AIMS Mathematics, 7(8): 15331–15347.
DOI: 10.3934/math.2022840
Received: 05 January 2022
Revised: 28 April 2022
Accepted: 11 May 2022
Published: 17 June 2022

http://www.aimspress.com/journal/Math

Research article

Outlier detection in gamma regression using Pearson residuals: Simulation and an application

Muhammad Amin[1], Saima Afzal[2], Muhammad Nauman Akram[1], Abdisalam Hassan Muse[3], Ahlam H. Tolba[4] and Tahani A. Abushal[5,*]

[1] Department of Statistics, University of Sargodha, Sargodha, Pakistan
[2] Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan
[3] Department of Mathematics (Statistics Option), Pan African University, Institute for Basic Sciences, Technology and Innovation (PAUSTI); Nairobi, 62000-00200, Kenya
[4] Mathematics Department, Faculty of Science, Mansoura University, Mansoura 35516, Egypt
[5] Department of Mathematical Science, Faculty of Applied Science, Umm AL-Qura University, Makkah, 21961, Saudi Arabia

* **Correspondence:** Email: taabushal@uqu.edu.sa.

Abstract: In data analysis, the choice of an appropriate regression model and outlier detection are both very important in obtaining reliable results. Gamma regression (GR) is employed when the distribution of the dependent variable is gamma. In this work, we derived new methods for outlier detection in GR. The proposed methods are based upon the adjusted and standardized Pearson residuals. Furthermore, a comparison of available and proposed methods is made using a simulation study and a real-life data set. The results of simulation and real-life application the evidence better performance of the adjusted Pearson residual based outlier detection approach.

Keywords: adjusted Pearson residuals; gamma regression; outlier detection; Pearson residuals
Mathematics Subject Classification: 62J12, 62J20

1. Introduction

Regression analysis is the main tool to study the relationship between and dependence of one or more variables on one or more independent variables. These relationships can be observed in every research area. Regression analysis has a wide variety of applications in different fields [1–4] and Sarhan et al. [5]. Regression results are reliable only if the quality of data is good and the selected regression model is correct. If the quality of data and the regression model are not appropriate, then the results of

measuring such relationships are incorrect. The quality of data refers to the outlier free data set [2, 6]. The correct regression model refers to identifying the distribution of the response variable.

An outlier is a point that is far from the rest of the data points [7–9]. In the regression context, Desgagné [10] stated that the observations with more distant regression errors conflict with most of the errors originating from the assumed normal distribution. An outlier may or may not affect the regression inferences. An outlier may be in one or multiple variables. Outlier detection in univariate analysis has been done by several researchers i.e. many studies in the literature have focused on univariate analysis [7–9, 11, 12]. Outlier diagnostics in the linear regression model (LRM) has also gained much attention from researchers [13,14]. Balasooriya et al. [15] compared some well-known outlier detection methods by using the LRM. They concluded that all methods do not agree with each other for the detection of outliers.

Regression analysis is used to determine the model for forecasting/prediction purposes. There is a variety of regression models, e.g., LRMs, generalized linear models (GLMs) and non-linear models, Gamma regression (GR) is employed, when the distribution of the dependent variable is gamma. GR has a variety of applications in the literature with examples in health sciences, industries and environment, for more details see [16–23].

Outlier detection using univariate gamma response without considering any independent variable is also available in the literature [24–28]. Shayib and Young [29] first studied the extreme residuals in GR and proposed the Pearson and Anscombe residuals with modified forms; they concluded that the modified forms of these residuals are not good.

The detection of outliers in the GR model has not been addressed in the literature. This paper deals with outlier detection in the GR model by using a new approach i.e., an adjusted Pearson residuals (PRs) approach. Outlier detection with the adjusted form of residuals (other than PRs) was first studied by Tiao and Guttman [30]. This approach was also proposed for some of the GLM responses. Cordeiro [31] introduced the adjusted PRs for the Poisson regression model. In addition, adjusted Wald residuals and PRs for beta regression have been suggested by various authors [32,33]. In these studies, most of the researchers focused on the adjusted PRs probability distributions. They observed that the adjusted residuals performed better than others.

The main objectives of the current research were to propose some outlier diagnostics based upon Pearson (standardized and adjusted) residuals in GR, modify some available LRM-based outlier detection methods for GR and then make a comparison of these modified and proposed outlier detection methods with the help of simulations and a real data set.

2. Materials and methods

The probability density function of the gamma response variable (y) is given by

$$f(y;\nu,\tau) = \frac{1}{\tau^{\nu}\Gamma(\nu)} y^{\nu-1} e^{-\frac{y}{\tau}},\ y>0,\ \nu>0,\ \tau>0. \tag{2.1}$$

The mean and variance of Eq (2.1) are respectively given as $E(y) = \nu\tau$ and $Var(y) = \nu\tau^2$. According to Hardin and Hilbe [34], Eq (2.1) can be transformed with parameters $\nu = \phi^{-1}$ and $\tau = \mu\phi$. Given these parameters, the gamma density for y is given by

$$f(y;\mu,\phi) \quad = \quad \frac{1}{\Gamma\left(\frac{1}{\phi}\right)}\left(\frac{1}{\mu\phi}\right)^{\frac{1}{\phi}} y^{\frac{1}{\phi}-1}\, e^{-\frac{y}{\mu\phi}}\,,\ y>0,\mu>0,\phi>0. \tag{2.2}$$

The mean and variance of Eq (2.2) are respectively given as $E(y)=\mu$ and $Var(y)=\phi\mu^2$.

For the ith observation, let $x_{i1},\ldots x_{\text{ip}}$ represent the p independent variables. Then, the GR for the mean of the response variable y is given by

$$g(\mu_i)=\eta_i=\ \mathbf{x_i^T}\beta, \qquad i=1,\ldots,n.$$

where $\mathbf{x_i^T}=\left(1,x_{i1},\ldots,x_{\text{ip}}\right)$, $\beta=\left(\beta_0,\beta_1,\ldots,\beta_p\right)$ is the vector of regression coefficients including intercept and g(.) is the link function. This link function in the GR can be reciprocal or log.

Let l_i be the log-likelihood function of the response variable of Eq (2.2) which is mathematically defined by

$$l_i \quad = \quad l_i(\mu_i,\phi)=\sum_{i=1}^{n}\left\{\frac{\frac{y_i}{\mu_i}+\ln(\mu_i)}{-\phi}+\frac{1-\phi}{\phi}\ln(y_i)-\frac{\ln(\phi)}{\phi}-ln\left[\Gamma\left(\frac{1}{\phi}\right)\right]\right\}. \tag{2.3}$$

Let $\widehat{\beta}$, $\widehat{\mu}$ and $\widehat{\phi}$ be the maximum likelihood estimates (MLEs) which are obtained by maximizing the log-likelihood of Eq (2.3) using the Netwon-Raphson iterative method. The MLE of β is computed by solving the system of equations. For this purpose, we equate the first derivative of Eq (2.3) to zero; then, we have

$$U(\beta) \quad = \quad \frac{\partial l}{\partial\beta}=-\frac{1}{\phi}\left(\mathbf{y}-(\mathrm{X}\beta)^{-\mathbf{1}}\right)\mathbf{X}=0, \tag{2.4}$$

where $U(\beta)$ is the score vector of the order $(p+1)\times 1$. Since Eq (2.4) is nonlinear in β, Newton-Raphson methods can be employed for the estimation of β [34]. Suppose β^m is the approximated MLE of β at the mth iteration; then, the iterative reweighted method [35] gives the following expression

$$\beta^{m+1} \quad = \quad \beta^m+\{I(\beta^m)\}^{-1}U(\beta^m), \tag{2.5}$$

where $I(\beta^m)$ is the $(p+1)\times(p+1)$ fisher information matrix at the mth iteration. Applying convergence in deviance to Eq (2.5), the unknown parameters can be computed as

$$\widehat{\beta}=\left(\mathbf{X^T}\widehat{\mathbf{W}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X^T}\widehat{\mathbf{W}}\widehat{\mathbf{z}}, \tag{2.6}$$

where $\widehat{z_i}=\widehat{\eta_i}+\frac{y_i-\widehat{\mu_i}}{\widehat{\mu_i^2}}$ is the adjusted response variable, $\widehat{\mathbf{W}}=diag\left(\widehat{\mu_1^2},\ldots,\widehat{\mu_n^2}\right)$ and $\widehat{\mu_i}=\frac{1}{x_i^T\widehat{\beta}}$.

Several types of the GLM residuals are available in the literature [33] but we consider the most popular PR for the detection of an outlier.

The PRs for GR are defined by

$$\chi_i \quad = \quad \frac{y_i-\mu_i}{\sqrt{V(\mu_i)}}=\frac{y_i-\mu_i}{\mu_i}\,. \tag{2.7}$$

The standardized PRs are characterized as

$$\chi_i^{'} = \frac{\chi_i}{\sqrt{\phi(1-h_{\rm ii})}}, \tag{2.8}$$

where $h_{\rm ii} = diag\left(\mathbf{H} = \widehat{\mathbf{w}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X^T}\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X^T}\widehat{\mathbf{W}}^{\frac{1}{2}}\right)$.

Generally, $E(\chi_i)$ to order $O\left(n^{-1}\right)$ does not converge to zero and Var (χ_i) does not tend to one. Here n is the sample size. To handle such a situation, we require some adjustments to these residuals. To do this, Cox and Snell [36] obtained some matrix formulae for the adjusted residuals.

Various criteria are available in the literature [2,12] for testing the quality of regression models. These criteria include mean quadratic error prediction (MEP), the Akaike information criterion (AIC), standard errors and coefficients of determination. As our study is concerned with the GR model, we consider some different criteria for testing the goodness of the GR model after diagnosing the outlying points. These include the Pearson chi-square statistic $\left(\chi^2\right)$, MEP, the AIC, Efron's pseudo r-squared criterion $\left(R^2_{Efron}\right)$ and the dispersion parameter $\left(\hat{\phi}\right)$; these criteria are computed according to the following relations:

$\chi^2 = \sum_{i=1}^{n}\chi_i^2$, where χ_i is the ith PR defined in Eq (2.7).

$MEP = \sum_{i=1}^{n}\frac{V^{1/2}(\mu_i)\chi_i^2}{(1-h_{ii})n}$, where $V(\mu_i) = \mu_i^2$ is the variance function of the GR model.

$AIC = \frac{-l_i(M_K)+2p}{n}$, where l_i is the log likelihood function of the GR model defined in Eq (2.3).

$R^2_{Efron} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$,

$\hat{\phi} = \frac{\chi^2}{n-p-1}$.

3. Outlier detection method

This section comprises two subsections. In the first subsection, the proposed outlier detection methods based on the PRs of the GR model are presented. The second one comprises a review of some existing outlier detection methods.

3.1. Proposed outlier detection methods

The PRs have a significant contribution in regression diagnostics. Here we propose an outlier detection method based upon PRs.

3.1.1. Standardized PRs

In regression analysis, the standardized residuals are generally used for the detection of outliers. So, here we consider standardized PRs, which have been defined in Section 2 by Eq (2.8) as

$$\chi_i^{'} = \frac{\chi_i}{\sqrt{\phi(1-h_{\rm ii})}}. \tag{3.1}$$

On the basis of standardized PRs, the ith point is considered to be an outlier if $\left|\chi_i^{'}\right| > 3$.

3.1.2. Jackknife PRs

There are some analytical methods that are used for the detection of outliers. One of these analytical methods is the use of jackknife residuals. Cook and Weisberg [37] suggested that the outliers can be detected with the help of jackknife residuals. They defined the jackknife residuals for LRMs as

$$e_{Ji} = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}, \tag{3.2}$$

where r_i is the standardized residual of the LRMs. The decision rule for the detection outlier is that if $|e_{Ji}| > t_{\frac{\alpha}{2n}}$ (with $n-p-1$ degrees of freedom), then there is an indication for the existence of outliers. The application of these residuals for the detection of outliers in chemometrics with reference to an LRM has been studied by Meloun and Militky [2]. Now, we modify Eq (3.2) for the GR by following Amin et al. [1] and obtaining

$$\chi_{Ji} = \chi_i^{'} \sqrt{\frac{n-p-1}{n-p-\chi_i^{'2}}}. \tag{3.3}$$

To identify outliers of the GR, we propose the cut-off point for the jackknife PR to be if $\left|\chi_{J,i}\right| > t_{(1-\alpha)}(n-p-1)$, then the ith observation is declared as an outlier t is the student t-distribution with $(n-p-1)$ degrees of freedom.

3.1.3. Adjusted PRs

Amin et al. [1] proposed the adjusted PRs in the inverse Gaussian regression for the detection of single influential points in chemometrics. These residuals are proposed for the GR as follows:

$$\chi_i^A = \frac{\chi_i - r_i}{\sqrt{v_i}}, \tag{3.4}$$

where $r_i = (E(R_i))^T = -\frac{\sqrt{\phi}}{2}(\mathbf{I}-\mathbf{H})J\mathbf{z}$, where $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X^T}\mathrm{WX})^{-1}\mathbf{X^T}\mathbf{W}^{\frac{1}{2}}$, $\mathbf{J} = diag\left(2\mu^2\right)$ and $\mathbf{z} = (z_{11},\ldots,z_{\mathrm{nn}})^T$is a vector; also, $v_i = (\mathrm{Var}(R_i))^T = 1 + \frac{\phi}{2}(\mathrm{QHJ}-\mathbf{T})\mathbf{z}$, where $\mathbf{Q} = diag(2)$ and $\mathbf{T} = diag\left(\left(2\phi^{-1}+6\right)\mu^2\right)$.

Note that $\widehat{r_i}$ and $\widehat{v_i}$ are computed by using $\widehat{\mu_i}$ instead of μ_i.

Amin et al. [1] stated that the adjusted residuals can be used for the detection of an outlier. The decision rule to declare the ith observation is an outlier is that $\left|\chi_i^A\right| > 2$.

3.2. Available outlier detection methods

In the literature, numerous methods have been recommended for the detection of outliers. We consider a few of them for comparison with the adjusted PR for the GR model.

3.2.1. Z-method

For the identification of outliers in univariate cases, the Z-score is introduced based on the median and inter-quartile range (IQR) as below:

$$z_i = \frac{y_i - \mathrm{Median}(y_i)}{IQR(y_i)}. \tag{3.5}$$

The Z-score method declares that the *i*th observation is an outlier if $|Z| > 3$.

3.2.2. Modified Z-method

The modified Z-statistic (MZS) has been proposed for the detection of outliers based on the median [38]; MZS is defined as

$$Z_i^* = \frac{y_i - Median(y_i)}{\text{Median}\,|y_i - Median(y_i)|}, \tag{3.6}$$

where Median|.| represents the median absolute deviation from the median. One can conclude that the *i*th observation is an outlier if $Z_i^* > 3.50$.

3.2.3. Grubb's test

This test was introduced by Grubbs [13] for the detection of a single outlier in the univariate response variable. The Grubb's (G) test only detects single outliers. Therefore, it suspects that most of the observations are outliers. The G statistic can be defined as $G = \frac{\text{Max}|y_i - \overline{y}|}{S}$, where s is the sample standard deviation. For our assumed model, the G statistic is modified as

$$G = \frac{|y_i - \overline{y}|}{S}. \tag{3.7}$$

The decision rule of the G statistic for detecting outliers is given as $\text{G} > \sqrt{\frac{t^2_{\left(\frac{\alpha}{2n}\right),n-2}}{n-2+\left(t^2_{\left(\frac{\alpha}{2n}\right),n-2}\right)}}$, where α is the level of significance and n represents the sample size. The above decision rule may be unable to find the appropriate outlier. So, we propose another decision rule for Grubb's method to diagnose an outlier a GR model. For the *i*th data point, if $G \geq 2$, then this data point is declared as an outlier.

4. Results and discussion

A comparison of the proposed methods of outlier detection with already available methods through the use of simulations and a real-life data set is presented in this section.

4.1. Simulation study

This section explains the simulation experiment conducted to study the performance of two types of GR residuals, i.e., the standardized and adjusted residuals to detect outliers. We generated the dependent variable of the GR model with reciprocal link function to be $y_i \sim \text{Gamma}\,(\mu_i, \phi)\,; i = 1, ..., n$, where $\mu_i = (\beta_0 + \beta_1 x_{i1} + ... + \beta_4 x_{i4})^{-1}$, where x_i's are generated from two probability distributions, i.e., uniform (0, 1) and standard normal distributions. The true values of the regression parameters vector β were selected as the normalized eigenvector corresponding to the largest eigenvalue of the $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ matrix such that $\beta^T\beta$=1 [39]. The sample sizes were generated as $n = 25, 50, 75, 100, 125, 150, 175$ *and* 200 and we took on the values of dispersion, i.e., $\phi = 0.33, 0.67$ *and* 2. A single outlier was generated for the dependent variable, i.e., the 8^{th} observation was replaced as $y_8 = y_8 + a_0$, where $a_0 = \overline{y_i} + 3\,(V\,(y_i))$. The multiple outliers in the dependent variable are generated as $y_{ii} = y_{ii} + a_0$, where ii= 8, 15, 20 three outliers. The performance

of these diagnostics was assessed by using gamma—produced samples for the identification of a generated outlier(s). The simulated results were computed with the help of the R-statistical language. The simulation study was replicated 1000 times to find the outliers in percentages.

In Table 1, the performance of our proposed methods is gauged for the detection of outliers based on the standard normal generated x's, and by using PRs with dispersion and sample sizes. It can be observed that for $\phi < 1$, the performance of the χ_i^A method was better than that of the other methods in diagnosing the generated single outlier. For this dispersion, when sample sizes were increased to 100, the performance of the $\chi_i^{'}$, χ_i^A and χ_J outlier detection methods was improved. The performance of the Z, Z^* and G outlier diagnostic methods was not affected by the increase in sample size. It can also be observed that the performance of all outlier diagnostic methods increased with increasing in dispersion. Moreover, when $\phi = 2$ and $n > 50$, the performance of all the diagnostic methods seems to have been identical. This indicates that sample size and dispersion have some significant effect on our proposed method χ_i^A in the detection of a single outlying point. The comparison of the outlier diagnostic methods revealed that χ_i^A is better than other methods. It can be seen that the detection was better with the proposed method than with all of the other diagnostic methods, when the values of X were generated from $U(0,\ 1)$. For further details, see Table 2. It can be seen that the performances of the available methods were not good as compared to the performance of our proposed method.

When we applied these methods for the detection of multiple outlying points with standard normally generated independent variables, the performance of our proposed method seemed to be better and more consistent than those of the other methods. The detection performance of the other diagnostic methods was reduced to 50%. On the other hand, the outlying diagnostic performance of all methods was reduced to some extent when the independent variables were generated from the uniform distribution. The performance of all outlier diagnostic methods increased rapidly as the dispersion crossed 1.0 (see Tables 1–4). We found that our proposed method performed better than the other methods including the case when the X's were generated from the standard normal distribution instead of the x's being generated from the uniform distribution.

Based on our findings (Tables 1–4), we can rank the methods as the adjusted PR being the first and the Grubb method being the second best method for outlier detection. Moreover, the results show that the performance of outlier detection increases with an increase in sample size. In studying the effect of dispersions on the outlier detection methods, we have found that outlier detection methods are affected directly by the dispersion. This means that upon increasing the dispersions, the outlier detection efficiencies of different methods are increased. The maximum numbers outliers were detected for larger dispersions.

Table 1. Comparison for single outlier detection methods for single outlier detection when $x^{'}s \sim N(0, 1)$.

ϕ	n	$\chi^{'}$	χ^{A}	χ_J	Z	Z^*	G
0.33	25	0.00	89.40	26.70	18.60	28.30	61.60
	50	0.50	93.30	34.70	13.90	25.20	64.40
	75	1.80	97.90	42.30	15.70	26.00	68.90
	100	1.70	98.60	42.80	12.50	24.00	69.90
	125	3.00	99.00	42.40	13.50	24.50	67.20
	150	3.90	99.00	43.90	12.40	23.60	66.70
	175	4.80	99.20	44.10	14.20	24.40	67.90
	200	4.70	99.40	44.10	13.30	25.40	68.20
0.67	25	0.10	99.30	56.00	41.10	54.10	91.30
	50	8.80	100.00	76.80	39.50	57.60	95.00
	75	15.30	100.00	84.50	39.60	58.40	95.10
	100	20.60	100.00	87.50	38.80	58.60	96.60
	125	24.30	100.00	87.60	37.20	55.80	97.00
	150	30.50	100.00	87.90	41.60	60.70	96.60
	175	26.70	100.00	88.70	37.40	59.20	96.30
	200	30.30	100.00	89.40	38.60	59.40	96.10
2.00	25	19.60	100.00	99.30	95.70	98.30	100.00
	50	87.30	100.00	100.00	98.20	99.60	100.00
	75	96.60	100.00	100.00	98.20	100.00	100.00
	100	98.70	100.00	100.00	98.10	99.70	100.00
	125	99.50	100.00	100.00	99.40	100.00	100.00
	150	99.70	100.00	100.00	99.40	99.90	100.00
	175	99.50	100.00	100.00	99.50	99.90	100.00
	200	99.80	100.00	100.00	98.60	100.00	100.00

Table 2. Comparison for single outlier detection methods for single outlier detection when $x^{'}s \sim U(0, 1)$.

ϕ	n	$\chi^{'}$	χ^{A}	χ_J	Z	Z^*	G
0.33	25	0.00	93.60	27.80	15.10	23.80	54.30
	50	0.10	94.90	33.50	12.60	22.30	62.30
	75	2.20	97.60	34.60	10.30	21.20	61.70
	100	2.50	98.00	38.10	12.90	23.00	62.50
	125	2.70	98.30	41.10	12.40	22.30	65.10
	150	2.80	98.40	38.60	10.90	21.60	63.60
	175	4.00	98.30	42.00	14.70	24.30	63.40
	200	2.80	97.60	39.10	9.90	22.00	62.80
0.67	25	0.30	99.70	62.00	35.30	49.00	88.20
	50	7.50	99.90	73.80	36.20	53.40	93.70
	75	13.20	100.00	78.30	33.20	51.50	93.60
	100	20.00	100.00	82.70	34.70	51.70	94.40
	125	18.00	100.00	84.30	33.10	53.70	95.80
	150	22.60	100.00	86.30	34.40	54.20	96.30
	175	23.60	100.00	83.50	32.90	51.40	95.10
	200	24.10	100.00	85.50	33.60	53.20	94.80
2.00	25	19.60	100.00	99.70	93.80	97.20	100.00
	50	84.80	100.00	100.00	95.70	99.00	100.00
	75	95.90	100.00	100.00	97.00	99.50	100.00
	100	97.30	100.00	100.00	97.50	99.50	100.00
	125	98.90	100.00	100.00	97.90	99.80	100.00
	150	99.30	100.00	100.00	98.90	99.90	100.00
	175	99.70	100.00	100.00	99.20	100.00	100.00
	200	99.40	100.00	100.00	99.00	99.80	100.00

Table 3. Comparison of outlier detection methods for multiple-outlier detection when $x's \sim N(0, 1)$.

ϕ	n	χ'	χ^A	χ_J	Z	Z^*	G
0.33	25	0.00	85.80	14.63	8.50	16.23	34.60
	50	0.20	95.33	27.50	10.33	20.73	51.53
	75	0.90	97.10	31.37	10.63	21.30	55.60
	100	1.73	98.27	35.50	11.33	21.90	60.40
	125	2.60	98.73	40.80	12.53	24.10	63.07
	150	3.53	98.50	40.97	11.63	21.87	63.70
	175	3.90	99.10	42.30	11.77	23.30	64.73
	200	3.93	98.93	41.50	11.43	22.23	63.80
0.67	25	0.00	98.33	28.30	22.87	37.60	53.53
	50	1.30	99.93	58.83	32.20	49.83	81.63
	75	5.73	100.00	69.07	32.60	52.30	88.57
	100	10.07	100.00	75.57	34.87	55.30	91.40
	125	14.90	100.00	80.83	34.90	55.77	93.37
	150	17.43	100.00	83.67	35.60	56.87	94.13
	175	19.77	100.00	84.17	34.07	54.33	93.97
	200	21.57	100.00	86.30	36.53	57.83	94.87
2.00	25	0.00	100.00	57.97	86.07	93.50	87.40
	50	12.03	100.00	96.07	95.77	98.80	99.83
	75	48.13	100.00	99.83	97.50	99.50	100.00
	100	74.57	100.00	99.87	98.50	99.73	100.00
	125	87.00	100.00	99.97	98.77	99.93	100.00
	150	93.03	100.00	100.00	98.87	99.87	100.00
	175	95.53	100.00	100.00	99.20	99.97	100.00
	200	97.17	100.00	100.00	99.17	99.93	100.00

Table 4. Comparison of outlier detection methods for multiple-outlier detection when $x's \sim U(0,1)$.

ϕ	n	χ'	χ^A	χ_J	Z	Z^*	G
0.33	25	0	76.57	23.63	8.37	15.8	31.5
	50	0.27	93.67	31.1	9.33	18.8	47.07
	75	0.77	95	32.8	9.23	19.23	51.63
	100	1.7	96.97	34.53	9.13	18.03	55.23
	125	1.87	97.3	37.1	9.77	19.4	57
	150	2.33	97.93	39	10	21.03	59.37
	175	3.1	98.3	40.13	10.53	21.13	60.13
	200	2.87	98.4	39.47	10.27	21.07	61.17
0.67	25	0	89.93	40.97	20.83	33.23	50.87
	50	2.37	99.83	62.37	26.13	43.07	78.47
	75	7.07	99.93	69.5	28.67	46.63	85.27
	100	11.1	100	75.17	29.57	47.67	88.6
	125	12.93	100	78.03	30.73	49.73	90.7
	150	15.6	100	79.97	29.67	48.4	92.1
	175	16.77	100	81.4	31.3	51	91.67
	200	19.73	100	83.9	31.43	52.37	93.13
2	25	0.2	98.87	61.93	80.83	90.67	85.1
	50	23.57	100	95.03	92.93	97.43	99.8
	75	53.6	100	99.17	95.6	99.03	99.93
	100	73.97	100	99.77	96.63	99.43	99.9
	125	83.43	100	99.77	97.37	99.8	99.97
	150	90.17	100	100	97.6	99.57	100
	175	92.73	100	100	98.33	99.77	100
	200	95.13	100	100	98.37	99.73	100

4.2. Application: ARDENNES dataset

Now, we will evaluate the performance of the proposed methods with the help of a real application. For this purpose, we applied the *ARDENNES* data taken from Barnard et al. [40]. The main use for this data set was to determine the first etch biopsy, i.e., in the beginning of a layer of extracted incisor enamel (y) based on two explanatory variables for the data on 55 children. These explanatory variables included the etched depth (x_1), which was estimated from the amount of calcium removed for the duration of the etch biopsy as the first explanatory variable. The age of the child (x_2) which had been transformed to the decimal system from years and months was considered as the second explanatory variable. Mallet et al. [41] applied linear regression and other models to this data set using log (y) as a response variable. After fitting the LRM, they explored the outlier detection analysis of this data and found that the 23rd, 48th and 52nd points were the outliers. However, this data set is not well fitted to the normal distribution since the trend of the dependent variable is positively skewed. From the distribution of the fitting test, we observed that the GR model is well fitted to this data set, the results are reported in Table 5.

Table 5. Fitting the Probability Distributions of the Response Variable.

Distribution Fitting	Normal		Gamma		Inverse Gaussian		Weibull	
Tests	Statistic	P-Value	Statistic	P-Value	Statistic	P-Value	Statistic	P-Value
Anderson-Darling	1.0945	0.0066	0.3553	**0.4699**	0.437	0.4152	0.5883	0.1279
Cramér-von Mises	0.1405	0.0311	0.0539	**0.4617**	0.0872	0.2991	0.0725	0.2542
Pearson chi-square	10.6364	0.1553	5.5454	**0.5937**	7	0.4288	7.7273	0.3572

So the appropriate regression model to determine the etch biopsy based on these two explanatory variables is the GR model.

The fitted GR is given by

$$\hat{\mu}_i = (0.00073[0.0008, N] + 0.0003x_1[0.00012, S] - 0.0009x_2[0.000083, N])^{-1},$$

where the square brackets contain the standard errors of the estimated parameters. The letter N represents the non-significance and S represents the significance of the regression coefficients. The fitted GR model achieved the following results MEP=267.08, AIC=847.19, and Pearson chi-square statistic= 11.61. After fitting the GR model, we next computed the outlier statistics, which are plotted in Figure 1.

From Figure 1, we can observe that the proposed outlier detection methods detected the 9$^{\text{th}}$, 23$^{\text{rd}}$, 48$^{\text{th}}$ and 52$^{\text{nd}}$ points as outliers. The 9$^{\text{th}}$ point was not detected as an outlier in the original work due to misidentification of the regression model but this did not affect the GR estimates. Hossain and Naik [42] also indicated that an outlier may or may not affect the regression estimates. So, in this case this outlying point had minimal effect on the GR estimates while R^2_{Efron} also decreased because outliers are related to the response variable. R^2_{Efron} only improved due to elimination of the influential points which are related to the explanatory variables for more details see [5,43,44]. All detected outliers indicated that these children had high lead levels in their enamel, but there were no extreme points in the explanatory variables. Moreover, these outliers were also identified successfully by using the Jackknife residuals and adjusted PRs.

After deleting the identified outliers, the refitted GR model was given by

$$\hat{\mu}_i = (0.00183[0.00069, S] + 0.0003x_1[0.0001, S] - 0.0002x_2[0.000071, S])^{-1},$$

with $MEP = 169.85$, $AIC = 756.15$ and $Pearson\ chi - square\ statistic = 6.89$. These results indicate how much the variation decreased after deleting the identified outliers, e.g., the Pearson chi-square statistic (variation measure) was reduced to 59% (see Table 6). Another noticeable point is that the two independent variables were found to be significant after deleting the identified outliers. From Table 6, it can be observed that an individually high outlier is the 48$^{\text{th}}$ point, which affected the fitted GR results. Collectively, all detected outlying points affected the GR estimates of β_0 and β_2 respectively.

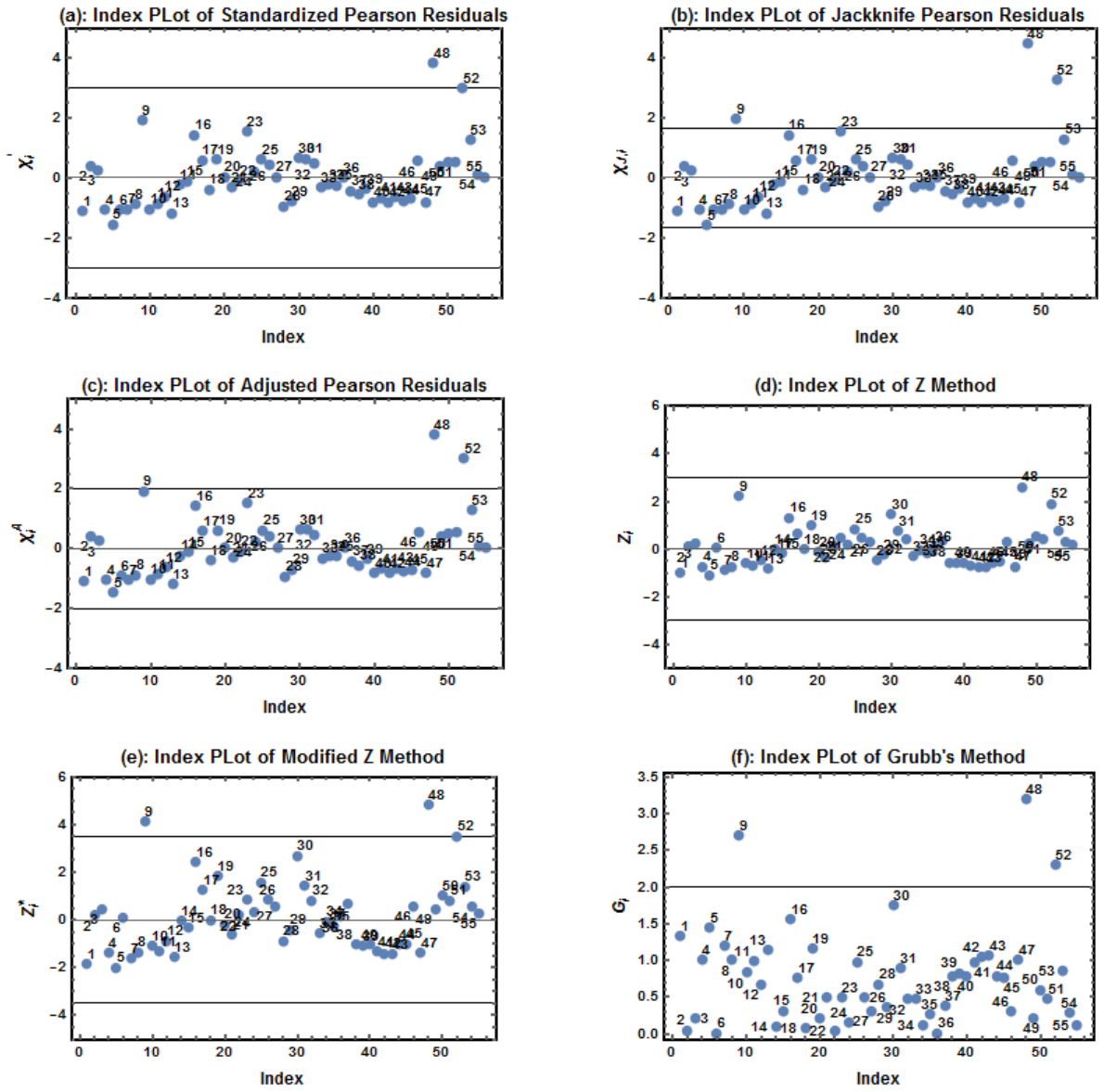

Figure 1. Index plots for outlier detection methods.

Table 6. Change (%) in the GR results after deleting the outlying points.

Outlying Points	β_0	β_1	β_2	MEP	$\hat{\phi}$	R^2_{Efron}	AIC
9	5.78	-9.06	-6.62	-4.24	-2.91	-12.28	-2.27
23	15.56	9.04	23.87	0.23	1.8	4.39	-1.96
48	**72.49**	**-0.34**	**66.86**	**-2.27**	**-19.58**	**52.58**	**-3.15**
52	22.84	8.27	28.64	-0.23	-9.25	27.63	-2.64
9,23,48,52	149.68	9.78	145.04	3.04	-48.02	119.11	-10.91

5. Conclusions

Outlier detection in regression models is an important step to getting reliable and valid results. These detections methods are based on some diagnostic test statistics, which can be calculated based on the regression results. To detect the outliers, the first and most important step is the choice of an appropriate regression model, because, sometimes, outliers may arise due to an inappropriate regression model. For the selection of an appropriate regression model, one should test the distribution of the response variable. If the probability distribution of the dependent variable is gamma, the appropriate choice of model is the GR model. In this paper, we proposed outlier diagnostics based on the use of the Pearson (standardized and adjusted) residuals in GR model. Some available LRM-based outlier detection methods were modified for the GR model. These modified methods were compared with our proposed outlier detection methods with the help of simulations and a real data set. These results indicate that our proposed methods for the detection of outliers are better than available methods in terms of improving the results of the selected model to enable better decisions in statistics and other disciplines.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by (Grant Code 22UQU4310063DSR05).

Conflict of interest

The authors declare that they have no conflicts of interest regarding the publication of this article.

References

1. M. Amin, M. Amanullah, M. Aslam, Empirical evaluation of the inverse Gaussian regression residuals for the assessment of influential points, *J. Chemometr.*, **30** (2016), 394–404. https://doi.org/10.1002/cem.2805
2. M. Meloun, J. Militký, Detection of single influential points in OLS regression model building, *Anal. Chim. Acta*, **439** (2001), 169–191. https://doi.org/10.1016/S0003-2670(01)01040-6
3. K. A. Mogaji, Geoelectrical parameter-based multivariate regression borehole yield model for predicting aquifer yield in managing groundwater resource sustainability, *J. Taibah Univ. Sci.*, **10** (2016), 584–600. https://doi.org/10.1016/j.jtusci.2015.12.006
4. O. S. Alshamrani, Construction cost prediction model for conventional and sustainable college buildings in North America, *J. Taibah Univ. Sci.*, **11** (2017), 315–323. https://doi.org/10.1016/j.jtusci.2016.01.004
5. A. M. Sarhan, A. I. El-Gohary, A. Mustafa, A. H. Tolba, Statistical analysis of regression competing risks model with covariates using Weibull sub-distributions, *Int. J. Reliab. Appl.*, **20** (2019), 73–88.
6. J. Burger, P. Geladi, Hyperspectral NIR image regression part II: Dataset preprocessing diagnostics, *J. Chemometr.*, **20** (2006), 106–119. https://doi.org/10.1002/cem.986

7. D. L. Massart, L. Kaufman, P. J. Rousseeuw, A. Leroy, Least median of squares: A robust method for outlier and model error detection in regression and calibration, *Anal. Chem. Acta*, **187** (1986), 171–179. https://doi.org/10.1016/S0003-2670(00)82910-4
8. E. Hund, D. L. Massart, J. Smeyers-Verbeke, Robust regression and outlier detection in the evaluation of robustness tests with different experimental designs, *Anal. Chem. Acta.*, **463** (2002), 53–73. https://doi.org/10.1016/S0003-2670(02)00337-9
9. P. J. Rousseeuw, M. Debruyne, S. Engelen, M. Hubert, Robustness and outlier detection in chemometrics, *Crit. Rev. Anal. Chem.*, **36** (2006), 221–242. https://doi.org/10.1080/10408340600969403
10. A. Desgagné, Efficient and robust estimation of regression and scale parameters, with outlier detection, *Comput. Stat. Data Anal.*, **155** (2021), 1–19. https://doi.org/10.1016/j.csda.2020.107114
11. V. Barnett, T. Lewis, *Outliers in statistical data*, Chichester, UK: Wiley, 1994.
12. W. J. Dixon, Analysis of extreme values, *Ann. Math. Stat.*, **21** (1950), 488–506.
13. F. E. Grubbs, Procedures for detecting outlying observations in samples, *Technometrics*, **11** (1969), 1–21.
14. B. Rosner, Percentage points for a generalized ESD many-outlier procedure, *Technometrics*, **25** (1983), 165–172.
15. U. Balasooriya, Y. K. Tse, Y. S. Liew, An empirical comparison of some statistics for identifying outliers and influential observations in linear regression models, *J. Appl. Stat.*, **14** (1987), 177–184. https://doi.org/10.1080/02664768700000022
16. J. F. Lawless, *Statistical models and methods for life time data*, New York: Wiley, 2003.
17. D. Jearkpaporn, D. C. Montgomery, G. C. Runger, C. M. Borror, Model based process monitoring using robust generalized linear models, *Int. J. Prod. Res.*, **43** (2005), 1337–1354. https://doi.org/10.1080/00207540412331299693
18. M. L. Segond, C. Onof, H. S. Wheater, Spatial temporal disaggregation of daily rainfall from a generalized linear model, *J. Hydrol.*, **331** (2006), 674–689. https://doi.org/10.1016/j.jhydrol.2006.06.019
19. R. N. Das, J. Kim, GLM and joint GML techniques in hydrogeology: An illustration, *Int. J. Hydrol. Sci. Technol.*, **2** (2012), 185–201.
20. R. De Marco, F. Locatelli, I. Cerveri, M. Bugiani, A. Marinoni, G. Giammanco, Incidence and remission of asthma: A retrospective study on the natural history of asthma in Italy, *J. Allergy Clin. Immun.*, **110** (2002), 228–235. https://doi.org/10.1067/mai.2002.125600
21. M. Faddy, N. Graves, A. Pettitt, Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, Gamma and log-normal distributions, *Value Health*, **12** (2009), 309–314. https://doi.org/10.1111/j.1524-4733.2008.00421.x
22. Y. Murakami, T. Okamura, K. Nakamura, K. Miura, H. Ueshima, The clustering of cardiovascular disease risk factors and their impacts on annual medical expenditure in Japan: Community-based cost analysis using Gamma regression models, *BMJ Open*, **3** (2013), 1–6.

23. D. Griffie, L. James, S. Goetz, B. Balotti, Y. H. Shr, M. Corbin, et al., Outcomes and economic benefits of Penn State extension's dining with diabetes program, *Prev. Chronic Dis.*, **15** (2018), 1–13. https://doi.org/10.5888/pcd15.170407
24. N. Kumar, S. Lalitha, Testing for upper outliers in gamma sample, *Commun. Stat.-Theory Methods*, **41** (2012), 820–828. https://doi.org/10.1080/03610926.2010.531366
25. M. J. Nooghabi, H. J. Nooghabi, P. Nasiri, Detecting outliers in gamma distribution, *Commun. Stat. Theory Methods*, **39** (2010), 698–706. https://doi.org/10.1080/03610920902783856
26. A. C. Kimber, Tests for a single outlier in a gamma sample with unknown shape and scale parameters, *J. Roy. Stat. Soc. Ser. C*, **28** (1979), 243–250. https://doi.org/10.2307/2347194
27. A. C. Kimber, Discordancy testing in gamma samples with both parameters unknown, *J. Roy. Stat. Soc. Ser. C*, **32** (1983), 304–310. https://doi.org/10.2307/2347953
28. T. Lewis, N. R. J. Fieller, A recursive algorithm for null distribution for outliers: I. Gamma samples, *Technometrics*, **21** (1979), 371–376.
29. M. A. Shayib, D. H. Young, The extreme residuals in gamma regression, *Commun. Stat. Theory Methods*, **20** (1991), 561–577. https://doi.org/10.1080/03610929108830515
30. G. C. Tiao, I. Guttman, Analysis of outliers with adjusted residuals, *Technometrics*, **9** (1967), 541–559.
31. G. M. Cordeiro, On Pearson's residuals in generalized linear models, *Stat. Probabil. Lett.*, **66** (2004), 213–219. https://doi.org/10.1016/j.spl.2003.09.004
32. M. R. Urbano, C. G. Demtrio, G. M. Cordeiro, On Wald residuals in generalized linear models, *Commun. Stat. Theory Methods*, **41** (2012), 741–758. https://doi.org/10.1080/03610926.2010.529537
33. T. Anholeto, M. C. Sandoval, D. A. Botter, Adjusted Pearson residuals in beta regression models, *J. Stat. Comput. Simul.*, **84** (2014), 999–1014. https://doi.org/10.1080/00949655.2012.736993
34. J. W. Hardin, J. W. Hilbe, *Generalized linear models and extensions*, Stata Press Publication: Texas, 2012.
35. P. J. Green, Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, *J. Roy. Stat. Soc.: Ser. B*, **46** (1984), 149–170. https://doi.org/10.1111/j.2517-6161.1984.tb01288.x
36. D. R. Cox, E. J. Snell, A general definition of residuals (with discussion), *J. Roy. Stat. Soc.: Ser. B*, **30** (1968), 248–275.
37. R. D. Cook, S. Weisberg, *Residuals and influence in regression*, Chapman Hall, New York, 1982.
38. B. Iglewicz, D. C. Hoaglin, *How to detect and handle outliers*, Milwaukee: ASQC Quality Press, 1993.
39. S. Ahmad, M. Aslam, Another proposal about the new two-parameter estimator for linear regression model with correlated regressors, *Commun. Stat.-Simul. Comput.*, **51** (2022), 3054–3072. https://doi.org/10.1080/03610918.2019.1705975
40. T. E. Barnard, K. S. Booksh, R. G. Brereton, D. H. Coomans, S. N. Deming, Y. Hayashi, *Chemometrics in environmental chemistry-statistical methods*, Vol. 2, Springer-Verlag Berlin Heidelberg New York, 1995.

41. Y. L. Mallet, D. H. Coomans, O. Y. de Vel, Robust non-parametric methods in multiple regressions of environmental data, In: *Chemometrics an environmental chemistry-statistical methods*, 1995. https://doi.org/10.1007/978-3-540-49148-4_6

42. A. Hossain, D. N. Naik, A comparative study on detection of influential observations in linear regression, *Stat. Pap.*, **32** (1991), 55–69. https://doi.org/10.1007/BF02925479

43. T. A. Abushal, Parametric inference of Akash distribution with Type-II censoring with analyzing of relief times of patients, *AIMS Math.*, **6** (2021), 10789–10801. https://doi.org/10.3934/math.2021627

44. T. A. Abushal, A. H. Abdel-Hamid, Inference on a new distribution under progressive-stress accelerated life tests and progressive type-II censoring based on a series-parallel system, *AIMS Math.*, **7** (2022), 425–454. https://doi.org/10.3934/math.2022028

AIMS Press