Decision Analytics Journal 14 (2025) 100557

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

Advanced outlier detection methods for enhancing beta regression robustness

Oktsa Dwika Rahmashari, Wuttichai Srisodaphol *

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand

ARTICLE INFO

Keywords:
Beta regression
Outlier detection
Tukey–Pearson residuals
Parameter estimation
Model reliability
Statistical techniques

ABSTRACT

Beta regression is a valuable statistical technique for modeling response variables within the standard unit interval (0, 1), where values represent rates, proportions, or probabilities. However, outliers in beta regression can severely impact parameter estimates and model performance, leading to predicted values that deviate significantly from actual observations. Detecting and managing these outliers is essential to ensure model reliability and accuracy. In this study, we propose three novel outlier detection methods: Tukey-Pearson Residual (TPR), Iterative Tukey-Pearson Residual (ITPR), and Iterative Tukey-MinMax Pearson Residual (ITMPR). These methods integrate the principles of Tukey's boxplot with Pearson residuals, providing robust frameworks for detecting outliers in beta regression models. Extensive simulation studies and real-world data applications were conducted to evaluate their performance against existing outlier detection techniques in the literature. The results indicate that the ITPR method achieves the highest levels of precision and reliability, making it the most effective among the proposed methods. The TPR and ITMPR methods also exhibit strong performance, closely aligning with existing techniques. These findings highlight the potential of the proposed methods to enhance the robustness of beta regression analysis and its practical applications.

1. Introduction

The beta distribution is well-known for its flexibility as a probabilistic model, particularly when applied to constrained data within an open interval. Its primary application lies in modeling data on the standard unit interval (0, 1), where values are often interpreted as rates or proportions. Building on previous contributions from statisticians, various regression methodologies have been developed to harness the unique properties of the beta distribution. Notable contributions include methodological advancements that focus on modeling beta-distributed response variable [1,2] and the influential regression framework known as beta regression, proposed by Ferrari and Cribari-Neto [3]. The beta regression model employs the beta distribution to relate the mean and precision of the response variable to independent variables, following a framework similar to the generalized linear models developed by McCullagh and Nelder [4]. Beta regression has been used in various fields where analyzing proportions or rates is important, such as in medical research [5–7], pathology [8,9], natural sciences [10], and social sciences [11–13].

Outliers frequently appear in real-world data across various fields [14,15], and a key concern in empirical research is whether these outlying observations might undermine the validity of the findings [16]. Outliers can be found in regression analysis, as various variables are being modeled and analyzed. These are data points that deviate from the overall pattern of the data and can significantly impact regression estimators [17,18]. In beta regression, the presence of outliers can significantly distort model predictions and reduce overall reliability. Yellareddygari et al. [8] highlighted this issue in their study on postharvest pink rot development, where the predicted probabilities from beta regression closely matched the true probabilities, except for a few observations identified as outliers. For these outliers, the predicted values deviated substantially from the actual responses, underscoring the sensitivity of beta regression to atypical data points due to the bounded nature of the response variable within the interval (0, 1).

Accurate detection of outliers is essential to minimize their impact, ensuring robust model estimation and reliable inference. Residual analysis, which examines the differences between observed and predicted values, is a fundamental technique for identifying potential outliers. While raw response residuals are often unsuitable due to heteroskedasticity [19], standardized (or Pearson) and deviance residuals in beta regression provide more reliable diagnostic tools [3]. The literature highlights various methods for addressing outliers in beta regression, including residual analysis and Cook's distance [3]. Serdahl [20] introduced the rules of thumb method, suggesting that standardized

* Corresponding author.
E-mail address: wuttsr@kku.ac.th (W. Srisodaphol).

https://doi.org/10.1016/j.dajour.2025.100557
Received 30 December 2024; Received in revised form 12 February 2025; Accepted 22 February 2025
Available online 25 February 2025

residuals with absolute values over 2 in regression analysis may indicate potential outliers. Espinheira et al. [21] applied the rules of thumb method to compare their proposed residuals in beta regression. Muñoz-Pichardo et al. [22] introduced a Jackknife-after-Bootstrap transformation for case-deletion diagnostics in beta regression, combining jackknife and bootstrap resampling techniques. This method, referred to as Jackknife-after-Bootstrap Cook's distance, builds upon the approaches proposed by [23,24], demonstrating its effectiveness as a complement to traditional techniques like Cook's distance in beta regression.

Despite their utility, the rules of thumb method is not specifically tailored for beta regression, as it relies on normal distribution assumptions and does not account for the unique characteristics of beta regression, such as asymmetrical error distributions and parameter constraints within the (0, 1) range. Similarly, the Jackknife-after-Bootstrap Cook's distance depends on fixed cut-off thresholds, which may not fully capture the complexity of beta regression models. To address these limitations, this research introduces novel methods for detecting outliers in beta regression, leveraging Tukey's boxplot by [25] in conjunction with Pearson residuals. Pearson residuals standardize the variance of the response variable, making it easier to identify observations that deviate significantly from the model's predictions. They have been widely used for developing outlier detection methods in regression models, including gamma regression models [26] and logistic regression models [27,28], and have shown good performance in identifying groups of outliers. Tukey's boxplot effectively summarizes these residuals, enabling the detection of outliers based on the interquartile range (IQR) without relying on specific distributional assumptions. Our study extends the use of Pearson residuals and Tukey's boxplot to beta regression, leading to more accurate outlier detection and enhanced model reliability.

Our contributions are:

- Introducing novel methods for detecting outliers in beta regression by utilizing Tukey's boxplot and Pearson residuals.
- Assessing the performance of our approach by comparing it to the rules of thumb and Jackknife-after-Bootstrap Cook's distance (JaB-Cook's distance) methods to evaluate its effectiveness.

Section 2 presents the materials and methods, which include a review of the relevant literature and a detailed outline of the proposed methods. Section 3 reports the results of simulations and real data applications to demonstrate the effectiveness of the methods. Section 4 provides a discussion of the findings, including their implications and comparisons with existing approaches. Finally, Section 5 concludes the research by summarizing the key findings.

2. Materials and methods

2.1. Related work

2.1.1. Beta regression

Beta regression is a statistical technique specifically designed for modeling response variables restricted to the interval (0, 1) [3]. The model assumes that the response variable follows a beta distribution, which is characterized by the mean (μ) and the precision parameter (ϕ), denoted as $Y \sim \text{Beta}(\mu, \phi)$. The beta probability density function is defined for a random variable Y that follows a beta distribution [3]. It is given by:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0 < y < 1, \quad (1)$$

with $0 < \mu < 1$ and $\phi > 0$. The mean and variance of Y, which follows a beta distribution are given as:

$$E(Y) = \mu, \quad (2)$$

$$VAR(Y) = \frac{\mu(1-\mu)}{(1+\phi)}, \quad (3)$$

where for a fixed μ, a higher ϕ corresponds to a smaller variance in y. The precision parameter may also be influenced by external explanatory variables or latent factors [19,29]. To incorporate the effect of independent variables, the model employs the link function, which relates μ_i to the independent variables as

$$g(\mu_i) = x_i^T \beta = \eta_i, \quad (4)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a $(k+1)\times 1$ vector of unknown regression parameter $(k < n)$,$x_i = (1, x_{i1}, \ldots, x_{ik})^T$ is the vector of k independent variables, and $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ is the linear combination of the independent variables, weighted by their respective coefficients. As an alternative, the Pearson residual in beta regression was proposed and is defined as [3]

$$r_i = \frac{y_i - \mu_i}{\sqrt{VAR(y_i)}}, \quad (5)$$

for $i = 1, 2, \ldots, n$. A well-fitting model should show no detectable pattern when plotting the Pearson residuals against the observation index, while a detectable trend in the plot of r_i against $\hat{\eta}_i$ may indicate link function misspecification [3].

2.1.2. Tukey's boxplot

Tukey [25] introduced the boxplot, a graphical method for visually understanding data distribution. To identify outliers using a boxplot, the lower and upper fences are first calculated using the following formula

$$Lower\ fence = Q_1 - (1.5IQR), \quad (6)$$

$$Upper\ fence = Q3 + (1.5IQR). \quad (7)$$

Any data points falling outside these fences are considered outliers. The boxplot has been effectively utilized in outlier detection across various fields, such as in the classification task of data mining where prior outlier preprocessing is necessary [30]. It has also been applied in detecting outliers in multivariate data by filtering observations into normal data and potential outliers [31], as well as in classifying outliers in big functional data by separating outlier indices from typical observations [32]. The IQR is also applied in regression, such as in identifying leverage points in soil data with quantile regression forests [33], ensuring stable inliers for reference points.

2.1.3. Performance evaluation

Three performance measurements are used to evaluate outlier detection methods: the probability of successfully detecting all true outliers (*pout*), the probability of the masking effect (*pmask*), and the probability of the swamping effect (*pswamp*). These measurements were introduced by [34] and have been employed in several studies to assess the effectiveness of outlier detection methods [35,36]. Each performance measurement's formula is defined as

$$pout = \frac{n_{success}}{n_{out}}, \quad (8)$$

$$pmask = \frac{n_{failure}}{n_{out}}, \quad (9)$$

$$pswamp = \frac{n_{false}}{n - n_{out}}, \quad (10)$$

where $n_{success}$ is the number of observations that are successfully detected as outliers, $n_{failure}$ is the number of outliers are identified as inliers, n_{false} is the number of inliers that are identified as outliers, n_{out} is the total number of outliers, and n is the total number of observations. An outlier detection method is considered good if *pout* is close to one, while the values of *pmask* and *pswamp* are close to zero.

2.2. Proposed methods

This section presents the three proposed methods for detecting outliers using Tukey's boxplot and Pearson residual.

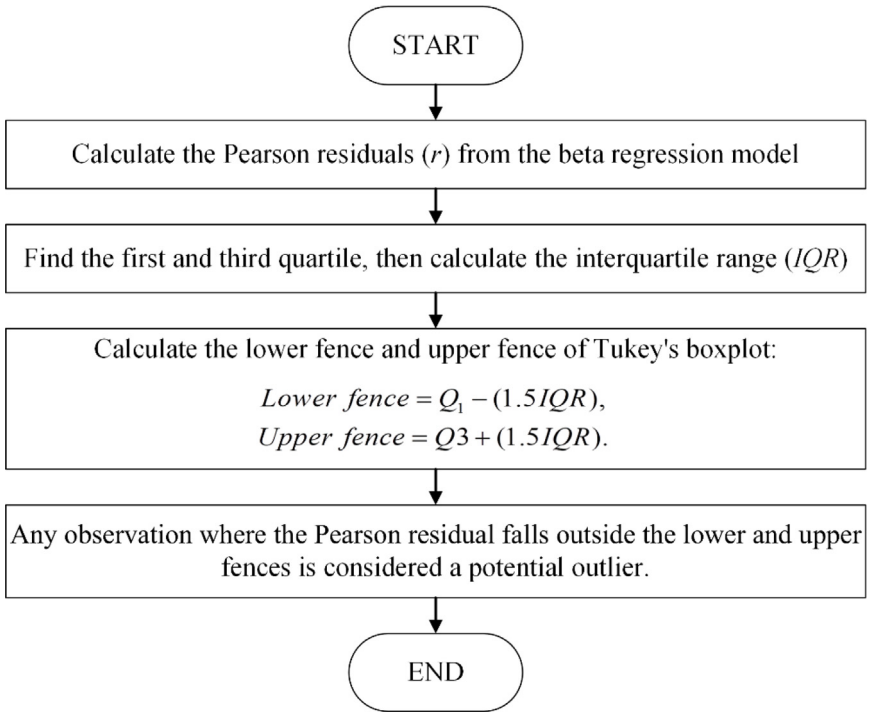

Fig. 1. Concept of TPR method.

2.2.1. Tukey-pearson residual (TPR) method

The first proposed method uses Tukey's boxplot along with Pearson residual to identify outliers, which is referred to as the Tukey-Pearson residual method (or, as we named it, the TPR method). The concept of this method is shown in Fig. 1.

2.2.2. Iterative Tukey-Pearson residual (ITPR) method

The second proposed method is an extension of the TPR method, which is referred to as the Iterative Tukey-Pearson residual (ITPR) method. This method uses Tukey's boxplot to separate the Pearson residuals into two groups: Group 1 (G_1) and Group 2 (G_2), allowing the calculation of a cut-off point based on the members of G_1. The concept of this method is shown in Fig. 2.

2.2.3. Iterative Tukey-MinMax pearson residual (ITMPR) method

The third proposed method divides the observations into two groups, Group 1 (G_1) and Group 2 (G_2), using the minimum and maximum values of Pearson residuals, a process referred to as the MinMax criterion. Tukey's boxplot is used to determine the cut-off point for identifying outliers. We have designated this approach as the Iterative Tukey-MinMax Pearson residual (ITMPR) method. The concept of this method is shown in Fig. 3.

3. Results

In this section, several applications based on simulation data and real data are included to illustrate the performance of the proposed methods in this paper and to compare them to existing methods.

3.1. Simulation study

Monte Carlo simulation studies, consisting of 1,000 repetitions, were performed using a beta regression model with a logit link function. Two data conditions were analyzed: uncontaminated and contaminated, and six scenarios with varying sample sizes $n = 40, 100, 300$ were considered. The first three scenarios involved one independent variable, with mean response values near zero, near one, and between 0 and 1. The remaining three scenarios involved three independent variables, with the same categories of mean response values. This division into three categories (near zero, near one, and between 0 and 1) follows a common framework found in previous research [37], ensuring a comprehensive evaluation under different conditions. All independent variables were generated from a uniform distribution, following the approach used in [22,38,39], but with a different parameter setting. The response values were generated from beta distribution with $\mu_i = \text{logit}^{-1}(x_i^T\beta)$ and $\phi = 300$.

3.1.1. Uncontaminated data

(i) The independent variable was generated from the standard uniform distribution $(x_{i1} \sim UNIF(0,1))$. The regression parameters were set as $\beta_0 = -2.5$ and $\beta_1 = -1.2$, resulting in mean responses near zero.

(ii) The independent variable was generated from the standard uniform distribution $(x_{i1} \sim UNIF(0,1))$. The regression parameters were set as $\beta_0 = 4$ and $\beta_1 = -0.8$, resulting in mean responses near one.

(iii) The independent variable was generated from the standard uniform distribution $(x_{i1} \sim UNIF(0,1))$. The regression parameters were set as $\beta_0 = 0.6$ and $\beta_1 = -0.7$, resulting in mean responses between 0 and 1.

(iv) The independent variables were generated from the standard uniform distribution $(x_{i1}, x_{i2}, x_{i3} \sim UNIF(0,1))$. The regression parameters were set as $\beta_0 = -2.3, \beta_1 = -1.1, \beta_2 = -0.7$, and $\beta_3 = -0.1$, resulting in mean response values close to zero.

(v) The independent variables were generated from the uniform distribution: $x_{i1} \sim UNIF(6,10)$, $x_{i2} \sim UNIF(2,6)$, and $x_{i3} \sim UNIF(0.1, 0.9)$. The regression parameters were set as $\beta_0 = 0.1$, $\beta_1 = 0.5$, $\beta_2 = -0.2$ and $\beta_3 = -0.5$, resulting in mean responses near one.

(vi) The independent variables were generated from the uniform distribution: $x_{i1} \sim UNIF(3,4), x_{i2} \sim UNIF(1.2, 1.4)$, and $x_{i3} \sim UNIF(0,1)$. The regression parameters were set as $\beta_0 = -1$, $\beta_1 = 0.5, \beta_2 = -0.5$ and $\beta_3 = 0.3$, resulting in mean responses between 0 and 1.

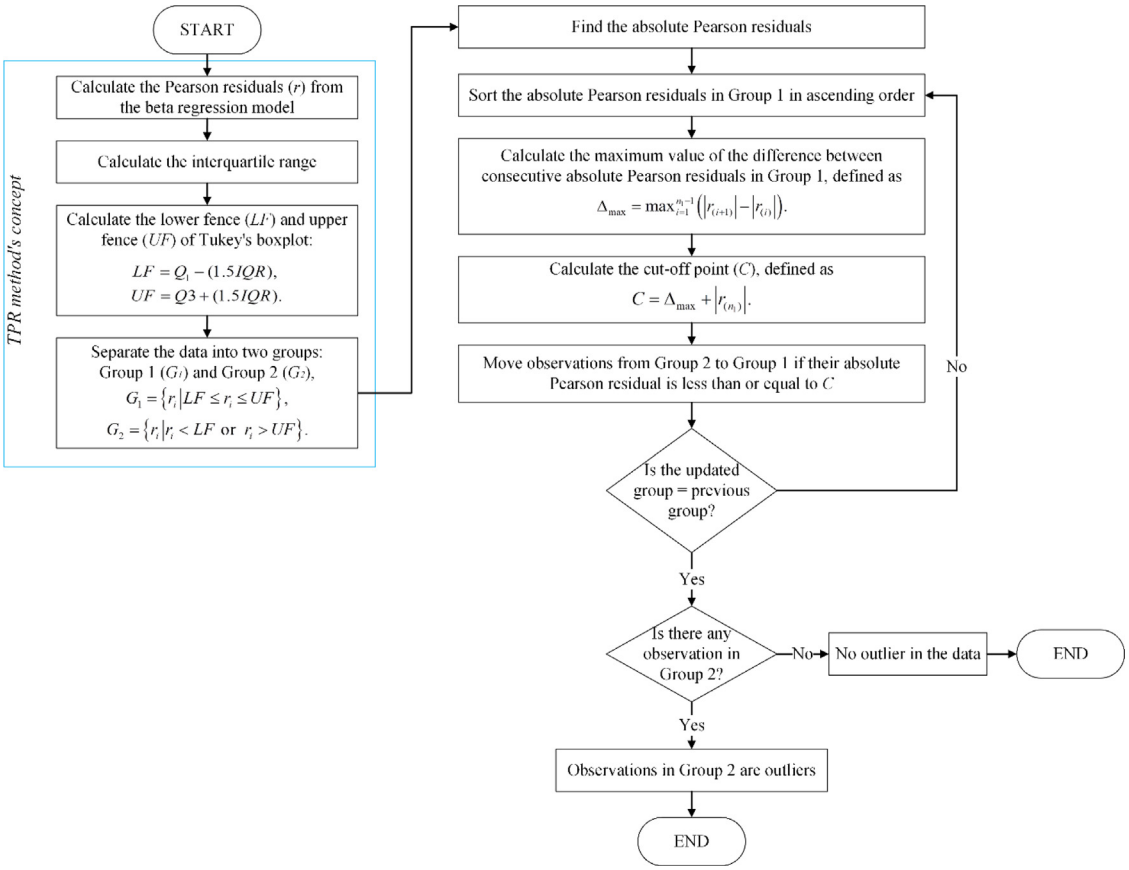

Fig. 2. Concept of ITPR method.

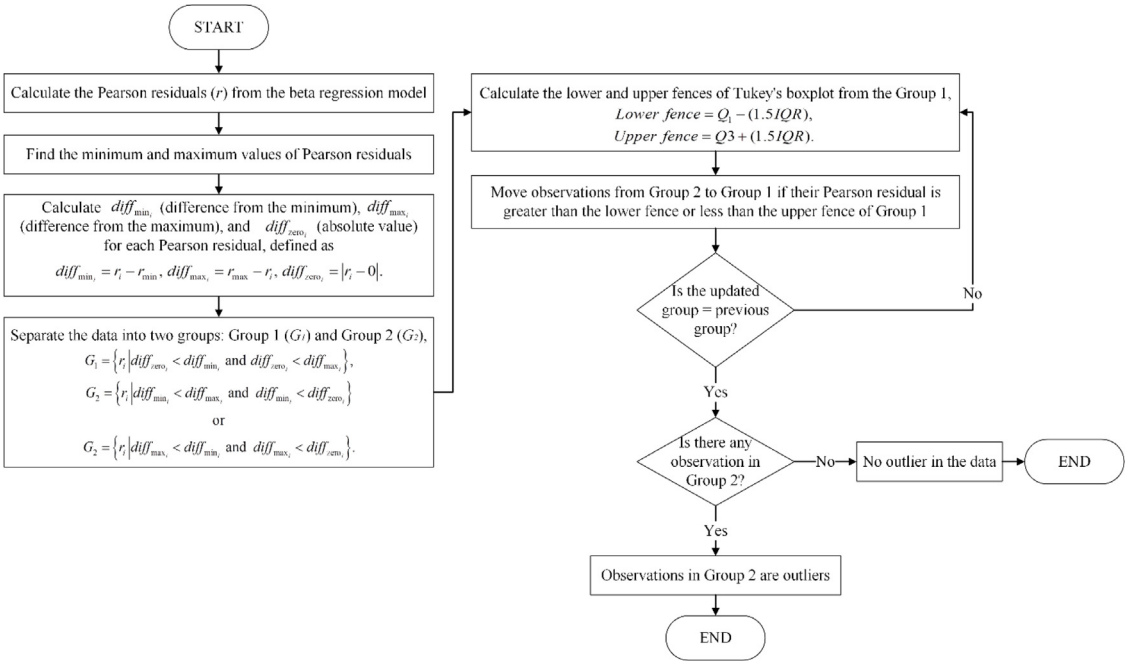

Fig. 3. Concept of ITMPR method.

Table 1
The performance evaluation in uncontaminated data.

Scenario	n	*pswamp*				
		TPR	ITPR	ITMPR	Rules of thumb	JaB-Cook's distance
i	40	0.0186	0.0145	0.0458	0.0410	0.0547
	100	0.0137	0.0107	0.0197	0.0410	0.0517
	300	0.0122	0.0093	0.0148	0.0420	0.0505
ii	40	0.0226	0.0176	0.0481	0.0411	0.0540
	100	0.0171	0.0132	0.0246	0.0409	0.0512
	300	0.0150	0.0113	0.0187	0.0406	0.0505
iii	40	0.0156	0.0104	0.0477	0.0437	0.0541
	100	0.0098	0.0070	0.0135	0.0439	0.0520
	300	0.0073	0.0051	0.0086	0.0447	0.0508
iv	40	0.0193	0.0151	0.0454	0.0415	0.0533
	100	0.0142	0.0112	0.0199	0.0412	0.0515
	300	0.0132	0.0101	0.0162	0.0415	0.0507
v	40	0.0191	0.0143	0.0465	0.0412	0.0550
	100	0.0146	0.0113	0.0215	0.0424	0.0528
	300	0.0128	0.0095	0.0158	0.0426	0.0507
vi	40	0.0145	0.0102	0.0406	0.0438	0.0514
	100	0.0093	0.0066	0.0144	0.0447	0.0522
	300	0.0076	0.0056	0.0089	0.0451	0.0506

Table 2
Outliers' positions for 5% and 10% contamination.

n	Outliers' position (i)	
	5% contamination	10% contamination
40	7, 38	7, 12, 25, 38
100	4, 7, 48, 68, 98	4, 7, 25, 34, 48, 52, 68, 77, 89, 98
300	4, 7, 48, 68, 98, 112, 146, 167, 183, 202, 206, 258, 272, 281, 288	4, 7, 25, 34, 48, 52, 68, 77, 89, 98, 107, 112, 128, 133, 146, 154, 167, 169, 183, 199, 202, 206, 236, 238, 258, 266, 272, 281, 288, 291

This simulation study applied three proposed outlier detection methods and two existing methods across different scenarios and sample sizes. For the uncontaminated data, the performance of each method was assessed using the *pswamp* value, where the number of outliers is zero ($n_{out} = 0$).

According to Table 1, the proposed methods produce lower *pswamp* values than the existing methods in all scenarios, except for the ITMPR method in a small sample size ($n = 40$). This indicates that the proposed methods are more effective in minimizing the misidentification of inliers. The ITPR method has the lowest *pswamp* value, ranging from 0.0051 to 0.0176, indicating its ability to misidentify fewer inliers than other methods, while the TPR method has a *pswamp* value ranging from 0.0073 to 0.0226. The ITMPR method shows the widest *pswamp* range of 0.0086 to 0.0481, indicating a lack of robustness due to its sensitivity to variations in the data. The existing methods show higher *pswamp* ranges, with the rules of thumb method ranging from 0.0406 to 0.0451 and the JaB-Cook's distance method ranging from 0.0505 to 0.0550.

3.1.2. Contaminated data

The data was generated under the same conditions as the uncontaminated data, where outliers were added by modifying the mean parameter μ at specific positions, following the method outlined in [40], but using a different value for the mean parameter. Three contamination cases were considered: single outlier, 5% outliers, and 10% outliers. For the single outlier case, the outlier was positioned at observation 10 ($i = 10$). Table 2 provides the specific positions of outliers for the 5% and 10% contamination cases.

The scenarios of contaminated data are as follows:

(i) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(-2.5 - 1.2x_{i1} + 2.3\right)$.
(ii) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(4 - 0.8x_{i1} - 3.5\right)$.
(iii) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(0.6 - 0.7x_{i1} + 3\right)$.
(iv) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(-2.3 - 1.1x_{i1} - 0.7x_{i2} \ -0.1x_{i3} + 2.4\right)$.
(v) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(0.1 + 0.5x_{i1} - 0.2x_{i2} \ -0.5x_{i3} - 2.75\right)$.
(vi) Modify the mean parameter μ at outlier position i using $\mu_i = \text{logit}^{-1}\left(-1 + 0.5x_{i1} - 0.5x_{i2} \ +0.3x_{i3} + 2.6\right)$.

For the contaminated data, the performance of each method was assessed using three metrics: *pout*, *pmask*, and *pswamp*. The performance evaluation values presented in Table 3 highlight the ability of various methods to detect outliers in contaminated data, specifically with a single outlier. For most methods and scenarios, the *pout* equals 1, while the *pmask* is consistently 0, indicating perfect detection of the single outlier without any masking effects. An exception is observed with the JaB-Cook's distance method in scenarios (iii) and (vi) with small sample sizes ($n = 40$), where the *pout* values are 0.9990 and 0.9860, respectively. This indicates that, in these specific cases, the JaB-Cook's distance method failed to detect the single outlier.

The *pswamp* values indicate significant variations in performance among the methods. The ITPR method achieves the lowest *pswamp* values among the proposed methods, ranging from 0.0008 to 0.01, indicating high precision in minimizing the misidentification of inliers as outliers. The TPR method also performs well, with values between

Table 3
The performance evaluation in contaminated data by single outlier.

Scenario	n	TPR			ITPR			ITMPR			Rules of thumb			JaB-Cook's distance		
		pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
i	40	1	0	0.0147	1	0	0.0039	1	0	0.0293	1	0	0.0001	1	0	0.0024
	100	1	0	0.0126	1	0	0.0072	1	0	0.0182	1	0	0.0024	1	0	0.0167
	300	1	0	0.0123	1	0	0.0087	1	0	0.0153	1	0	0.0148	1	0	0.0378
ii	40	1	0	0.0157	1	0	0.0008	1	0	0.0395	1	0	0	1	0	0.0002
	100	1	0	0.0148	1	0	0.0038	1	0	0.0209	1	0	0.0001	1	0	0.0037
	300	1	0	0.0146	1	0	0.0100	1	0	0.0183	1	0	0.0049	1	0	0.0234
iii	40	1	0	0.0114	1	0	0.0046	1	0	0.0312	1	0	0	0.9990	0.0010	0.0005
	100	1	0	0.0087	1	0	0.0049	1	0	0.0126	1	0	0	1	0	0.0035
	300	1	0	0.0076	1	0	0.0051	1	0	0.0089	1	0	0.0018	1	0	0.0181
iv	40	1	0	0.0157	1	0	0.0048	1	0	0.0327	1	0	0.0002	1	0	0.0034
	100	1	0	0.0132	1	0	0.0078	1	0	0.0184	1	0	0.0030	1	0	0.0169
	300	1	0	0.0126	1	0	0.0092	1	0	0.0156	1	0	0.0142	1	0	0.0374
v	40	1	0	0.0165	1	0	0.0027	1	0	0.0353	1	0	0	1	0	0.0027
	100	1	0	0.0132	1	0	0.0053	1	0	0.0188	1	0	0.0011	1	0	0.0119
	300	1	0	0.0121	1	0	0.0080	1	0	0.0150	1	0	0.0090	1	0	0.0320
vi	40	1	0	0.0114	1	0	0.0059	1	0	0.0298	1	0	0.0001	0.9860	0.0140	0.0023
	100	1	0	0.0093	1	0	0.0052	1	0	0.0137	1	0	0.0001	1	0	0.0055
	300	1	0	0.0072	1	0	0.0048	1	0	0.0085	1	0	0.0032	1	0	0.0218

Table 4
The performance evaluation in contaminated data by 5% outliers.

Scenario	n	TPR			ITPR			ITMPR			Rules of thumb			JaB-Cook's distance		
		pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
i	40	1	0	0.0119	1	0	0.0015	1	0	0.0372	1	0	0	1	0	0.0001
	100	1	0	0.0087	1	0	0.0004	1	0	0.0188	1	0	0	1	0	0.0005
	300	1	0	0.0069	1	0	0.0002	1	0	0.0150	1	0	0	1	0	0.0005
ii	40	1	0	0.0122	1	0	0.0003	1	0	0.1089	1	0	0	1	0	0.0001
	100	1	0	0.0095	1	0	0.0001	1	0	0.0259	1	0	0	1	0	0.0005
	300	1	0	0.0090	1	0	0.0000	1	0	0.0195	1	0	0	1	0	0.0006
iii	40	1	0	0.0077	1	0	0.0040	1	0	0.0293	1	0	0	0.9665	0.0335	0.0003
	100	1	0	0.0051	1	0	0.0025	1	0	0.0120	1	0	0	1	0	0.0005
	300	1	0	0.0044	1	0	0.0021	1	0	0.0083	1	0	0	1	0	0.0003
iv	40	1	0	0.0120	1	0	0.0017	1	0	0.0355	1	0	0	1	0	0.0005
	100	1	0	0.0088	1	0	0.0003	1	0	0.0192	1	0	0	1	0	0.0003
	300	1	0	0.0075	1	0	0.0002	1	0	0.0162	0.9999	0.0001	0	0.9999	0.0001	0.0003
v	40	1	0	0.0129	1	0	0.0019	1	0	0.0471	1	0	0	1	0	0.0006
	100	1	0	0.0092	1	0	0.0006	1	0	0.0217	1	0	0	1	0	0.0003
	300	1	0	0.0076	1	0	0.0003	1	0	0.0163	1	0	0	1	0	0.0003
vi	40	1	0	0.0088	0.9975	0.0025	0.0046	1	0	0.0284	1	0	0.0001	0.9280	0.0720	0.0018
	100	1	0	0.0055	1	0	0.0030	1	0	0.0117	1	0	0	0.9978	0.0022	0.0005
	300	1	0	0.0045	1	0	0.0023	1	0	0.0084	1	0	0	0.9999	0.0001	0.0003

0.0072 and 0.0165, while the ITMPR method exhibits greater variability, with values ranging from 0.0085 to 0.0395. In comparison, the rules of thumb method shows lower *pswamp* values, ranging from 0 to 0.0148. Its highest value is lower than those of the TPR and ITMPR methods but slightly higher than the ITPR method. This suggests that, although the rules of thumb method is less precise than ITPR, it performs comparably to the other proposed methods in minimizing the misidentification of inliers as outliers. The *pswamp* values of the proposed methods are lower than those of the JaB-Cook's distance method for larger sample sizes, indicating that the proposed methods perform better in minimizing the misidentification of inliers as outliers.

Table 4 highlights the effectiveness of various methods for identifying outliers in simulated data with 5% contamination. The results reveal that the TPR and ITMPR methods consistently achieve *pout* values of 1 across all scenarios, demonstrating their robust capability to detect all outliers. In contrast, the ITPR and rules of thumb methods do not consistently maintain pout values of 1 in certain scenarios. For example, the ITPR method recorded a *pout* value of 0.9975 in scenario (vi) with $n = 40$ and the rules of thumb methods recorded a *pout* value of 0.9999 in scenario (iv) with $n = 300$. This indicates a slight limitation in its outlier detection performance. The JaB-Cook's distance method shows the most marked decline in *pout*, reaching a low of 0.9280 in scenario (vi) with $n = 40$, highlighting its reduced reliability in detecting outliers under these conditions. Most methods exhibit negligible *pmask* values, with the TPR and ITMPR having *pmask* values of 0 in all scenarios, indicating that no outliers were misidentified as inliers. The ITPR and rules of thumb methods achieve a *pmask* close to 0 in a specific scenario, scenario (vi) with $n = 40$ and scenario (iv) with $n = 300$, respectively. This indicates a tendency to misidentify outliers as inliers. The JaB-Cook's distance method has *pmask* values near 0 in more scenarios than other methods.

The ITPR method achieves the most favorable *pswamp* values among the proposed methods, within a range between 0 and 0.0046, indicating its superior precision. The TPR method demonstrates slightly higher *pswamp* values, ranging between 0.0044 and 0.0129. The ITMPR method has the widest range of *pswamp* values (0.0083 to 0.1089), indicating increased variability in precision under specific conditions and a tendency to misidentify inliers as outliers. The rules of thumb method consistently achieves *pswamp* values of 0 in almost all scenarios, indicating perfect identification of inliers. The JaB-Cook's distance method shows a *pswamp* range of 0.0001 to 0.0018, indicating better identification of inliers than the proposed methods.

Table 5 evaluates the performance of various outlier detection methods on simulated data with 10% contamination. The proposed TPR

Table 5
The performance evaluation in contaminated data by 10% outliers.

Scenario	n	TPR			ITPR			ITMPR			Rules of thumb			JaB-Cook's distance		
		pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
i	40	1	0	0.0063	1	0	0.0006	1	0	0.0793	1	0	0	0.6395	0.3605	0
	100	1	0	0.0041	1	0	0.0002	1	0	0.0194	0.9995	0.0005	0	0.4998	0.5002	0
	300	1	0	0.0032	1	0	0.0001	1	0	0.0158	0.9995	0.0005	0	0.4878	0.5122	0
ii	40	1	0	0.0068	1	0	0.0004	1	0	0.7654	1	0	0	0.6555	0.3445	0
	100	1	0	0.0046	1	0	0.0001	1	0	0.4372	1	0	0	0.5107	0.4893	0
	300	1	0	0.0041	1	0	0	1	0	0.0512	1	0	0	0.4931	0.5069	0
iii	40	0.9973	0.0028	0.0048	0.9863	0.0138	0.0026	0.9990	0.001	0.0415	0.6228	0.3773	0	0.5480	0.4520	0.0008
	100	1	0	0.0027	0.9990	0.0010	0.0015	1	0	0.0107	0.6562	0.3438	0	0.5188	0.4812	0
	300	1	0	0.0022	1	0	0.0014	1	0	0.0080	0.6522	0.3478	0	0.5056	0.4944	0
iv	40	1	0	0.0062	0.9998	0.0003	0.0011	1	0	0.0620	0.9953	0.0048	0	0.6940	0.3060	0
	100	1	0	0.0047	1	0	0.0003	1	0	0.0222	0.9986	0.0014	0	0.5414	0.4586	0
	300	1	0	0.0034	1	0	0.0001	1	0	0.0171	0.9990	0.0010	0	0.5072	0.4928	0
v	40	1	0	0.0088	0.9995	0.0005	0.0022	1	0	0.1192	0.9930	0.0070	0	0.6870	0.3130	0.0001
	100	1	0	0.0052	1	0	0.0007	1	0	0.0277	0.9933	0.0067	0	0.5307	0.4693	0
	300	1	0	0.0042	1	0	0.0002	1	0	0.0189	0.9927	0.0073	0	0.5044	0.4956	0
vi	40	0.9878	0.0123	0.0061	0.9400	0.0600	0.0036	0.9955	0.0045	0.0390	0.6478	0.3523	0.0001	0.5695	0.4305	0.0024
	100	1	0	0.0028	0.9946	0.0054	0.0018	1	0	0.0104	0.7762	0.2238	0	0.5694	0.4306	0.0002
	300	1	0	0.0025	1	0	0.0015	1	0	0.0082	0.8411	0.1589	0	0.5311	0.4689	0

and ITMPR methods exhibit strong detection capabilities, consistently achieving *pout* values of 1 across most scenarios. Their performance decreases slightly in specific cases, such as scenarios (iii) and (vi) in small sample sizes, where the TPR method records *pout* values of 0.9973 and 0.9878, and the ITMPR method records 0.9990 and 0.9955. The ITPR method also performs well but does not achieve *pout* values of 1 in scenarios with smaller sample sizes, such as in scenarios (iii) and (vi). The rules of thumb method significantly underperform, yielding *pout* values below 1, such as scenario (iii) and scenario (vi), indicating its limited ability to detect outliers. The JaB-Cook's distance method demonstrates the weakest detection capabilities, with low *pout* values and a maximum of 0.6940 in scenario (vi) with $n = 40$.

The *pmask* values are generally low across all proposed methods, indicating that most methods do not misidentify outliers as inliers. While TPR and ITMPR achieve *pmask* values of 0 in nearly all scenarios, the ITPR method shows some masking effects in smaller samples. The rules of thumb method shows high *pmask* values in scenarios (iii) and (vi), indicating a tendency to misidentify outliers, whereas the JaB-Cook's distance method struggles with higher *pmask* values across all conditions, indicating its challenges in effectively detecting outliers. The ITPR and TPR methods demonstrate good precision with slightly higher *pswamp* values, ranging from 0 to 0.0036 and 0.0022 to 0.0088, respectively. The ITMPR method shows a wider range of pswamp values, from 0.0080 to 0.7654, indicating variability in performance depending on different contamination patterns. There is a tendency for the ITMPR method to misidentify many inliers, such as in scenario (ii) with $n = 40$, where it achieves a *pswamp* value of 0.7654. The rules of thumb and JaB-Cook's distance methods achieve consistently low *pswamp* values, reaching 0 across various scenarios, indicating strong precision.

3.2. Real data applications

The application of real-world data demonstrates the practical utility of the proposed method, underscoring its relevance and adaptability across diverse scenarios. This study utilizes breast cancer data and gasoline yield data to illustrate the method's effectiveness.

3.2.1. Breast cancer data

The breast cancer data was obtained from the University of Wisconsin Hospitals [41] and accessed through the UC Irvine Machine Learning Repository (https://doi.org/10.24432/C5DW2B). The research used the radius as the response variable, as it is associated with tumor spread; as the cell radius increases, so does the surface area [5]. The independent variables encompass significant factors influencing the radius, including diagnosis, perimeter, area, and compactness. Most cases (62.7%) in this data correspond to benign tumors, with only 37.3% (212 observations) corresponding to malignant tumors. Given this distribution, the observations with malignant tumors are treated as outliers. The evaluation focuses on the effectiveness of the method in accurately identifying observations diagnosed with malignant tumors.

Fig. 4 shows that the TPR method identified 35 observations as outliers. Twenty outliers are observations with malignant tumors (*pout* = 0.094), while the remaining ones are benign tumors (*pswamp* = 0.042). The TPR method also failed to identify 192 observations with malignant tumors as outliers (*pmask* = 0.906). Fig. 5 shows that the ITPR method identified 33 observations as outliers, where 18 outliers corresponded to malignant tumor cases (*pout* = 0.085) and the rest to benign tumors (*pswamp* = 0.042). The method failed to identify 194 malignant tumor cases as outliers (*pmask* = 0.915). Fig. 6 demonstrates that the ITMPR method identified 207 observations as outliers, with 148 of them corresponding to malignant tumor cases (*pout* = 0.698) and the remaining associated with benign tumors (*pswamp* = 0.165). The method failed to identify 64 observations with malignant tumor as outliers (*pmask* = 0.302). Fig. 7 illustrates that the rules of thumb method identified 8 observations as outliers, including 3 associated with malignant tumor cases (*pout* = 0.014) and 5 associated with benign tumors (*pswamp* = 0.014). The method also identified 209 observations with malignant tumors as inliers (*pmask* = 0.986). Fig. 8 shows that the JaB-Cook's method identified 19 observations as outliers, 14 of which were linked to malignant tumor cases (*pout* = 0.061) and 5 associated with benign tumors (*pswamp* = 0.017). The method also identified 209 malignant tumor observations as inliers (*pmask* = 0.939).

3.2.2. Gasoline yield data

The gasoline yield data was collected by [42] and documented by [19] in the 'betareg' package titled "GasolineYield", which consists of 32 observations and six variables. This data was also analyzed by [40], providing further insights into diagnostic plots in beta regression models. This study analyzed yield as the response variable, using temp10 and temp as independent variables. Potential outliers were initially identified using established rules of thumb; observations 9 and 10 were removed, and the model was refitted. To introduce true outliers into the data, the response values of two observations with the highest Pearson residuals (observations 2 and 12) were modified by adjusting them to $y_i = y_i + \max(y)$.

Fig. 9 shows that the TPR method identified observations 2, 12, and 29 as outliers, indicating this method successfully identified outliers but

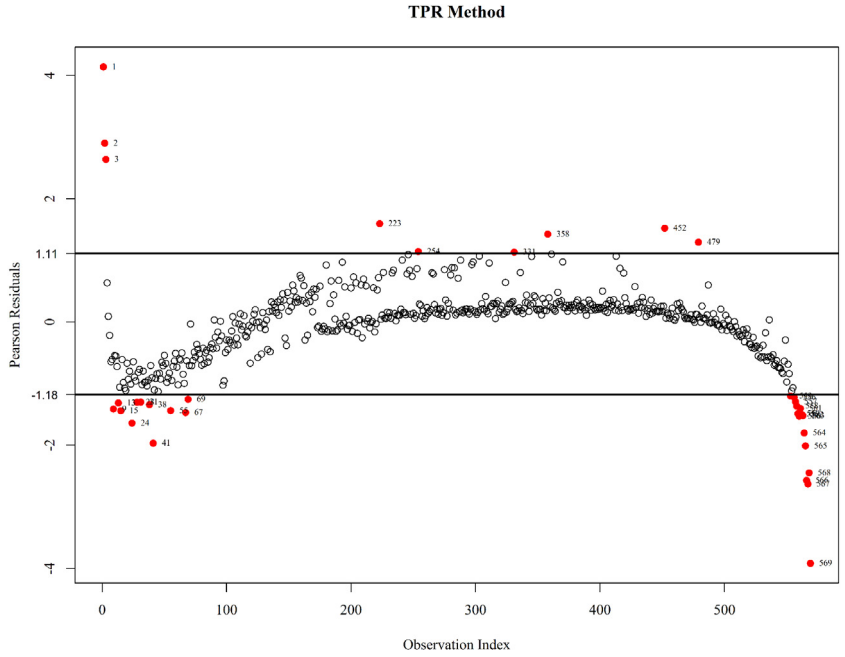

Fig. 4. Breast cancer plot where the red dots are the observations identified as outliers using the TPR method with cut-off points of −1.18 and 1.11.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

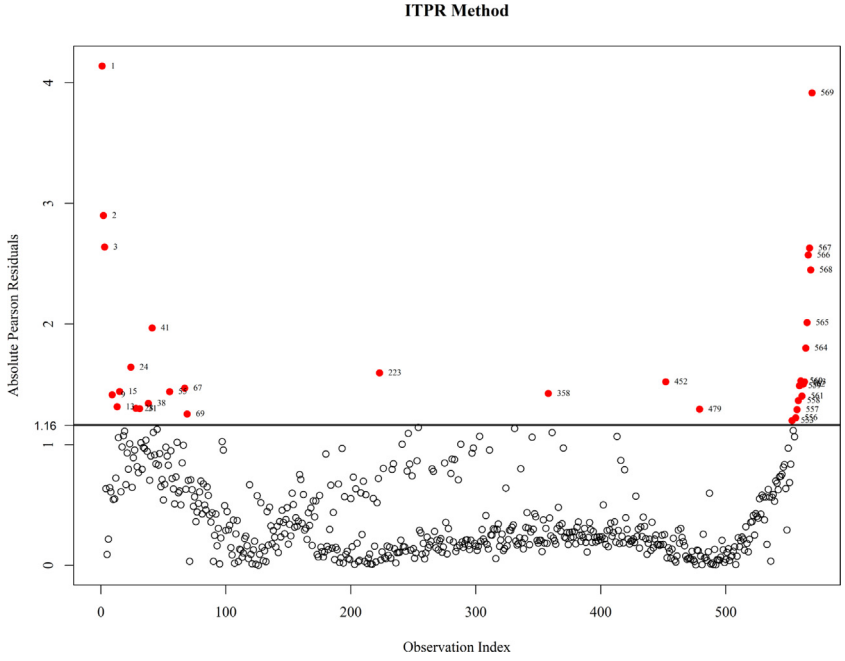

Fig. 5. Breast cancer plot where the red dots are the observations identified as outliers using the ITPR method with a cut-off point of 1.16.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

also identified an inlier as an outlier. Fig. 10 demonstrates that the ITPR method identified observations 2 and 12 as outliers, indicating perfect outlier detection without masking and swamping effect. From Fig. 11, the finding shows that the ITMPR also identified observations 2, 12, and 29 as outliers, the same as the TPR method. Fig. 12 illustrates that the rules of thumb method perfectly identified the outliers, as evidenced by only identifying observations 2 and 12 as outliers. While other methods could identify the outliers, Fig. 13 shows that JaB-Cook's method did not identify any outliers.

4. Discussion

The simulation study demonstrates that the ITPR method consistently detected fewer outliers in uncontaminated data, showcasing superior robustness across varying sample sizes compared to other methods. When contamination was introduced, the ITPR, TPR, and ITMPR methods outperformed the rules of thumb and JaB-Cook's distance methods, achieving *pout* values approaching 1 and *pmask* values nearing 0, despite exhibiting a slightly higher swamping effect. In real

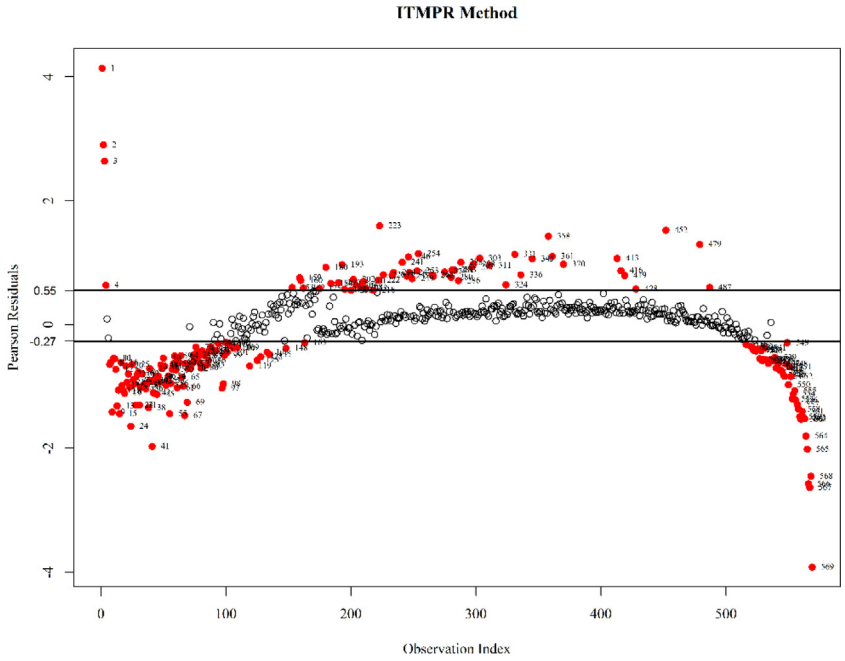

Fig. 6. Breast cancer plot where the red dots are the observations identified as outliers using the ITMPR method with cut-off points of −0.27 and 0.55.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

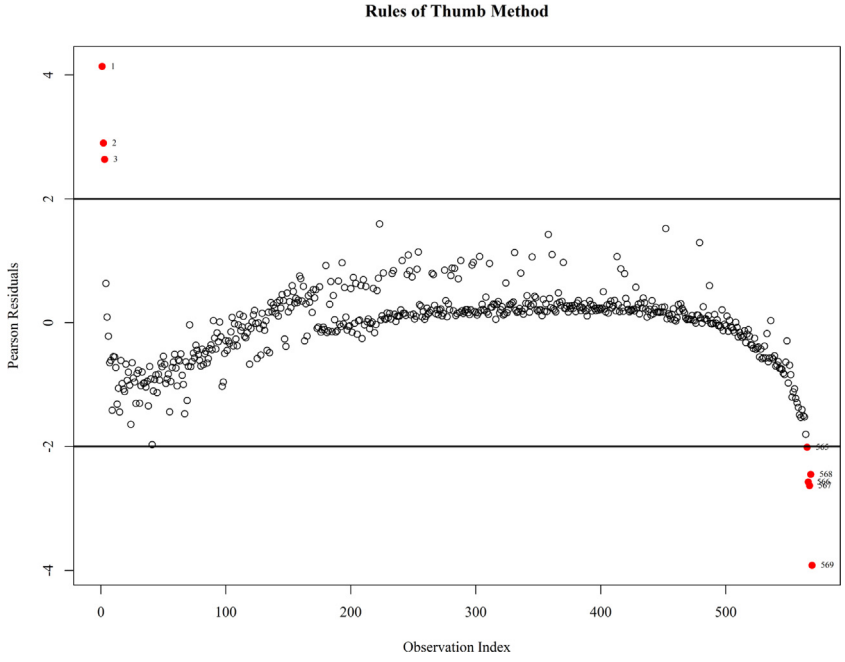

Fig. 7. Breast cancer plot where the red dots are the observations identified as outliers using the rules of thumb method.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

data applications, the ITPR, TPR, and ITMPR methods demonstrated superior performance in identifying outliers, affirming their practical applicability and effectiveness.

The effectiveness of these methods relies on certain assumptions regarding the distribution of Pearson residuals in beta regression. Pearson residuals are typically centered around zero in well-specified models [43]. According to [3], a plot of the residuals against the index of the observations should show no detectable pattern, indicating a well-fitting model. The methods assume this behavior, with Pearson residuals exhibiting symmetric properties under such conditions. These considerations are crucial when applying the methods to detect outliers, as misidentification can occur if the residual distribution deviates significantly from these assumptions.

Among the three proposed methods, the ITPR method stands out as the most precise and consistent, demonstrating an exceptional ability to correctly identify true outliers while minimizing the misclassification of inliers as outliers. The TPR method also shows strong performance, balancing precision and robustness, making it a reliable alternative.

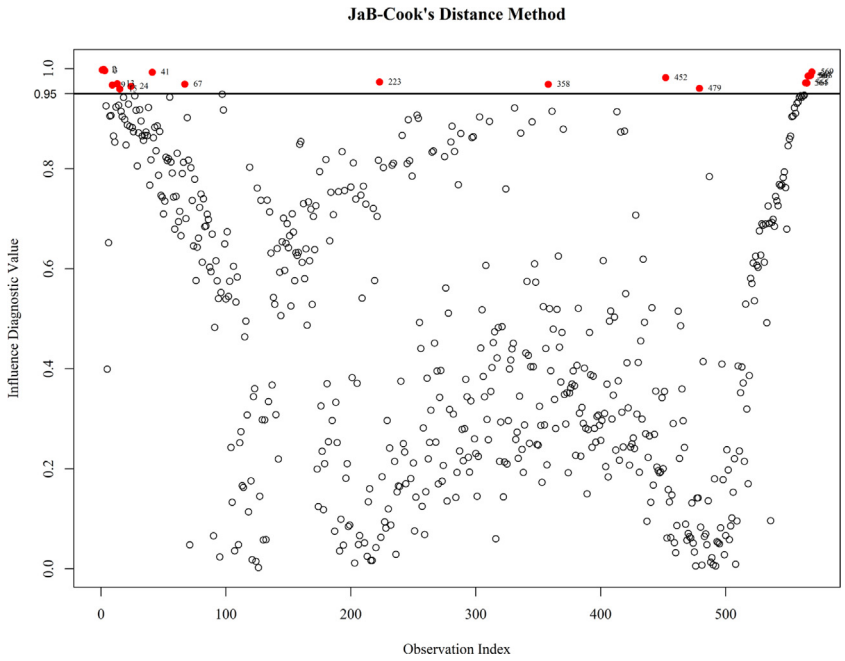

Fig. 8. Breast cancer plot where the red dots are the observations identified as outliers using the JaB-Cook's distance method with $p = 0.95$.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

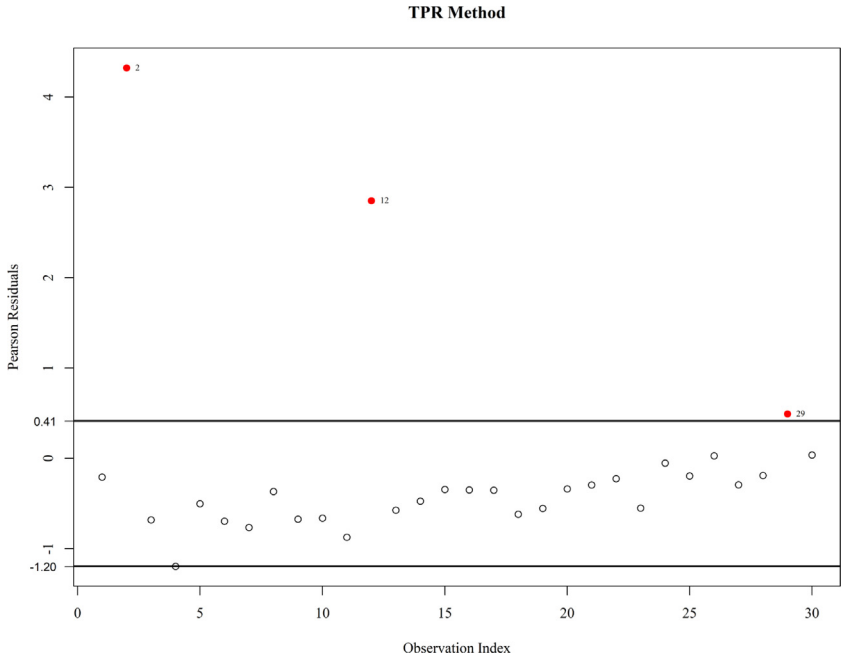

Fig. 9. Gasoline yield plot where the red dots are the observations identified as outliers using the TPR method with cut-off points of −1.20 and 0.41.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

While the ITMPR method is effective in detecting outliers, we observed some variability, especially due to the influence of the extreme values in Group 2. The MinMax criterion used to define Group 2 can bias the method towards extreme values, leading to the over-identification of outliers. This reduces the overall reliability of the ITMPR method. To address this issue, we could modify the criteria for transferring data between groups, ensuring that extreme values do not unduly affect the outlier detection process. Exploring other robust methods may also improve the ITMPR method's ability to distinguish between true outliers and extreme but valid values, reducing the swamping effect.

In contrast, the rules of thumb method is robust under normal conditions but declines in performance when the data contains more than 5% contamination. This finding aligns with [20,44], which observed that approximately 95% of observations fall within two standard deviations of the regression line. Similarly, the JaB-Cook's distance proves effective in identifying outliers; however, its performance diminishes under high contamination levels (more than 5% contamination).

Although the proposed methods exhibit some swamping effect, they demonstrate considerable promise in high-risk applications such as fraud detection, where missing an outlier could have significant consequences. For example, [45] highlighted the importance of reliable

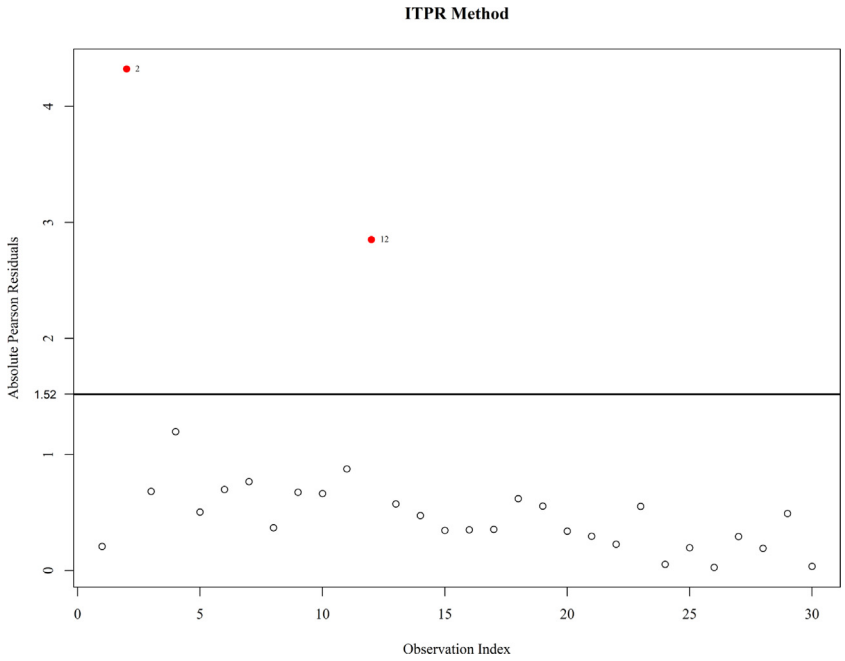

Fig. 10. Gasoline yield plot where the red dots are the observations identified as outliers using the ITPR method with a cut-off point of 1.52.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

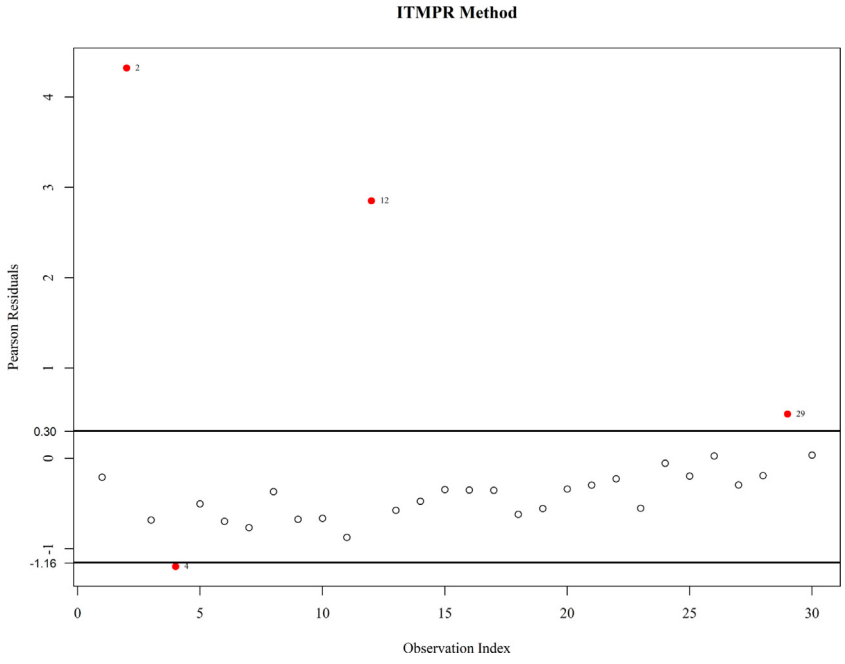

Fig. 11. Gasoline yield plot where the red dots are the observations identified as outliers using the ITMPR method with cut-off points of −1.16 and 0.30.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outlier detection in fraud detection in financial transactions. Their study underscores the critical need for robust anomaly detection techniques, which aligns with our research on improving outlier detection in beta regression.

5. Conclusions

This paper proposes three novel outlier detection methods, TPR, ITPR, and ITMPR, in the beta regression model by utilizing Tukey's boxplot and Pearson residual. The performance of the proposed methods is considered in the simulation study using Monte Carlo simulation and real data applications. The performance evaluations used in this paper are the probability of all outliers being successfully detected, masking rate, and swamping rate. The findings of this study underscore the effectiveness and reliability of the proposed ITPR, TPR, and ITMPR methods in detecting outliers under varying conditions of contamination and data complexity compared to the existing methods. The ITPR method emerges as the most precise and consistent method across both simulation studies and real data applications, with robust performance in identifying true outliers while minimizing the misidentification of inliers. The TPR and ITMPR methods also show strong potential, with the former offering a balanced approach and the latter requiring further

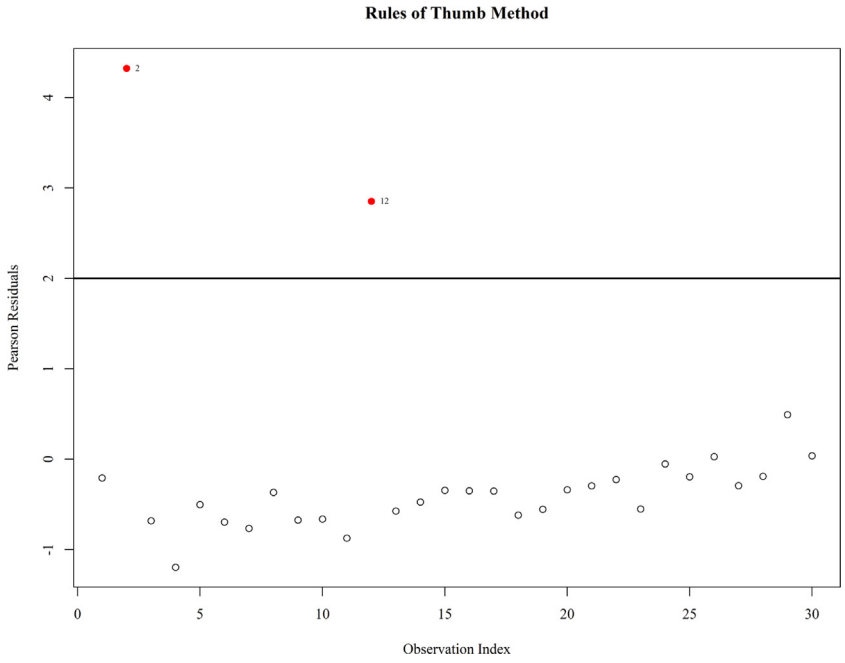

Fig. 12. Gasoline yield plot where the red dots are the observations identified as outliers using the rules of thumb method.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

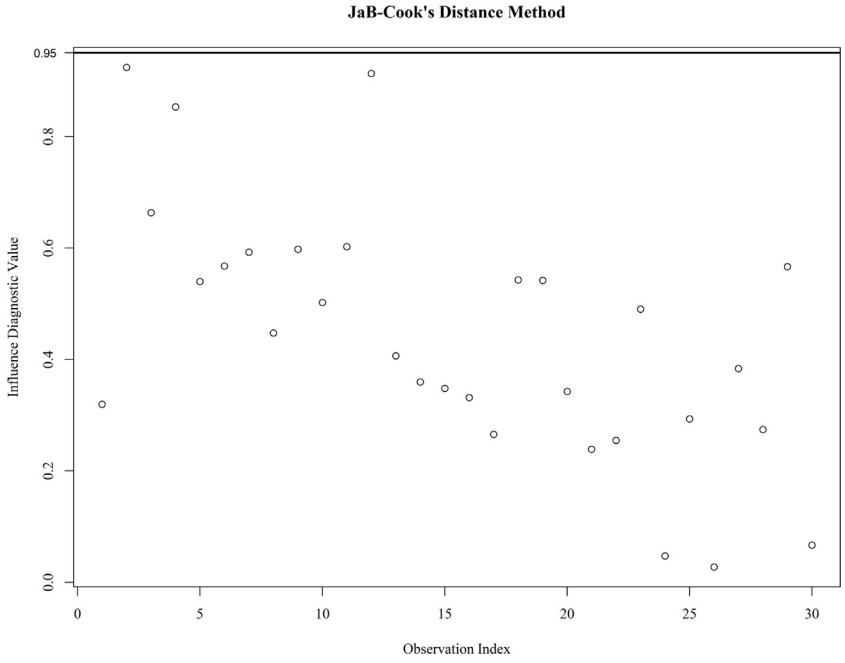

Fig. 13. Gasoline yield plot where the red dots are the observations identified as outliers using the JaB-Cook's distance method with $p = 0.95$.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

investigation due to its susceptibility to higher swamping effects. While the proposed methods exhibit some degree of swamping effect, this characteristic makes them particularly useful in situations where the cost or risk of missing an outlier is significantly high. In fields like fraud detection, manufacturing safety, or health monitoring, it is generally more acceptable to incorrectly identify some regular data as a concern than to miss a genuine issue.

However, the performance of these methods is based on certain assumptions, such as the symmetry of Pearson residuals in well-specified models and the assumption that the response variable follows a beta distribution. If these assumptions are not met, the methods' effectiveness may be compromised. Future work should focus on refining the ITMPR method, addressing the limitations of JaB-Cook's distance, and exploring the broader applicability and effectiveness of these methods across different data structures and applications.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Oktsa Dwika Rahmashari reports financial support was provided by Khon Kaen University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a Khon Kaen University Scholarship, Khon Kaen, Thailand for ASEAN and GMS Countries' personnel. The authors would like to express their gratitude to the Department of Statistics, Faculty of Science, Khon Kaen University for the financial support that made completing this work possible.

Appendix A. Supplementary material

The research data for this article, including the script and code related to this study, are available on Mendeley Data. https://data.mendeley.com/preview/mytyzttg4w?a=2bef80ad-6314-4c85-834c-f41caaa03ed6

Data availability

Data will be made available on request.

References

[1] R. Kieschnick, B.D. McCullough, Regression analysis of variates observed on (0, 1): percentages, proportions and fractions, Stat. Model. 3 (2003) 193–213, http://dx.doi.org/10.1191/1471082X03st053oa.
[2] P. Paolino, Maximum likelihood estimation of models with beta-distributed dependent variables, Political Anal. 9 (2001) 325–346, http://dx.doi.org/10.1093/oxfordjournals.pan.a004873.
[3] S.L.P. Ferrari, F. Cribari-Neto, Beta regression for modelling rates and proportions, J. Appl. Stat. 31 (2004) 799–815, http://dx.doi.org/10.1080/0266476042000214501.
[4] P. McCullagh, J.A. Nelder, Generalized Linear Models, second ed., Routledge, New York, 1989.
[5] M.M. Abo El Nasr, A.A. Abdelmegaly, D.A. Abdo, Performance evaluation of different regression models: application in a breast cancer patient data, Sci. Rep. 14 (2024) 12986, http://dx.doi.org/10.1038/s41598-024-62627-6.
[6] F. Cribari-Neto, A beta regression analysis of COVID-19 mortality in Brazil, Infect. Dis. Model. 8 (2023) 309–317, http://dx.doi.org/10.1016/j.idm.2023.02.005.
[7] C.J. Swearingen, B.C. Tilley, R.J. Adams, Z. Rumboldt, J.S. Nicholas, D. Bandyopadhyay, R.F. Woolson, Application of beta regression to analyze ischemic stroke volume in NINDS rt-PA clinical trials, Neuroepidemiology 37 (2011) 73–82, http://dx.doi.org/10.1159/000330375.
[8] S.K.R. Yellareddygari, J.S. Pasche, R.J. Taylor, S. Hua, N.C. Gudmestad, Beta regression model for predicting the development of pink rot in potato tubers during storage, Plant Dis. 100 (2016) 1118–1124, http://dx.doi.org/10.1094/pdis-06-15-0696-re.
[9] S. Kokilavani, V. Geethalakshmi, R. Pangayarselvi, J. Bhuvaneswari, G. Sudhakar, S. Subbulakshmi, P. Priyanka, Timmanna, S.K. Bal, Beta regression model for predicting development of powdery mildew in black gram, J. Agrometeorol. 25 (2023) 577–582, http://dx.doi.org/10.54386/jam.v25i4.2343.
[10] E.A. Geissinger, C.L.L. Khoo, I.C. Richmond, S.J.M. Faulkner, D.C. Schneider, A case for beta regression in the natural sciences, Ecosphere 13 (2022) 1–16, http://dx.doi.org/10.1002/ecs2.3940.
[11] D. Blane, G. Netuveli, S.M. Montgomery, Quality of life, health and physiological status and change at older ages, Soc. Sci. Med. 66 (2008) 1579–1587, http://dx.doi.org/10.1016/j.socscimed.2007.12.021.
[12] M. Babajanpour, Z. Iraji, H. Sadeghi-Bazargani, M. Asghari-Jafarabadi, Utilizing beta regression in predicting the underlying factors of motorcycle rider behavior, J. Biostat. Epidemiol 7 (2021) 7–24, http://dx.doi.org/10.18502/jbe.v7i1.6291.
[13] S. Zhang, M. Abdel-Aty, Risky driver identification using beta regression based on naturalistic driving data, Transp. Res. Rec. 2678 (2024) 325–334, http://dx.doi.org/10.1177/03611981231179475.
[14] V.J. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2004) 85–126, http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9.
[15] F. Kamalov, H.H. Leung, Outlier detection in high dimensional data, J. Inf. Knowl. Manag. 19 (2020) 1–16, http://dx.doi.org/10.1142/S0219649220400134.
[16] X. Jiao, F. Pretis, Testing the presence of outliers in regression models*, Oxf. Bull. Econ. Stat. 84 (2022) 1452–1484, http://dx.doi.org/10.1111/obes.12511.
[17] V. Barnett, T. Lewis, Outliers in Statistical Data, third ed., Wiley, New York, 1994.
[18] S. Chatterjee, A.S. Hadi, Influential observations, high leverage points, and outliers in linear regression, Statist. Sci. 1 (1986) 379–393, http://dx.doi.org/10.1214/ss/1177013622.
[19] F. Cribari-Neto, A. Zeileis, Beta regression in R, J. Stat. Softw. 34 (2010) 1–24, http://dx.doi.org/10.18637/jss.v034.i02.
[20] E. Serdahl, An Introduction To Graphical Analysis of Residual Scores and Outlier Detection in Bivariate Least Squares Regression Analysis, New Orleans, 1996.
[21] P.L. Espinheira, S.L.P. Ferrari, F. Cribari-Neto, On beta regression residuals, J. Appl. Stat. 35 (2008) 407–419, http://dx.doi.org/10.1080/02664760701834931.
[22] J.M. Muñoz-Pichardo, J.L. Moreno-Rebollo, R. Pino-Mejías, M.D.C. de la Vega, Influence measures in beta regression models through distance between distributions, AStA Adv. Stat. Anal. 103 (2019) 163–185, http://dx.doi.org/10.1007/s10182-018-00332-2.
[23] M.A. Martin, S. Roberts, Jackknife-after-bootstrap regression influence diagnostics, J. Nonparametr. Stat. 22 (2010) 257–269, http://dx.doi.org/10.1080/10485250903287906.
[24] U. Beyaztas, A. Alin, Jackknife-after-bootstrap as logistic regression diagnostic tool, Commun. Stat. Simul. Comput. 43 (2014) 2047–2060, http://dx.doi.org/10.1080/03610918.2013.783068.
[25] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.
[26] M. Amin, S. Afzal, M.N. Akram, A.H. Muse, A.H. Tolba, T.A. Abushal, Outlier detection in gamma regression using Pearson residuals: simulation and an application, AIMS Math. 7 (2022) 15331–15347, http://dx.doi.org/10.3934/math.2022840.
[27] A.H.M.R. Imon, A.S. Hadi, Identification of multiple outliers in logistic regression, Comm. Statist. Theory Methods 37 (2008) 1697–1709, http://dx.doi.org/10.1080/03610920701826161.
[28] H. Midi, S.B. Ariffin, Modified standardized pearson residual for the identification of outliers in logistic regression model, J. Appl. Sci. 13 (2013) 828–836, http://dx.doi.org/10.3923/jas.2013.828.836.
[29] A.B. Simas, W. Barreto-Souza, A.V. Rocha, Improved estimators for a general class of beta regression models, Comput. Statist. Data Anal. 54 (2010) 348–366, http://dx.doi.org/10.1016/j.csda.2009.08.017.
[30] C.S.K. Dash, A.K. Behera, S. Dehuri, A. Ghosh, An outliers detection and elimination framework in classification task of data mining, Decis. Anal. J. 6 (2023) 1–8, http://dx.doi.org/10.1016/j.dajour.2023.100164.
[31] T. Thanwiset, W. Srisodaphol, Statistical method for finding outliers in multivariate data using a boxplot and multiple linear regression, Sains Malays. 52 (2023) 2725–2732, http://dx.doi.org/10.17576/jsm-2023-5209-20.
[32] O.T. Ojo, A. Fernández Anta, R.E. Lillo, C. Sguera, Detecting and classifying outliers in big functional data, Adv. Data Anal. Classif. 16 (2022) 725–760, http://dx.doi.org/10.1007/s11634-021-00460-9.
[33] S. Rose, S. Nickolas, S.M. Sunoj, S. S. Sangeetha, A self-learning algorithm for identifying the leverage points in soil data using quantile regression forests, Decis. Anal. J. 10 (2024) 1–13, http://dx.doi.org/10.1016/j.dajour.2023.100375.
[34] D.M. Sebert, D.C. Montgomery, D.A. Rollier, A clustering algorithm for identifying multiple outliers in linear regression, Comput. Statist. Data Anal. 27 (1998) 461–484, http://dx.doi.org/10.1016/S0167-9473(98)00021-8.
[35] N.S. Zulkipli, S.Z. Satari, W.N.S.W. Yusoff, The effect of different similarity distance measures in detecting outliers using single-linkage clustering algorithm for univariate circular biological data, Pak. J. Stat. Oper. Res. 18 (2022) 561–573, http://dx.doi.org/10.18187/pjsor.v18i3.3982.
[36] S.S.S.A. Mutalib, S.Z. Satari, W.N.S.W. Yusoff, A new single linkage robust clustering outlier detection procedures for multivariate data, Sains Malays. 52 (2023) 2431–2451, http://dx.doi.org/10.17576/jsm-2023-5208-19.
[37] G.H.A. Pereira, On quantile residuals in beta regression, Commun. Stat. Simul. Comput. 48 (2019) 302–316, http://dx.doi.org/10.1080/03610918.2017.1381740.
[38] L.C. Chien, Multiple deletion diagnostics in beta regression models, Comput. Statist. 28 (2013) 1639–1661, http://dx.doi.org/10.1007/s00180-012-0370-9.
[39] A.V. Rocha, A.B. Simas, Influence diagnostics in a general class of beta regression models, Test 20 (2011) 95–119, http://dx.doi.org/10.1007/s11749-010-0189-z.
[40] L.C. Chien, Diagnostic plots in beta-regression models, J. Appl. Stat. 38 (2011) 1607–1622, http://dx.doi.org/10.1080/02664763.2010.515677.
[41] W. Wolberg, O. Mangasarian, N. Street, W. Street, Breast Cancer Wisconsin (Diagnostic) [Dataset], UCI Machine Learning Repository, 1993, http://dx.doi.org/10.24432/C5DW2B.
[42] N.H. Prater, Estimate gasoline yields from crudes, Pet. Refin. 35 (1956) 236–238.
[43] T. Anholeto, M.C. Sandoval, D.A. Botter, Adjusted Pearson residuals in beta regression models, J. Stat. Comput. Simul. 84 (2014) 999–1014, http://dx.doi.org/10.1080/00949655.2012.736993.
[44] C. Lewis-Beck, M. Lewis-Beck, Applied Regression: An Introduction, second ed., SAGE Publications, Inc, 2017.
[45] J.K. Afriyie, K. Tawiah, W.A. Pels, S. S. Addai-Henne, H.A. Dwamena, E.O. Owiredu, S.A. Ayeh, J. Eshun, A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions, Decis. Anal. J. 6 (2023) 1–12, http://dx.doi.org/10.1016/j.dajour.2023.100163.