *Pak. J. Statist.* **25**, 647-662, 2009.
[19] P. McCullagh and J.A. Nelder, *Generalized linear models*, London: Chapman and Hall, 1989.
[20] K. Pammer and A. Kevan, *The contribution of visual sensitivity, phonological processing, and nonverbal IQ to children's reading*, *Sci. Stud. Read.* **11**, 33-53, 2007.
[21] G.H.A. Pereira, *On quantile residuals in beta regression*, *Commun. Stat. Simul. Comput.* **48**, 302-316, 2019.
[22] D. Pregibon, *Logistic regression diagnostics*, *Ann. Statist.* **9**, 705-724, 1981.
[23] J.S. Preisser and B.F. Qaqish, *Deletion diagnostics for generalised estimating equations*, *Biometrika* **83**, 551-562, 1996.
[24] A.V. Rocha and A.B. Simas, *Influence diagnostics in a general class of beta regression models*, *Test* **20**, 95-119, 2011.
[25] M. Smithson and J. Verkuilen, *A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables*, *Psychol. Methods* **11**, 54, 2006.
[26] W. Thomas and R.D. Cook, *Assessing influence on regression coefficients in generalized linear models*, *Biometrika* **76**, 741-749, 1989.
[27] M.A. Ullah and G.R. Pasha, *The origin and developments of influence measures in regression*, *Pak. J. Stat.* **25**, 2009.
[28] D.A. Williams, *Generalized linear model diagnostics using the deviance and single case deletions*, *J. R. Stat. Soc. Ser. C. Appl. Stat.* **36**, 181-191, 1987.
[29] F.C. Xie and B.C. Wei, *Diagnostics analysis in censored generalized Poisson regression model*, *J. Stat. Comput. Simul.* **77**, 695-708, 2007.
[30] L. Xu, S.Y. Lee, and W.Y. Poon, *Deletion measures for generalized linear mixed effects models*, *Comput. Statist. Data Anal.* **51**, 1131-1146, 2006.
[31] H.T. Zhu and S.Y. Lee, *Local influence for incomplete data models*, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63**, 111-126, 2001.
[32] M. Qasim, K. Månsson, and B.M.G. Kibria, *On Some Beta Ridge Regression Estimators: Method, Simulation and Application*, *J. Stat. Comput. Simul.* **91**, 1699-1712, 2021.
[33] A.B. Simas, W. Barreto-Souza, and A.V. Rocha, *Improved estimators for a general class of beta regression models*, *Comput. Statist. Data Anal.* **54**, 48-66, 2010.
[34] C.M. Hurvich and C.L. Tsai, *Regression and time series model selection in small samples*, *Biometrika* **76** (2), 297-307, 1989.
[35] F. Cribari-Neto, J. J. Santana-e-Silva, and K. L. P. Vasconcellos, *Beta regression misspecification tests*, *J. Stat. Plan. Inference* **233**, 106193, 2024.
[36] J. A. Khan, A. Akbar, and B. M. G. Kibria, *Behavior of Residuals in Cook's Distance for Beta Ridge Regression Model (BRRM)*, *Int. J. Appl. Math. Comput. Sci. Syst. Eng.* **5**, 202-208, 2023.
[37] N.H. Prater, *Estimate gasoline yields from crudes*, *Pet. Refin.* **35**, 236238, 1956.
[38] A.C. Atkinson, *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, New York: Oxford University Press, 1985.
[39] A.J. Lemonte, S.L. Ferrari, and F. Cribari-Neto, *Improved likelihood inference in BirnbaumSaunders regressions*, *Comput. Stat. Data Anal.* **54**, 13071316, 2010.
[40] S. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, *J. Appl. Stat.* **31**, 799815, 2004.

[41] R. Ospina, P. L. Espinheira, L. A. Arias, C. M. Xavier, V. Leiva, and C. Castro, *New Statistical Residuals for Regression Models in the Exponential Family: Characterization, Simulation, Computation, and Applications*, *Mathematics* **12**(20), 3196, 2024. https://doi.org/10.3390/math12203196.