Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.1994

Research Article

Influence diagnostic in beta ridge regression model: Simulation and application

Javaria Ahmad KHAN[1,*], Atif AKBAR[1], B M Golam KIBRIA[2]

[1]Department of Statistics, Bahuddin Zakariya University, Multan, 60800, Pakistan
[2]Department of Mathematics and Statistics, Florida International University, Miami, FL 33199, USA

ARTICLE INFO

Article history
Received: 29 January 2024
Revised: 26 April 2024
Accepted: 31 October 2024

Keywords:
Beta Regression; Cook's Distance; Crude Oil Conversion; Influence Diagnostics; Multicollinearity; Outliers; Residuals; Ridge Regression; Shrinkage Parameter

ABSTRACT

When regressors exhibit a linear correlation in the Beta regression model (BRM), the Beta ridge regression model (BRRM) is employed to mitigate the impact of multicollinearity on maximum likelihood estimation (MLE) in the BRM. The traditional MLE approach is particularly sensitive to such correlations, leading to unstable and unreliable parameter estimates. To identify the influential data points, Cook's distance is utilized as a classic diagnostic tool to detect outliers. This paper emphasizes the dual challenges posed by multicollinearity and outliers, proposing a set of estimators for the shrinkage parameter in BRM using Cook's distance in combination with varied residual types. For illustrative purposes, Monte Carlo simulations and real data are presented.

The empirical results indicate that certain classes of weighted residuals (W, SW, ASW, and SW2) demonstrate superior performance in detecting outliers, with a high percentage of detection across different scenarios. Additionally, findings suggest that the outlier detection is not only influenced by the residual weighting scheme but is also inclined by the severity of multicollinearity and the selected value of the shrinkage parameter k. These results provide a baseline for further refinement of BRM, potentially in the selection of optimal shrinkage parameter and residual type, involving multicollinearity and contaminated data.

Cite this article as: Khan JA, Akbar A, Kibria BMG. Influence diagnostic in beta ridge regression model: Simulation and application. Sigma J Eng Nat Sci 2026;44(1):490−500.

INTRODUCTION

When multicollinearity occurs in a Beta regression model (BRM), the beta ridge regression model (BRRM) serves as a remedial tool in such a situation [1, 2]. In multiple regression, when the predictors are highly interrelated, multicollinearity occurs, which is one of the intense problems in regression analysis. Subsequently, the confidence interval becomes wider, making maximum likelihood estimates (MLE) unstable and inefficient [3], the variance of the MLE turns very high and such an estimator provides unreliable inference [4]. The literature documents several shrinkage and biased estimation strategies for handling multicollinearity, notably the modified ridge-type, the Liu and the Liu-type, the two-parameter, the Dawoud–Kibria, and the James–Stein type

***Corresponding author.**
*E-mail address: jakhan0@yahoo.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey

estimators [5–10]. These estimators aim to reduce the variance and improve estimation stability when multicollinearity challenges standard regression techniques. Many researchers examined their performance in different situations, like Taha [11] discussed the multiple diagnostic measures for multicollinearity and further analyzed the Poisson, logistic, negative binomial (NB), zero-inflated negative binomial (ZINB),and zero-inflated Poisson (ZIP) models. Estimation methods using maximum likelihood (ML), ridge, Liu, and Liu-type estimators. Zubair and Adenomon [12] compared different estimators for the linear regression model when multicollinearity is present with autocorrelation. Abonazel et al. [13] suggested a new estimator for binary data using a Probit regression model when multicollinearity is present in the data. Saputri et al. [14] also made a comparison among LASSO, MLE, and the Liu Estimator. When MLE methods are dealing with multicollinearity in the multinomial logistic regression model, and many others.

Correspondingly, the Extreme or anomaly value in the response/output variable results in an outlier while the predictor (input variable) results as an influential point. Outliers or Influential points are those data points that are far from nearby values and noticeably distinct from the remaining observations. Outliers can influence the coefficients of the regression model, distorting the relationship between variables, and also in terms of bias. Several established statistical methods exist for the identification of outliers, e.g., Cook's distance, index plots, and potential residual plots, see Hadi [15]. The presence of outliers is a crucial problem; usually, researchers suggest excluding such values, but exclusion can cause a loss of information [16]. A lot of literature is available, where studies focus only on detecting such observations with their influence. Some of them are: Sarkar et al. [17] performed an empirical study on a logistic regression model. Their study aims to identify potential outliers by using graphical methods and different measures of standardized residuals. Zakariya et al. [18] suggested a new estimator for the detection of outliers and named the Coefficient of Determination Ratio (CDR), which is then compared with some standard measures of influence, namely: Cook's distance, covariance ratio, studentised residuals, leverage values, and difference in fits standardized. Baba et al. [19] proposed a new detection method for outliers for the spatial regression model and compared that with classical methods like Cook's distance and some other robust methods. Similarly, Kumar et al [20] detect outliers by using a class of Cook's distance in survey-weighted linear regression and references their in.

Therefore, integrating detection and robust handling of both multicollinearity and outliers is crucial for reliable inference in regression models, supporting robust, precise estimations without compromising data integrity. Sinan and Alkan [21], Ibrahim and Yahya [22], Pati et al. [23], Majid et al. [24, 25], Arum et al. [26], Lukman et al. [27], and many others considered the multicollinearity and outlier problem concomitantly, but for the linear regression model (LRM). Pfaffenberger and Dielman [28] also discussed this combat problem in their book. They discussed the possible construction of a robust method that performs equally well in the presence of problems. For the generalized linear model (GLM), Arum et al. [29] examined this set of problems for the Poisson regression and Khan et al. [30] for BRRM. To the best of our knowledge, BRM has not been discussed to combat the problem of outliers in the existence of multicollinearity with different multiple shrinkage parameters.

This article will explain how residuals affect the stability and diagnostic performance of Cook's distance in BRRM under multicollinearity and their impact on coefficients and mean squared error (MSE). It will help to tackle the combined problem with the optimum type of residual in Cook's distance.

In the rest of this paper, Section 2 discusses the BRRM, Cook's distance with residuals, and biasing parameters. In Section 3, a Monte Carlo simulation study has been presented. An analysis of real-life data is presented in Section 4 for illustrative purposes, followed by concluding remarks in Section 5.

BETA REGRESSION MODEL

The BRRM, Cook's distance, and the considered residuals will be described in this section.

The Beta Ridge Regression Model

Consider the response vector $Y = (y_1, y_2, \dots, y_n)'$, where each y_i denotes an independent response (output) observation for $i = 1, 2, \dots, n$; which follows Beta (α, ϕ) distribution; where α is shape parameter and ϕ is scale parameter. The *pdf* of the Beta distribution is given as;

$$f(y;\alpha,\phi) = \frac{y^{\alpha-1}(1-y)^{\phi-1}}{B(\alpha,\phi)}, \quad y \in (0,1), \alpha > 0, \phi > 0 \qquad (1)$$

where $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\phi)}{\Gamma(\alpha+\phi)}$. The mean and variance of beta distribution are $\frac{\alpha}{\alpha+\phi}$ and $\frac{\alpha\phi}{(\alpha+\phi)^2(\alpha+\phi+1)}$, respectively with link function $g(\mu_i)=\eta_i=x_i^T\beta=logit(\mu_i)$. $x_i^T=(1, x_{i1}, \dots, x_{ip})'$ is the matrix of $(p+1)$explanatory/input variables and $\beta=(\beta_0, \beta_1, \dots, \beta_p)'$ is a vector of regression coefficients [31].

Ferrari and Cribari-Neto [32] and Espinheira et al. [33, 34], used the iterative reweighted least-squares (IRLS) algorithm in the Beta maximum likelihood (BML) estimator to estimate the parameter β, which is obtained as,

$$\hat{\beta}_{BML}=\left(X'\hat{W}X\right)^{-1}X'\hat{W}z,$$

where $z=\hat{\eta}+\hat{W}^{-1}\hat{T}\left(y^*-\mu^*\right)$ and $\hat{W}=diag\left(\hat{W}_1,\dots,\hat{W}_n\right)$, which are estimated in the final iteration.

In the multiple linear regression model, it is assumed that the explanatory variables are not associated. However, in real life, the explanatory variables may be linearly correlated, which causes the problem of multicollinearity [3]. To alleviate the impact of multicollinearity in the Beta regression model, Abonazel and Taha [1] and Qasim et al. [2] introduced the beta ridge regression (BRR) estimator as a substitute to the beta maximum likelihood (BML) estimator which is expressed as

$$\hat{\beta}_{BRR}=\left(X'\widehat{W}X+kI_p\right)^{-1}X'\widehat{W}z, \qquad k>0 \tag{2}$$

where k is known as the shrinkage or ridge or biasing parameter. The computation of the shrinkage or ridge parameter k is an important issue for the researcher, which will be discussed later.

Cook’s Distance and Residuals

To identify the outlier, Cook [35] proposed the Cook’s distance (CD) for the LRM, and Pregibon [36] utilized this approach for GLM. Cook’s distance (CD) quantifies the extent to which omitting the ith observation alters the estimated regression model. For the BRM, the CD statistic can be written as,

$$CD_i=\frac{\left(\hat{\beta}_{ML}-\hat{\beta}_{ML(i)}\right)' X'\widehat{W}X\left(\hat{\beta}_{ML}-\hat{\beta}_{ML(i)}\right)}{(k+1)\hat{\phi}}, \tag{3}$$

where, $\hat{\beta}_{ML}$ is the estimated BRM coefficients vector and $\hat{\beta}_{ML(i)}$ is the estimated BRM coefficients vector after deleting the ith observation. The following simplified form of equation (3) is given by McCullagh and Nelder [37].

$$CD_i=\frac{{\pi_i}^2}{(k+1)}\frac{h_{ii}}{1-h_{ii}}, \tag{4}$$

where, π_i is the ith residual, which is explained in Table 1. In the BRRM, 2*mean (Cook’s distance) [38] is used as the cutoff point for the detection of an outlier. The highest value of CD is a sign that the ith observation is the outlier. The following Table 1 presents the residuals that are considered in CD.

Choosing Biasing Parameter

Determining the ideal shrinkage/biasing parameter (k) is not possible but can only be estimated from the data. There are several methods to estimate the shrinkage parameter (see, e.g., [41, 42). Unlike linear regression, the effect of multicollinearity in GLM has not been significantly discussed in the literature using the ridge regression approach. Exceptionally, Månsson and Shukur [43] introduced some ridge parameters for the logistic regression model, and Algamal and Alanaz [44] compared different biasing parameters for Poisson ridge regression. Månsson and Shukur [45] established a Poisson ridge regression (PRR) estimator and used several biasing parameters. Since PRR can have a severe bias, Qasim et al. [46] suggested bias-adjusted PRR estimators. Amin et al. [47] examined the performance of inverse Gaussian ridge regression estimators. Lukman et al. [48] and Amin et al. [49] recommended some methods for estimating the shrinkage parameter based on the Gamma ridge regression. Khan et al. [30] used BRRM, but they selected ‘k’ randomly. Here, some well-known shrinkage estimators are given in Table 2.

Table 1. Summary of residuals and their notations

Sr. #	Residual	Notation	Reference
1	Pearson (P)	$r_t=\frac{y_t-\hat{f}(x_t)}{\sqrt{\hat{f}(x_t)}}$	Davison and Snell [39]
2	Deviance (D)	$r_t=sign\left(y_t-\mu_t\right)\sqrt{d_t}$	
3	Working (Wor)	$r_t=\frac{y_t-\hat{f}(x_t)}{\hat{f}(x_t)}$	
4	Response (R)	$r_t=y_t-\hat{f}(x_t)$	
5	Weighted (W)	$r_t^*=\frac{y_t^*-\hat{\mu}_t^*}{\sqrt{\phi v_t}},$	Espinheira et al. [33]
6	Standardized Weighted (SW)	$r_t^{\omega}=\frac{y_t^*-\hat{\mu}_t^*}{\sqrt{v_t}}$	
7	Standardized Weighted 2 (SW2)	$r_t^{\omega\omega}=\frac{r_t^{\omega}}{\sqrt{(1-h_{tt})}}$	
8	Adjusted Standardized Weighted (ASW)	$r_t^{\omega a}=\frac{r_i-\hat{E}(r_t^{\omega})}{\sqrt{\widehat{Var}(r_t^{\omega})}}$	Anholeto et al. [40]
9	Adjusted Pearson (AP)	$r_t^{a}=\frac{r_i-\hat{E}(r_i)}{\sqrt{\widehat{Var}(r_i)}}$	

Table 2. Summary of shrinkage estimators

Sr. #	*k*	Reference
1	$k_1 = \dfrac{(p-2)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	Thisted,[50]
2	$k_2 = \dfrac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	Hoerl and Kennard [51]
3	$k_3 = \dfrac{\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	Dwividi and Shrivastava [52]
4	$k_4 = Median\left\{\dfrac{\hat{\sigma}^2}{\hat{\alpha}_j{}^2}\right\}$	Kibria [41]
5	$k_5 = Median\left(\dfrac{1}{\sqrt{\dfrac{\hat{\sigma}_j{}^2}{\hat{\alpha}_j{}^2}}}\right)$	Muniz and Kibria [42]
6	$k_6 = Median\left(\dfrac{1}{\sqrt{\dfrac{\lambda max\hat{\sigma}_j{}^2}{(n-p)\hat{\sigma}_j{}^2+\lambda max\hat{\alpha}_j{}^2}}}\right)$	Muniz et al. [53]

SIMULATION STUDY

Simulation Technique

The following scheme is considered for the generation of simulated datasets:

I. The dependent variable of the BRM is generated from Beta distribution as $y_i = B(\alpha, \phi)$ for $i = 1, 2, \ldots, n$, where $\mu_i = E(y_i) = 3$ is arbitrary mean and $\phi = 5$ is dispersion parameter.

II. Two explanatory variables x_1 and x_2 are kept fixed throughout the whole simulation study. To introduce multicollinearity, Khan et al. [30] are followed. They used $x_{ij} = \sqrt{1-\rho^2}Z_{ij} + \rho Z_{ij}\ (p+1)$, $i = 1, 2, \ldots, n$; $j = 1, 2$ where Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4} are independent standard pseudo-random random numbers, and ρ is defined as the degree of multicollinearity, $\rho = 0.8, 0.9, 0.95$.

III. Then, we introduce the outliers in X_{ij} at 5th, 10th, 15th, 20th, and 25th point as $x_{ij} = a_0 + x_{ij}$, for $i = 5, 10, 15, 20, 25$; $j = 1, 2$, where $a_0 = \bar{x}_j + 100$.

IV. Sample size is considered $n = 25, 50, 100$, and 200 with 1000 replications.

V. The biasing parameter is chosen by using the methods given in Table 2.

VI. We used the detection rate as a percentage.

The simulated outlier detection rate in percentage of the BRM Cook's distance under consideration of different factors such as dispersion parameter, sample size,levels of multicollinearity, and different types of residuals, which are presented in Tables 3 to 8 for k_1, k_2, k_3, k_4, k_5, and k_6, respectively. For better insight, the average detection rate along with standard deviation for all methods over all n and ρ are obtained and presented at the end of Tables 3-8.

Table 3. Detection rate (%) of estimated outliers in the BRM with different residuals with k_1

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	**0.8**	22.2	22.0	26.0	22.1	24.1	24.1	24.3	23.0	20.6
	0.9	22.8	21.9	26.3	22.9	24.2	24.2	24.4	23.6	24.1
	0.95	25.0	24.1	26.1	25.1	24.6	24.6	26.3	25.1	22.8
	0.99	22.2	21.4	25.1	21.6	23.6	23.6	23.3	21.5	19.8
50	**0.8**	88.1	86.6	89.6	88.0	89.9	89.9	90.2	87.0	88.1
	0.9	88.1	87.8	88.4	86.9	88.1	88.1	88.8	86.2	87.1
	0.95	88.1	87.6	88.5	88.0	89.3	89.3	88.7	86.4	86.4
	0.99	85.6	85.6	86.4	85.6	87.9	87.9	88.0	84.3	85.1
100	**0.8**	99.0	99.1	98.9	99.1	99.4	99.4	99.2	98.7	98.7
	0.9	99.3	99.1	98.9	99.5	99.2	99.2	99.5	98.1	98.6
	0.95	99.2	99.1	99.3	99.3	99.9	99.9	99.8	99.5	98.7
	0.99	99.3	99.3	98.8	99.2	99.1	99.1	99.4	98.6	98.3
200	**0.8**	100	100	100	100	100	100	100	100.0	99.8
	0.9	100	100	100	100	100	100	100	100.0	100.0
	0.95	100	100	100	100	100	100	100	99.9	99.9
	0.99	100	100	100	100	100	100	100	100.0	99.8
Mean		77.4	77.1	78.3	77.3	78.1	78.1	78.2	77.0	76.7
SD		32.8387	33.090	31.609	32.8801	32.5043	32.50436	32.3332	32.52489	33.1884

Table 4. Detection rate (%) of estimated outliers in the BRM with different residuals with k_2

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	0.8	22.4	21.0	24.8	21.2	22.9	22.9	23.7	23.7	21.8
	0.9	20.8	18.4	23.7	21.6	23.7	23.7	22.2	21.8	19.3
	0.95	16.9	17.6	21.5	17.3	22.6	22.6	22.0	19.7	17.1
	0.99	15.7	16.7	19.0	15.0	20.5	20.5	19.4	18.3	15.9
50	0.8	88.5	85.1	89.1	88.9	89.7	89.7	89.4	87.7	86.0
	0.9	85.2	83.3	84.6	84.4	88.1	88.1	85.9	83.3	83.4
	0.95	84.0	82.7	84.7	83.6	86.6	86.6	85.7	83.3	82.1
	0.99	82.6	82.6	85.2	83.1	87.9	87.9	87.2	85.0	80.4
100	0.8	99.1	98.6	99.3	99.1	98.9	98.9	99.5	98.2	97.8
	0.9	99.1	98.7	99.1	99.2	99.7	99.7	99.4	98.8	98.1
	0.95	98.4	98.4	99.0	98.3	99.2	99.2	98.9	97.9	97.5
	0.99	98.4	98.0	98.3	98.7	99.5	99.5	98.8	98.8	97.2
200	0.8	99.9	100	99.9	100.0	100	100	100	99.9	99.6
	0.9	99.9	100	99.9	99.9	100	100	100	99.9	99.9
	0.95	99.8	100	100.0	99.8	100	100	100	99.9	99.6
	0.99	99.8	100	99.8	99.8	100	100	100	100.0	99.6
Mean		75.656	75.068	76.743	75.618	77.456	77.4562	77.006	76.012	74.706
SD		34.3913	34.4444	33.0259	34.4832	33.1877	33.1878	33.3449	33.4645	34.1913

Table 5. Detection rate (%) of estimated outliers in the BRM with different residuals with k_3

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	0.8	19.0	18.8	19.1	18.0	20.9	20.9	21.0	18.7	17.8
	0.9	18.3	17.9	18.8	17.3	21.3	21.3	20.7	19.0	16.6
	0.95	15.9	16.2	21.6	15.1	19.5	19.5	19.7	17.9	14.0
	0.99	14.1	14.5	22.5	15.9	19.9	19.9	18.8	17.9	16.8
50	0.8	85.2	85.7	86.3	86.3	87.3	87.3	88.1	84.5	84.2
	0.9	86.2	85.2	87.5	86.4	87.1	87.1	87.7	84.3	84.4
	0.95	86.4	86.1	86.6	86.8	87.9	87.9	87.1	84.4	84.5
	0.99	87.9	88.1	87.3	88.3	86.8	86.8	87.5	84.4	86.2
100	0.8	98.6	98.9	99.2	98.8	99.4	99.4	99.1	98.6	98.2
	0.9	98.7	99.0	99.4	98.6	99.1	99.1	99.1	98.7	98.4
	0.95	98.5	98.6	99.2	98.6	99.3	99.3	99.3	98.9	97.6
	0.99	99.0	98.9	98.6	98.9	98.7	98.7	98.6	98.2	98.8
200	0.8	99.9	99.9	99.7	99.9	99.8	99.8	99.7	99.5	99.7
	0.9	99.9	99.9	99.9	99.9	100.0	100.0	100.0	99.9	99.6
	0.95	99.9	99.9	99.9	99.8	99.8	99.8	99.8	99.8	99.6
	0.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8
Mean		75.468	75.475	76.6	75.537	76.675	76.675	76.637	75.293	74.7625
SD		35.4088	35.4087	33.8807	35.5526	33.9545	33.9545	34.1153	34.5127	35.3806

Table 6. Detection rate (%) of estimated outliers in the BRM with different residuals with k_4

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	**0.8**	14.9	6.5	20.5	14.3	23.2	23.2	17.6	22.4	11.9
	0.9	13.5	3.6	21.2	14.1	22.6	22.6	12.9	20.3	9.7
	0.95	11.3	4.4	18.3	12.9	21.9	21.9	13.2	20.2	7.7
	0.99	13.4	5.7	20.6	16.2	21.7	21.7	14.8	19.9	7.7
50	**0.8**	63.2	23.6	77.0	66.4	77.0	77.0	44.9	74.9	45.5
	0.9	57.0	18.7	76.8	59.5	75.3	75.3	40.8	71.6	39.4
	0.95	53.5	20.0	77.0	57.1	78.0	78.0	45.0	74.5	36.0
	0.99	50.8	29.9	77.6	55.0	81.0	81.0	51.7	76.9	35.5
100	**0.8**	77.4	46.4	86.0	78.4	83.2	83.2	65.6	82.0	68.4
	0.9	77.3	46.7	88.6	78.2	86.9	86.9	68.4	84.9	67.7
	0.95	71.1	41.9	85.3	73.2	84.4	84.4	63.3	82.3	57.7
	0.99	71.7	52.3	93.9	75.6	92.8	92.8	73.1	90.9	57.6
200	**0.8**	92.4	81.2	93.1	92.5	92.4	92.4	89.3	91.3	89.9
	0.9	83.4	73.6	87.0	83.9	86.6	86.6	81.3	84.7	78.8
	0.95	80.7	67.9	88.7	82.4	88.2	88.2	80.8	86.1	74.2
	0.99	82.9	76.0	94.5	84.4	94.7	94.7	90.4	93.7	74.9
Mean		57.156	37.4	69.131	59.006	69.368	69.36875	53.31875	67.2875	47.6625
SD		28.4823	27.2630	29.7933	28.5224	28.5714	28.57145	27.54289	28.45745	27.6908

Table 7. Detection rate (%) of estimated outliers in the BRM with different residuals with k_5

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	**0.8**	23.0	16.4	23.8	23.1	25.2	25.2	23.9	24.8	21.2
	0.9	23.7	15.1	26.1	23.4	25.0	25.0	24.1	24.6	20.7
	0.95	21.0	14.0	25.6	21.4	23.4	23.4	20.7	23.4	18.8
	0.99	17.4	8.8	23.5	21.1	25.0	25.0	19.7	23.9	12.1
50	**0.8**	83.5	69.1	88.7	85.0	91.0	91.0	86.4	89.1	74.7
	0.9	81.3	71.7	87.8	83.3	91.2	91.2	86.0	89.4	72.1
	0.95	79.9	69.1	86.0	82.7	89.6	89.6	85.6	85.4	74.1
	0.99	71.2	50.0	86.3	75.4	91.6	91.6	72.8	88.3	54.7
100	**0.8**	99.2	97.7	99.5	99.1	99.7	99.7	99.7	99.3	97.5
	0.9	98.6	97.5	98.9	98.7	99.3	99.3	99.0	98.5	96.2
	0.95	98.7	96.8	99.3	99.1	99.0	99.0	99.4	99.1	96.9
	0.99	97.2	92.0	99.2	98.1	99.8	99.8	98.7	99.3	92.7
200	**0.8**	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	99.5
	0.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0	99.7
	0.95	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.6
	0.99	99.8	99.0	100.0	99.9	100.0	100.0	100.0	100.0	99.2
Mean		74.6375	68.562	77.768	75.637	78.731	78.7312	75.9937	77.8062	70.6062
SD		33.0554	35.9579	32.0525	32.7649	32.4691	32.4691	33.0646	32.3556	33.8821

Table 8. Detection rate (%) of estimated outliers in the BRM with different residuals with k_6

n	ρ	P	AP	D	R	W	SW	ASW	SW2	Wor
25	**0.8**	18.5	12.0	20.5	17.5	23.7	23.7	22.4	20.6	14.8
	0.9	23.8	12.8	28.3	25.3	25.6	25.6	25.0	24.8	18.8
	0.95	22.4	13.8	25.0	22.2	27.0	27.0	22.6	26.8	19.4
	0.99	17.4	5.3	22.8	20.1	27.8	27.8	16.0	27.1	9.4
50	**0.8**	88.8	80.9	88.0	88.9	91.8	91.8	89.1	89.1	83.4
	0.9	86.9	79.3	88.1	88.1	89.9	89.9	86.5	87.8	81.7
	0.95	87.9	73.5	89.8	89.7	91.3	91.3	87.2	89.4	79.5
	0.99	71.7	50.8	86	76.6	91.6	91.6	74.5	89.2	57.8
100	**0.8**	98.5	98.6	99.1	99.1	99.2	99.2	99.2	98.7	97.0
	0.9	99.2	96.9	99.1	99.1	99.5	99.5	99.5	99.3	96.9
	0.95	97.1	94.6	97.9	97.4	98.9	98.9	98.5	97.6	93.9
	0.99	97.0	89.6	99.4	97.5	99.7	99.7	97.5	99.3	90.9
200	**0.8**	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	99.6
	0.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
	0.95	99.9	100.0	99.8	99.9	99.8	99.8	99.8	99.8	99.7
	0.99	99.8	99.5	99.9	100.0	100.0	100.0	99.9	99.9	99.5
Mean		75.55	69.218	77.731	76.3375	79.1125	79.1125	76.1062	78.0875	71.3875
SD		33.6581	37.1052	32.3560	33.4640	31.8705	31.8705	33.3501	32.09845	35.0324

RESULTS AND DISCUSSION

Tables 3-8 show the performance of Cook's distance in BRRM. From Tables 3-8, it can be seen that k_1 performed the best, followed by k_2, $k_{3,}$ and finally k_4 performed the worst in the sense of identifying outliers among 6 estimators. The performance of the shrinkage estimators (k_1 to k_6) also depends on the type of residuals. For example, k1 performed better for D Res in the sense of identifying outliers, while k_2-k_6 performed better for both W Res and SW residuals. It is interesting to note that the rate of identification of outliers is highly influenced by sample size; the percentage increases when the sample size increases and it detects all outliers when the sample size is very large. This behavior is almost the same for all the residuals.

By observing the results, it can be seen that ρ (levels of multicollinearity) affects the detection rate. When multicollinearity is high, it decreases the detection rate. This feature is clearer for a small sample size. It may be noticed that for a large sample size, all residuals performed almost equally well, but for a small sample size, W, SW, and ASW perform better.

In this study, SWeighted2 residual is one of the best options, which performs well with Cook's distance in the detection of outliers, according to the highest detection percentage. Also, Espinheira et al. [34] and Khan et al. [30] showed that the SWeighted2 residual is the best choice to be used in likelihood displacement (LD).

Application: Crude Oil Conversion Data

To examine the influence of outliers in real life; crude oil conversion data is used. Data is based on a data set from Prater [54]. It has four predictors, x_1,x_2,x_3, and $x_{4,}$with a dependent variable (y). This data set is used by Atkinson [55] and he applied LRM and observed that the error term is not symmetrical and transformed the dependent variable. Then, Lemonte et al. [56] considered that the dependent variable of this data follows a beta distribution. Ferrari and Cribari-Neto [32] used the data for the detection of outliers and found observation 4 as an influential. Then Qasim et al. [1] used the same data set and showed that there exist some multicollinearity issues, especially between variables x_2 and x_3. Khan et al. [30] considered this data with dual problems (outliers and multicollinearity) and detected several numbers of observations as outliers.

Now, we used that data to consider both problems and examine the impact of different residuals in Cook's distance and then their influence on the co-efficient (s). The Beta ridge regression model is then used to determine the considered residuals mentioned in Table 1.

Different shrinkage parameters are used for the detection of outliers which are presented in Table 3. It can be observed that P, AP, and D residuals detect almost the same residuals with any biasing parameter. Similarly, several outliers detected by the class of weighted residual are the same. Working residual detects a maximum number of outliers.

Another important feature of the result is that there is no impact of k in the model. Any type of mentioned k has no influence on the detection rate or on observations. Similar results are also found in Khan et al. [21], where they choose k randomly between 0-1, and no value has the power to detect outliers more firmly.

Next, the influence of these outliers is examined on the model. To generate Table 9, firstly, detected outliers are eliminated one by one from the model, and then combinedly, as detected by residuals. For this purpose, coefficients and MSE are considered. Results are available in Table 10, where it is found that no observation has as much influence, which makes the major change in results. Elimination of observation 7 makes the coefficient of gravity smaller. Observation 14 is one of those observations that are detected by almost every residual as an outlier. It is evident that the elimination of that observation has a major impact on MSE. Exclusion of observation 14, minimizes the MSE. Due to the deletion of observation 1, the MSE of the model increases; even the MSE is greater than the full model MSE.

Table 9. Outliers detected using different k

k / Res	k_1	k_2	k_3	k_4	k_5	k_6
P	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	2, 14, 31	1, 2, 14, 31	1, 2, 14, 31
AP	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31
D	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31	1, 2, 14, 31
Wor	1, 2, 11, 21, 25, 31	1, 2, 11, 21, 25, 31	1, 2, 11, 21, 25, 31	1, 11, 21, 25, 31	1, 2, 11, 21, 25, 31	1, 2, 11, 21, 25, 31
R	2, 3, 7, 14, 31	2, 3, 7, 14, 31	2, 3, 7, 14, 31	2, 14, 31	2, 3, 7, 14, 31	2, 3, 7, 14, 31
W	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	2, 11, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31
SW	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	2, 11, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31
ASW	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31
SW2	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31	1, 2, 14, 21, 31

Table 10. Parameter estimates and relative changes in estimates due to one-by-one exclusion and mean squared error (MSE)

	b_1	b_2	b_3	b_4	MSE
Complete (All)	-0.01220	-0.04003	-0.01830	0.01060	0.00080
1	-0.01334	-0.04337	-0.01847	0.01086	0.00082
2	-0.01401	-0.04304	-0.01828	0.01079	0.00076
3	-0.01360	-0.04066	-0.01807	0.01057	0.00078
7	-0.00951	-0.04932	-0.01874	0.01074	0.00075
11	-0.01305	-0.03368	-0.01793	0.01038	0.00079
14	-0.01438	-0.03308	-0.01853	0.01095	0.00069
21	-0.01099	-0.04309	-0.01811	0.01039	0.00079
25	-0.01134	-0.04350	-0.01813	0.01045	0.00078
31	-0.01092	-0.03931	-0.01871	0.01069	0.00076
P	-0.01669	-0.03937	-0.01915	0.01162	0.00059
AP					
D					
Wor	-0.01257	-0.04647	-0.01814	0.01058	0.00072
R	-0.01375	-0.04477	-0.01908	0.01135	0.00052
W	-0.01570	-0.04129	-0.01899	0.01143	0.00059
SW					
ASW					
SW2					

After that, all detected outliers are eliminated regarding residuals. Results showed that the removal of outliers detected by response residual provides the minimum MSE, as those outliers are also quite different than others.

CONCLUSION

This paper explores Cook's distance for thebeta ridge regression model with multiple residuals and different estimators of the shrinkage parameter k. The comparison of residuals using Cook's distance for outlier detection is assessed through a Monte Carlo simulation study and a real-lifedataset, leading to important conclusions. Firstly, for theBRRM,Cook's distance can assist in determining the choice of residuals based on the study's objectives. If the researcher's goal is outlier detection, then a class of weighted residuals is the optimal choice. Detection probabilities are higher for larger sample sizes. Additionally, it is noticed that the outlier detection is affected by the choice of the ridge parameter and influenced by the level of multicollinearity. Secondly, eliminating all outliers detected by response residuals results in a minimum mean squared error. Since the detection of outliers depends on the estimation of the shrinkage parameter k, and we only consider six estimators, it will be interesting to consider all available estimators in the literature.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Qasim M, Månsson K, Kibria BMG. On some beta ridge regression estimators: Method, simulation and application. J Stat Comput Simul 2021;91:1699–1712. [CrossRef]

[2] Abonazel MR, Taha IM. Beta ridge regression estimators: Simulation and application. Commun Stat Simul Comput 2023;52:4280–4292. [CrossRef]

[3] Frisch R. Statistical confluence analysis by means of complete regression systems. Oslo: Universitetets Økonomiske Instituut; 1934.

[4] Qasim M, Kibria BMG, Månsson K. A new Poisson Liu regression estimator: Method and application. J Appl Stat 2020;47:2258–2271. [CrossRef]

[5] Swindel BF. Good ridge estimators based on prior information. Commun Stat Theory Methods 1976;5:1065–1075. [CrossRef]

[6] Kejian L. A new class of biased estimate in linear regression. Commun Stat Theory Methods 1993;22:393–402. [CrossRef]

[7] Kejian L. Using Liu-type estimator to combat collinearity. Commun Stat Theory Methods 2003;32:1009–1020. [CrossRef]

[8] Özkale MR, Kaciranlar S. The restricted and unrestricted two-parameter estimators. Commun Stat Theory Methods 2007;36:2707–2725. [CrossRef]

[9] Abonazel MR, Dawoud I, Awwad FA. Dawoud–Kibria estimator for beta regression model: Simulation and application. Front Appl Math Stat 2022;8:775068. [CrossRef]

[10] Seifollahi S, Bevrani H. James–Stein type estimators in beta regression model: Simulation and application. Hacet J Math Stat 2023;1–20.

[11] Taha IMIM. Handling multicollinearity problem in generalized linear models. Cairo: Cairo University; 2019. Doctoral dissertation.

[12] Zubair MA, Adenomon MO. Comparison of estimators efficiency for linear regressions with joint presence of autocorrelation and multicollinearity. Sci World J 2021;16:103–109.

[13] Abonazel MR, Dawoud I, Awwad FA, Tag-Eldin E. New estimators for the probit regression model with multicollinearity. Sci Afr 2023;19:e01565. [CrossRef]

[14] Saputri SA, Herawati N, Sutrisno A, Nusyirwan, Nisa K. Comparison of MLE, LASSO, and Liu estimator methods to overcome multicollinearity in multinomial logistic regression: Simulation study. Glob Sci Acad Res J Multidiscip Stud 2024;3:8–12.

[15] Hadi AS. A new measure of overall potential influence in linear regression. Comput Stat Data Anal 1992;14:1–27. [CrossRef]

[16] Khan JA, Akbar A. Empirical performance of nonparametric regression over LRM and IGRM addressing influential observations. J Chemom 2019;33:e3143. [CrossRef]

[17] Sarkar SK, Midi H, Rana S. Detection of outliers and influential observations in binary logistic regression: An empirical study. J Appl Sci 2011;11:26–35. [CrossRef]

[18] Zakaria A, Howard NK, Nkansah BK. On the detection of influential outliers in linear regression analysis. Am J Theor Appl Stat 2014;3:100–106. [CrossRef]

[19] Baba AM, Midi H, Adam MB, Abd Rahman NH. Detection of influential observations in spatial regression model based on outliers and bad leverage classification. Symmetry (Basel) 2021;13:2030. [CrossRef]

[20] Kumar R, Biswas A, Singh D, Ahmad T. Detection of outliers in survey-weighted linear regression. Math Popul Stud 2024;31:147–164. [CrossRef]

[21] Sinan A, Alkan BB. A useful approach to identify the multicollinearity in the presence of outliers. J Appl Stat 2015;42:986–993. [CrossRef]

[22] Ibrahim SA, Yahya WB. Effects of outliers and multicollinearity on some estimators of linear regression model. In: Proceedings of the 1st International Conference; 2017. p.204–209.

[23] Pati KD. Using standard error to find the best robust regression in presence of multicollinearity and outliers. In: 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE; 2020. p.266–271. [CrossRef]

[24] Majid A, Aslam M, Ahmad S. Robust estimation of the distributed lag model with multicollinearity and outliers. Commun Stat Simul Comput 2022;1–15.

[25] Majid A, Ahmad S, Aslam M. A robust Kibria–Lukman estimator for linear regression model to combat multicollinearity and outliers. Concurr Comput Pract Exp 2023;35:e7533. [CrossRef]

[26] Arum KC, Ugwuowo FI, Oranye HE. Combating outliers and multicollinearity in linear regression model using robust Kibria–Lukman mixed with principal component estimator: Simulation and computation. Sci Afr 2023;e01566. [CrossRef]

[27] Lukman AF, Farghali RA, Kibria BMG. Robust–Stein estimator for overcoming outliers and multicollinearity. Sci Rep 2023;13:9066. [CrossRef]

[28] Pfaffenberger RC, Dielman TE. A comparison of regression estimators when both multicollinearity and outliers are present. In: Robust regression. Routledge; 2019. p.243–270. [CrossRef]

[29] Arum KC, Ugwuowo FI, Oranye HE. Robust modified jackknife ridge estimator for the Poisson regression model with multicollinearity and outliers. Sci Afr 2022;e01386. [CrossRef]

[30] Khan JA, Akbar A, Kibria BMG. Behavior of residuals in Cook's distance for beta ridge regression model (BRRM). Int J Appl Math Comput Sci Syst Eng 2023;5:202–208. [CrossRef]

[31] Forbes C, Evans M, Hastings N. Statistical distributions. Hoboken (NJ): John Wiley & Sons; 2011. [CrossRef]

[32] Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat 2004;31:799–815. [CrossRef]

[33] Espinheira PL, Ferrari S, Cribari-Neto F. On beta regression residuals. J Appl Stat 2008;35:407–419. [CrossRef]

[34] Espinheira PL, Ferrari S, Cribari-Neto F. Influence diagnostics in beta regression. Comput Stat Data Anal 2008;52:4417–4431. [CrossRef]

[35] Cook RD. Detection of influential observation in linear regression. Technometrics 1977;19:15–18. [CrossRef]

[36] Pregibon D. Logistic regression diagnostics. Ann Stat 1981;9:705–724. [CrossRef]

[37] McCullagh P, Nelder JA. Generalized linear models. London: Chapman & Hall; 1989. [CrossRef]

[38] Hardin JW, Hilbe JM. Generalized linear models and extensions. College Station (TX): Stata Press; 2007.

[39] Davison AC, Snell EJ, editors. Residuals and diagnostics. In: Statistical theory and modelling. London: Chapman & Hall; 1991.

[40] Anholeto T, Sandoval MC, Botter DA. Adjusted Pearson residuals in beta regression models. J Stat Comput Simul 2014;84:999–1014. [CrossRef]

[41] Kibria BMG. Performance of some new ridge regression estimators. Commun Stat Simul Comput 2003;32:419–435. [CrossRef]

[42] Muniz G, Kibria BMG. On some ridge regression estimators: An empirical comparison. Commun Stat Simul Comput 2009;38:621–630. [CrossRef]

[43] Månsson K, Shukur G. On ridge parameters in logistic regression. Commun Stat Theory Methods 2011;40:3366–3381. [CrossRef]

[44] Algamal ZY, Alanaz MM. Proposed methods in estimating the ridge regression parameter in Poisson regression model. Electron J Appl Stat Anal 2018;11:506–515.

[45] Månsson K, Shukur G. A Poisson ridge regression estimator. Econ Model 2011;28:1475–1481. [CrossRef]

[46] Qasim M, Månsson K, Amin M. Biased adjusted Poisson ridge estimators: Method and application. Iran J Sci Technol Trans A Sci 2020;44:1775–1789. [CrossRef]

[47] Amin M, Afzal S. New ridge estimators in the inverse Gaussian regression: Monte Carlo simulation and application to chemical data. Commun Stat Simul Comput 2022;51:6170–6187. [CrossRef]

[48] Lukman AF, Ayinde K, Kibria BMG. Modified ridge-type estimator for the gamma regression model. Commun Stat Simul Comput 2020;51:1–15. [CrossRef]

[49] Amin M, Qasim M, Amanullah M. Performance of some ridge estimators for the gamma regression model. Stat Pap 2020;61:997–1026. [CrossRef]

[50] Thisted RA. Ridge regression, minimax estimation, and empirical Bayes methods. Technical report no. 28. Stanford (CA): Division of Biostatistics, Stanford University; 1976.

[51] Hoerl AE, Kennard RW. Ridge regression: Biased estimation of nonorthogonal problems. Technometrics 1970;12:55–67. [CrossRef]

[52] Dwividi TD, Shrivastava VK. On the minimum mean square error estimators in a regression model. Commun Stat Theory Methods 1978;7:487–494. [CrossRef]

[53] Muniz G, Kibria BMG, Månsson K, Shukur G. On developing ridge regression parameters: A graphical investigation. SORT 2012;36:115–138.

[54] Prater NH. Estimate gasoline yields from crudes. Pet Refin 1956;35:236–238.

[55] Atkinson AC. Plots, transformations and regression: An introduction to graphical methods of diagnostic regression analysis. New York: Oxford University Press; 1985.

[56] Lemonte AJ, Ferrari SL, Cribari-Neto F. Improved likelihood inference in Birnbaum–Saunders regressions. Comput Stat Data Anal 2010;54:1307–1316. [CrossRef]