STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

An alternative Pearson residual-based method for outlier detection in a gamma regression model

Wuttichai Srisodaphol*, Lapasrada Polchumni

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand 40002

Abstract This study develops six novel outlier-detection methods for Gamma regression (TSPR, TJPR, TAPR, GASPR, GAJPR, and GAAPR) that combine Pearson residuals scaling with Tukey's boxplot rules and grouped absolute residuals, where thresholds are adapted from the residual behavior within successive groups of observations. These methods address outlier detection when the response is nonnegative and right-skewed, a common situation in applied settings. We compare the new procedures to six existing methods (SPRs, JPRs, APRs, Z, Z*, and G) using three performance metrics: pout (probability of detecting all true outliers), pmask (probability of masking true outliers as inliers), and pswamp (probability of misclassifying inliers as outliers). Performance is evaluated via simulation (uncontaminated data and contamination at 5% and 10%) and on a real dataset. Results show that GAJPR and GAAPR achieve the best detection power in simulations, while GASPR and GAJPR perform best on the real data; overall, GAJPR is the most effective method. The grouped absolute residuals approach prioritizes sensitivity and reduces masking but tends to increase false positives, so we recommend grouped-absolute-residuals screening followed by casewise validation or conservative re-testing before excluding observations.

Keywords Gamma regression, Outlier detection, Pearson residuals, Tukey's boxplot

AMS 2010 subject classifications 62J12, 62J20, 62P10

DOI: 10.19139/soic-2310-5070-3478

1. Introduction

Regression analysis is a tool used to describe the relationship between one or more response variables and one or more independent variables [2]. It is commonly used in research in many fields, such as biology, agriculture, and society. If the response variable data is right-skewed or non-negative, the gamma regression model is used to describe the relationship between the response variable and the independent variable [6, 11, 15]. Data suitable for modeling with the gamma regression model are available in various fields, such as studied the risk factors for cardiovascular disease that affect medical expenses of the population in Japan classified by gender and age. There are four cardiovascular risk factors: hypertension, high blood cholesterol, high blood sugar, and smoking [12], studied the factors affected the amount of dissolved oxygen, which consisted of four factors: total dissolved solids, magnesium, chloride content, and chemical oxygen demand [9], and studied the long-term unemployment rate and the index of labor market instability affecting the homicide rate [1].

A gamma regression model is a model that describes the relationship between a response variable (Y) and independent variables (X) when the response variable has a gamma distribution, which is right-skewed, or the value of the random variable Y is non-negative [15]. The probability density function of the gamma random variable (Y)

*Correspondence to: Wuttichai Srisodaphol (Email: wuttsr@kku.ac.th). Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand 40002.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

with shape parameters α and scale parameter β is given by

$$f(y|\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}y^{\alpha-1}exp(-\frac{y}{\beta});y>0,\alpha>0,\beta>0. \tag{1}$$

The mean and variance of Y are $E(Y)=\alpha\beta$ and $Var(Y)=\alpha\beta^2$.

According to Hardin and Hilbe [10], eq.(1) can be transformed with parameters $\alpha=\frac{1}{\phi}$ and $\beta=\mu\phi$, so eq.(1) has a probability density function given by

$$f(y|\mu,\phi)=\frac{1}{\Gamma(\frac{1}{\phi})}(\frac{1}{\mu\phi})^{\frac{1}{\phi}}y^{\frac{1}{\phi}-1}exp(-\frac{y}{\mu\phi});y>0,\mu>0,\phi>0. \tag{2}$$

Thus, the mean and variance of Y are $E(Y)=\mu$ and $Var(Y)=\phi\mu^2$.

The response variable has a gamma distribution, which is part of the exponential family. The gamma regression model consists of a random component Y, a systematic component or linear predictor η, and a link function $g(.)$, with a linear relationship between the random component converted to the link function and the systematic component. The mean function of the gamma regression model with the inverse link function [2] is given as

$$g(\mu_i)=\frac{1}{\mu_i}=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+...+\beta_pX_{ip};i=1,2,...,n. \tag{3}$$

where $X_{i1},X_{i2},...,X_{ip}$ are the i^{th} observation of the p independent variables,

β_0 is the Y-intercept when all the independent variables equal 0, which is the parameter of the regression model,

$\beta_1,\beta_2,...,\beta_p$ are the partial regression coefficients,

p is the number of independent variables.

Consequently, the initial assumptions in the gamma regression model are not as stringent as those in multiple linear regression. For instance, the errors are independent but are not normally distributed; however, the model assumes heteroscedasticity (where variance is proportional to the square of the mean) rather than homoscedasticity [15]. Multiple linear regression models use sample data to build the regression model, but sample data may contain outliers, which are observations that deviate significantly from other observations. Common causes of outliers include data entry errors, measurement errors, experimental errors, intentional (creation of outliers), data processing errors, sampling errors, and outliers that occur naturally in the data and are not errors [16]. Although outliers are often discarded, they can provide valuable insights and deserve further investigation. For example, some women live well with HIV for many years without treatment, which is considered an abnormality. Compared to most women who do not receive treatment and die quickly [8], this can be further investigated by detecting outliers. The multiple linear regression model has outlier detection methods, including scatter graph, boxplot, Williams graph, Rankit graph (Q-Q plot), graph of predicted residuals, predicted residuals, standardized residuals, studentized residuals, Jackknife residuals, Cook's distance, Different-in-fits (DFFITS), and Atkinson's measure [4].

In the gamma regression model, Amin et al. [2] presented an outlier detection method that utilizes the principle of Pearson residuals, specifically standardized Pearson residuals, Jackknife Pearson residuals, and adjusted Pearson residuals. The proposed methods are compared with outlier detection methods in the response variable, specifically the Z-method, modified Z-method, and Grubb's test, through data simulation and application to real data. The Z-method (Z) for detecting outliers using $|Z|>3$ where $Z_i=\frac{y_i-Median(y_i)}{IQR(y_i)}$. The modified Z-method (Z^*) for detecting outliers using $Z_i^*>3.5$ where $Z_i^*=\frac{y_i-Median(y_i)}{Median|y_i-Median(y_i)|}$. Grubb's test (G) for detecting outliers using $G\geq 2$ where $G=\frac{|y_i-\bar{y}_i|}{s}$ and $s=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$. The performance of each method was evaluated based on the percentage of outliers detected. The results indicated that the adjusted Pearson residuals method showed the highest performance in detecting outliers in simulated data. In contrast, when applied to real data, the Jackknife Pearson residuals method detected outliers more accurately than the adjusted Pearson residuals method. In data simulation, the efficiency of each method is measured by considering the percentage of outlier detection results. However,

considering only the percentage of outlier detection results in data simulation is not enough because it cannot be known whether the outlier detection results obtained are complete, exceed, or fall short of the specified outlier, and whether the outliers are true outliers. From the above, this study develops new methods for detecting outliers in the gamma regression model. The principles of Pearson residuals and Tukey's box plot [18]. Tukey's box plot is used with univariate data. In this research, we utilize Tukey's box plot to identify outliers based on standardized Pearson residuals, Jackknife Pearson residuals, and adjusted Pearson residuals using the first three methods. In the other three methods, we used absolute Pearson residuals and grouping [17]. We utilize the absolute standardized Pearson residuals, absolute Jackknife Pearson residuals, and absolute adjusted Pearson residuals to divide the data into two groups. We use the residuals from Group 1 to determine the cut-off point. Also, these methods are compared with the existing methods, namely standardized Pearson residuals, Jackknife Pearson residuals, adjusted Pearson residuals, Z-method, modified Z-method, and Grubb's test, using the probability that all outliers are successfully detected (pout), the probability that outliers are falsely detected as inliers (pmask), and the probability that inliers are detected as outliers (pswamp) as criteria to measure the performance of the outlier detection method. These three criteria can indicate whether the outlier detection method effectively identifies complete, excessive, or insufficient outliers.

2. Proposed methods for detecting outliers

The study proposes six methods for detecting outliers using Pearson residuals in the gamma regression model. The principles of Pearson residuals, Tukey's box plot, and absolute grouping are used to develop methods. The procedure for developing methods to detect outliers in the gamma regression model is as follows:

2.1 Method 1 (TSPR): Tukey's Boxplot on Standardized Pearson Residuals

This method identifies potential outliers by constructing a Tukey boxplot on standardized Pearson residuals (SPRs), with the following steps:

Step 1: Fit the gamma regression model and calculate standardized Pearson residuals (SPRs) χ_i' for $i = 1, 2, ..., n$,

$$\chi_i' = \frac{\chi_i}{\sqrt{\phi\left(1 - h_{ii}\right)}},$$

where ϕ is the dispersion parameter of the responsive variables when $\phi = E\left(d_i\right)$, $E\left(d_i\right)$ is the expected value of the dispersion statistic when $d_i = \frac{\left(Y_i - \mu_i\right)^2}{V(\mu)}$,

μ_i is the predicted value ith obtained from the regression model,

$V\left(\mu\right)$ is the variance of the predicted values obtained from the regression model when in the gamma regression model $V\left(\mu\right) = \mu^2$,

h_{ii} is the ith member on the diagonal of the hat matrix,

$\chi_i = \frac{y_i - \mu_i}{\sqrt{V(\mu)}} = \frac{y_i - \mu_i}{\mu_i}$ is the Pearson residuals ith,

y_i is the observed value of the response variable ith,

Step 2: Calculate the first quartiles (Q_1) and third quartiles (Q_3) of the standardized Pearson residuals (SPRs), then calculate the interquartile range (IQR), $\mathrm{IQR} = Q_3 - Q_1$.

Step 3: Calculate the cut-off points using the upper fence and lower fence,

$$\text{Upper fence} = Q_3 + (1.5 \times IQR)$$

$$\text{Lower fence} = Q_1 - (1.5 \times IQR)$$

Step 4: Identify the outliers by considering the standardized Pearson residuals (SPRs) that are greater than the upper fence or less than the lower fence.

2.2 Method 2 (TJPR): Tukey's boxplot on Jackknife Pearson Residuals

This method identifies potential outliers by constructing a Tukey boxplot on Jackknife Pearson residuals (JPRs), with the following steps:

Step 1: Fit the gamma regression model and calculate Jackknife Pearson residuals (JPRs) χ_{Ji} for $i = 1, 2, ..., n$,

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

when χ' is the standardized Pearson residuals ith,

n is the sample size,

p is the number of independent variables.

Step 2: Calculate the first quartiles (Q_1) and third quartiles (Q_3) of the Jackknife Pearson residuals (JPRs), then calculate the interquartile range (IQR), $\text{IQR} = Q_3 - Q_1$.

Step 3: Calculate the cut-off points using the upper fence and lower fence.

$$\text{Upper fence} = Q_3 + (1.5 \times IQR)$$

$$\text{Lower fence} = Q_1 - (1.5 \times IQR)$$

Step 4: Identify the outliers by considering the Jackknife Pearson residuals (JPRs) that are greater than the upper fence or less than the lower fence.

2.3 Method 3 (TAPR): Tukey's boxplot on Adjusted Pearson Residuals

This method identifies potential outliers by constructing a Tukey boxplot on adjusted Pearson residuals (JPRs), with the following steps:

Step 1: Fit the gamma regression model and calculate adjusted Pearson residuals (APRs) χ_i^A for $i = 1, 2, ..., n$,

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

when χ_i is the Pearson residuals ith,

$r_i = (E(R_i))^T$ is the transpose of the expected value of Pearson residuals ith when $(E(R_i))^T = -\frac{\sqrt{\phi}}{2}(\mathbf{I} - \mathbf{H})\mathbf{Jz}$,

$v_i = (Var(R_i))^T$ is the variance of Pearson residuals ith when $(Var(R_i))^T = 1 + \frac{\phi}{2}(\mathbf{QHJ} - \mathbf{T})\mathbf{z}$,

R_i is Pearson residuals ith,

$\mathbf{I}$ is an identity matrix,

$\mathbf{H}$ is the hat matrix when $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\frac{1}{2}}$,

$\mathbf{W}$ is wights matrix when $\mathbf{W} = diag\left(\mu_i^2\right)$,

$\mathbf{X}$ is a matrix of independent variables,

$\mathbf{X}^{\mathrm{T}}$ is the transpose of a matrix of independent variables,

$\mathbf{J} = diag\left(2\mu^2\right)$,

$\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$,

$\mathbf{Q} = diag\,(2)$,

$\mathbf{T} = diag\left(\left(2\phi^{-1} + 6\right)\mu^2\right)$.

Step 2: Calculate the first quartiles (Q_1) and third quartiles (Q_3) of the adjusted Pearson residuals (APRs), then calculate the interquartile range (IQR), $\text{IQR} = Q_3 - Q_1$.

Step 3: Calculate the cut-off points using the upper fence and lower fence.

$$\text{Upper fence} = Q_3 + (1.5 \times IQR)$$

$$\text{Lower fence} = Q_1 - (1.5 \times IQR)$$

Step 4: Identify the outliers by considering the adjusted Pearson residuals (APRs) that are greater than the upper fence or less than the lower fence.

2.4 Method 4 (GASPR): Grouping Absolute Standardized Pearson Residuals

This method groups the absolute values of standardized Pearson residuals into quantile-based intervals. It

examines the group-wise frequencies and statistics to identify abnormal concentrations of large residuals, indicative of outlying observations.

Step 1: Fit the gamma regression model and calculate the standardized Pearson residuals (SPRs) χ_i' for $i = 1, 2, ..., n$, and find the absolute value $|\chi_i'|$.

Step 2: Divide $|\chi_i'|$ into two groups, Group 1: $G_1 = \{|\chi_i'|, \text{where } |\chi_i'| < 1\}$, and Group 2: $G_2 = \{|\chi_i'|, \text{where } |\chi_i'| \geq 1\}$ [17].

Step 3: Sort all $|\chi_i'|$ of Group 1 (G_1) in ascending order, $G_1 = \left\{\left|\chi'_{(1)}\right|, \left|\chi'_{(2)}\right|, ..., \left|\chi'_{(n_1)}\right|\right\}$ where n_1 is the number of members in Group 1, then calculate the maximum difference between consecutive $\left|\chi'_{(i)}\right|$ in this sequence, defined as

$$\Delta_{\max} = \max_{i=1}^{n_1-1} \left(\left|\chi'_{(i+1)}\right| - \left|\chi'_{(i)}\right|\right).$$

Step 4: Calculate the cut-off point (C) defined as

$$C = \Delta_{\max} + \left|\chi'_{(n_1)}\right|,$$

which identifies a clear breakpoint between normal and extreme data points.

Step 5: Move the $|\chi_i'|$ in Group 2 to Group 1 when the $|\chi_i'|$ are less than or equal to the cut-off point C.

Step 6: Repeat Steps 3–5 until $|\chi_i'|$ in Group 2 can no longer be moved into Group 1; the remaining members of Group 2 are considered outliers.

2.5 Method 5 (GAJPR): Grouping Absolute Jackknife Pearson Residuals

This method groups the absolute values of Jackknife Pearson residuals into quantile-based intervals. It examines the group-wise frequencies and statistics to identify abnormal concentrations of large residuals, indicative of outlying observations.

Step 1: Fit the gamma regression model and calculate the Jackknife Pearson residuals (JPRs) χ_{Ji} for $i = 1, 2, ..., n$, and find the absolute values $|\chi_{Ji}|$.

Step 2: Divide $|\chi_{Ji}|$ into two groups, Group 1: $G_1 = \{|\chi_{Ji}|, \text{where } |\chi_{Ji}| < 1\}$, and Group 2: $G_2 = \{|\chi_{Ji}|, \text{where } |\chi_{Ji}| \geq 1\}$ [17].

Step 3: Sort all $|\chi_{Ji}|$ of group 1 (G_1) in ascending order, $G_1 = \left\{\left|\chi_{J(1)}\right|, \left|\chi_{J(2)}\right|, ..., \left|\chi_{J(n_1)}\right|\right\}$ where n_1 is the number of members in Group 1, then calculate the maximum difference between consecutive $\left|\chi_{J(i)}\right|$ in this sequence, defined as

$$\Delta_{\max} = \max_{i=1}^{n_1-1} \left(\left|\chi_{J(i+1)}\right| - \left|\chi_{J(i)}\right|\right).$$

Step 4: Calculate the cut-off point (C) defined as:

$$C = \Delta_{\max} + \left|\chi_{J(n_1)}\right|,$$

which identifies a clear breakpoint between normal and extreme data points.

Step 5: Move the $|\chi_{Ji}|$ in Group 2 to Group 1 when the $|\chi_{Ji}|$ are less than or equal to the cut-off point C.

Step 6: Repeat Steps 3–5 until $|\chi_{Ji}|$ in Group 2 can no longer be moved into Group 1; the remaining members of Group 2 are considered outliers.

2.6 Method 6 (GAAPR): Grouping Absolute Adjusted Pearson Residuals

This method groups the absolute values of adjusted Pearson residuals into quantile-based intervals. It examines the group-wise frequencies and statistics to identify abnormal concentrations of large residuals, indicative of outlying observations.

Step 1: Fit the gamma regression model and calculate the adjusted Pearson residuals (APRs) χ_i^A for $i = 1, 2, ..., n$, and find the absolute values $\left|\chi_i^A\right|$.

Step 2: Divide $\left|\chi_i^A\right|$ into two groups, Group 1: $G_1 = \left\{\left|\chi_i^A\right|, \text{where } \left|\chi_i^A\right| < 1\right\}$, and Group 2: $G_2 = \left\{\left|\chi_i^A\right|, \text{where } \left|\chi_i^A\right| \geq 1\right\}$ [17].

Step 3: Sort all $\left|\chi_i^A\right|$ of Group 1 (G_1) in ascending order, $G_1 = \left\{\left|\chi^A_{(??)}\right|, \left|\chi^A_{(??)}\right|, ..., \left|\chi^A_{(n_1)}\right|\right\}$ where n_1 is the number of members in Group 1, then calculate the maximum difference between consecutive $\left|\chi_i^A\right|$ in this sequence,

defined as:

$$\Delta_{\max} = \max_{i=1}^{n_1-1} \left(\left| \chi^A_{(i+1)} \right| - \left| \chi^A_{(i)} \right| \right)$$

Step 4: Calculate the cut-off point (C) defined as:

$$C = \Delta_{\max} + \left| \chi^A_{(n_1)} \right|,$$

which identifies a clear breakpoint between normal and extreme data points.

Step 5: Move the $\left|\chi^A_i\right|$ in Group 2 to Group 1 when the $\left|\chi^A_i\right|$ are less than or equal to the cut-off point C.

Step 6: Repeat Steps 3–5 until $\left|\chi^A_i\right|$ in Group 2 can no longer be moved into Group 1; the remaining members of Group 2 are considered outliers.

3. Comparison of outlier detection methods

In this study, the performance of the developed outlier detection methods is compared with existing outlier detection methods using two types of data (simulated data and real data).

3.1. Simulated data

We consider both uncontaminated and contaminated data. The variables, parameters, values, and performance criteria are specified as follows:

3.1.1 Uncontaminated data

1) Four independent variables(X), defined as:

$x_{i1} \sim Lognormal\left(4.5399 \times 10^{-5}, 1.73928 \times 10^{18}\right)$, $x_{i2} \sim N\left(25, 49\right)$, $x_{i3} \sim Exp\left(9\right)$, and $x_{i4} \sim N\left(40, 100\right)$

2) Parameter $\boldsymbol{\beta}$ when $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$ [3] that is $\beta_0 = 0.9317$, $\beta_1 = 0.1234$, $\beta_2 = 0.1850$, $\beta_3 = 0.0622$, $\beta_4 = 0.2798$.

3) A response variable (Y) has a gamma distribution using the inverse link function, where the parameters are μ_i and ϕ with $\mu_i = \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\right)^{-1}$ and dispersion $\phi = 0.33$, 0.67, and 2.

4) Sample size $n = 20, 40, 60, 80$, and 100.

5) The replication of the simulation to 1,000 iterations.

6) The criterion used to measure the effectiveness of the outlier detection method is pswamp, which is the probability that inliers are detected as outliers.

$$\text{pswamp} = \frac{\text{"false"}}{(n - out)},$$

where false is the number of inliers in the dataset that were detected as outliers, n is the total number of observations. The pswamp value ranges from 0 to 1, with a value equal to 0 indicating that no inliers were detected as outliers [19, 13, 7, 14].

3.1.2 Contaminated data

Pearson residuals are used as a practical device to identify observations that are extreme relative to the fitted model, thereby generating a controlled set of influential outliers for comparison purposes.

1) Four independent variables(X) defined as:

$x_{i1} \sim Lognormal\left(4.5399 \times 10^{-5}, 1.73928 \times 10^{18}\right)$, $x_{i2} \sim N\left(25, 49\right)$, $x_{i3} \sim Exp\left(9\right)$, and $x_{i4} \sim N\left(40, 100\right)$

2) Parameter $\boldsymbol{\beta}$ when $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$ [3] that is $\beta_0 = 0.9317$, $\beta_1 = 0.1234$, $\beta_2 = 0.1850$, $\beta_3 = 0.0622$, $\beta_4 = 0.2798$.

3) A response variable (Y) has a gamma distribution using the inverse link function, where the parameters are μ_i and ϕ with $\mu_i = \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\right)^{-1}$ and dispersion $\phi = 0.33$, 0.67, and 2.

4) Sample size $n = 20, 40, 60, 80$, and 100.

5) The replication of the simulation to 1,000 iterations.

6) Contaminated data in the response variable at 5% and 10% of the sample size (n).

6.1) Contaminated data at 5% of the sample size (n)

6.1.1) For $n = 20$, the observation with the maximum Pearson residuals will be contaminated by adding a_0 to that observation where $a_0 = \bar{y} + 3\,(V\,(y))$.

6.1.2) For $n = 40$, the two observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.1.3) For $n = 60$, the three observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.1.4) For $n = 80$, the four observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.1.5) For $n = 100$, the five observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.2 Contaminated data at 10% of sample size (n)

6.2.1) For $n = 20$, the two observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.2.2) For $n = 40$, the four observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.2.3) For $n = 60$, the six observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.2.4) For $n = 80$, the eight observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

6.2.5) For $n = 100$, the ten observations with the highest Pearson residuals will be contaminated by adding a_0 to those observations.

7) The criteria used to measure the performance of the outlier detection method are pout, pmask, and pswamp [19, 13, 7, 14],

$$\text{pout} = \frac{\text{"success"}}{\text{out}},$$

where success is the number of observations that successfully identified all outliers and out is the total number of outliers. The closer the pout value is to 1, the more effectively the proposed outlier detection identifies all the outliers.

$$\text{pmask} = \frac{\text{"failure"}}{\text{out}},$$

where failure is the number of outliers in the dataset that are detected as inliers. The pmask value ranges from 0 to 1, with a value equal to 0 indicating that no outliers were detected as inliers.

$$\text{pswamp} = \frac{\text{"false"}}{(n - out)},$$

where false is the number of inliers in the dataset that were detected as outliers, n is the total number of observations. The pswamp value ranges from 0 to 1, with a value equal to 0 indicating that no inliers were detected as outliers.

3.2. Real data

We compare the performance of the proposed outlier detection methods (TSPR, TJPR, TAPR, GASPR, GAJPR, and GAAPR) with existing methods: standardized Pearson residuals (SPRs), Jackknife Pearson residuals (JPRs), adjusted Pearson residuals (APRs), the Z-method (Z), the modified Z-method (Z*), and Grubb's test (G) using a dataset of 462 patients. The response variable is low-density lipoprotein (LDL); the four predictors are (i) tobacco use (amount per year), (ii) adiposity (body fat), (iii) alcohol consumption, and (iv) age. of the 462 observations, 160 patients with coronary heart disease are labeled as outliers [5].

4. Results on the simulated data

We propose six outlier detection methods (TSPR, TJPR, TAPR, GASPR, GAJPR, and GAAPR) based on Pearson residuals. These methods are compared with existing methods (SPRs, JPRs, APRs, Z, Z*, G) using simulated data under two scenarios: uncontaminated data and contaminated data, as follows:

4.1. Uncontaminated data

For uncontaminated data with sample sizes $n = 20, 40, 60, 80,$ and 100 and dispersion $\phi = 0.33, 0.67,$ and 2, the pswamp values are shown in Table 1.

Table 1. pswamp by sample size and dispersion (uncontaminated)

Method	$\phi = 0.33$					$\phi = 0.67$					$\phi = 2$				
	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
SPRs	0.002	0.007	0.008	0.009	0.011	0.001	0.009	0.012	0.013	0.014	0.001	0.010	0.016	0.019	0.021
JPRs	0.092	0.081	0.079	0.076	0.075	0.084	0.077	0.076	0.075	0.074	0.075	0.083	0.083	0.081	0.079
APRs	0.011	0.006	0.005	0.005	0.006	0.027	0.028	0.028	0.029	0.029	0.080	0.098	0.097	0.097	0.094
Z	0.006	0.005	0.005	0.004	0.005	0.007	0.011	0.012	0.013	0.013	0.011	0.027	0.041	0.051	0.054
Z*	0.035	0.033	0.033	0.031	0.033	0.047	0.067	0.070	0.071	0.071	0.079	0.180	0.218	0.230	0.236
G	0.037	0.042	0.042	0.041	0.042	0.044	0.050	0.050	0.050	0.050	0.049	0.057	0.058	0.057	0.056
TSPR	0.023	0.024	0.024	0.023	0.026	0.019	0.029	0.032	0.034	0.036	0.011	0.033	0.048	0.060	0.064
TJPR	0.037	0.029	0.027	0.026	0.028	0.031	0.035	0.036	0.037	0.038	0.021	0.040	0.052	0.062	0.066
TAPR	0.030	0.025	0.024	0.023	0.026	0.030	0.031	0.033	0.035	0.036	0.033	0.043	0.055	0.063	0.067
GASPR	0.195	0.172	0.168	0.174	0.172	0.183	0.148	0.140	0.140	0.136	0.165	0.133	0.139	0.134	0.130
GAJPR	0.213	0.179	0.173	0.177	0.175	0.204	0.153	0.144	0.142	0.137	0.186	0.135	0.140	0.135	0.130
GAAPR	0.072	0.058	0.056	0.055	0.055	0.114	0.106	0.106	0.106	0.107	0.200	0.172	0.169	0.165	0.162

From Table 1, when $\phi = 0.33$, Z at $n = 40, 60, 80,$ and 100 has the lowest pswamp, indicating it is most effective at educing the misidentification of inliers as outliers. At $n = 20$, SPRs outperforms Z (it has a lower pswamp), while APRs performs worse than Z. JPRs, Z*, G, TSPR, TJPR, TAPR, and GAAPR show moderate pswamp values, indicating higher misidentification rates than Z, SPRs, and APRs. In contrast, GASPR and GAJPR exhibit the highest pswamp values, indicating a greater misidentification rate.

When $\phi = 0.67$ and $\phi = 2$, SPRs has the lowest pswamp, showing superior ability to reduce misidentification of inliers as outliers. For $\phi = 0.67$ SPRs and Z have identical pswamp values at $n = 60$ and 80, but at $n = 100$, Z has a lower pswamp than SPRs. JPRs, APRs, Z*, G, TSPR, TJPR, and TAPR maintain moderate pswamp values, indicating higher misidentification rates than SPRs and Z. GASPR, GAJPR, and GAAPR show the highest pswamp values among the methods, indicating greater misidentification. Thus, SPRs is the most reliable and effective method for minimizing misidentification of inliers as outliers, particularly under higher dispersion ($\phi = 0.67$ and 2).

Among the developed methods, TSPR, TJPR, and TAPR perform well (moderate pswamp), whereas GASPR, GAJPR, and GAAPR demonstrate lower efficacy.

4.2. Contaminated data

For the 5% contaminated case, we simulate datasets with sample sizes n = 20, 40, 60, 80, and 100, and dispersion $\phi = 0.33, 0.67,$ and 2. The pout, pmask, and pswamp are reported in Tables 2-4, respectively.

Table 2 ($\phi = 0.33$) GAJPR attains the highest pout, indicating the strongest ability to detect outliers; at $n = 100$, GASPR matches GAJPR in pout. JPRs and GAAPR follow with lower pout, while Z*, G, TSPR, TJPR, and TAPR show moderate pout. SPRs, APRs, and Z have the lowest pout, indicating weaker detection. GAJPR also records the lowest pmask (best at avoiding masking); at $n = 80$ and 100, GASPR attains pmask equal to GAJPR's. Z*, G, TSPR, TJPR, and TAPR are moderate on pmask, whereas SPRs, APRs, and Z show the highest pmask values. On pswamp, Z shows the lowest values (near 0), demonstrating the best control of false positives; at $n = 20$, SPRs has an even lower pswamp than Z. JPRs, Z*, G, TSPR, TJPR, TAPR, and GAAPR are moderate on pswamp,

Table 2. pout, pmask, and pswamp for 5% contaminated data ($\phi = 0.33$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.106	0.894	0.000	0.352	0.648	0.001	0.399	0.601	0.001	0.421	0.579	0.002	0.434	0.566	0.001
JPRs	0.668	0.332	0.054	0.797	0.204	0.030	0.823	0.177	0.024	0.840	0.160	0.022	0.835	0.165	0.021
APRs	0.159	0.841	0.009	0.294	0.707	0.002	0.351	0.649	0.002	0.368	0.633	0.001	0.382	0.618	0.001
Z	0.081	0.919	0.001	0.081	0.920	0.001	0.088	0.912	0.000	0.085	0.916	0.000	0.081	0.919	0.000
Z*	0.316	0.684	0.019	0.436	0.565	0.011	0.482	0.518	0.010	0.489	0.511	0.007	0.518	0.482	0.007
G	0.405	0.595	0.016	0.585	0.416	0.012	0.634	0.366	0.011	0.652	0.348	0.010	0.662	0.338	0.009
TSPR	0.409	0.591	0.014	0.643	0.358	0.011	0.727	0.273	0.011	0.748	0.252	0.010	0.762	0.238	0.009
TJPR	0.532	0.468	0.022	0.691	0.310	0.013	0.747	0.253	0.012	0.768	0.233	0.011	0.774	0.226	0.010
TAPR	0.339	0.661	0.025	0.611	0.390	0.013	0.722	0.278	0.012	0.747	0.253	0.010	0.766	0.234	0.010
GASPR	0.691	0.309	0.148	0.815	0.186	0.111	0.842	0.158	0.104	0.856	0.144	0.097	0.854	0.146	0.099
GAJPR	0.706	0.294	0.165	0.818	0.182	0.116	0.843	0.157	0.106	0.857	0.144	0.099	0.854	0.146	0.100
GAAPR	0.591	0.409	0.059	0.782	0.218	0.031	0.827	0.173	0.026	0.844	0.156	0.024	0.839	0.161	0.022

Table 3. pout, pmask, and pswamp for 5% contaminated data ($\phi = 0.67$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.050	0.950	0.000	0.303	0.697	0.002	0.360	0.640	0.002	0.402	0.598	0.003	0.389	0.611	0.002
JPRs	0.615	0.385	0.056	0.847	0.154	0.031	0.884	0.116	0.027	0.905	0.095	0.025	0.909	0.091	0.023
APRs	0.296	0.704	0.021	0.617	0.384	0.013	0.698	0.302	0.010	0.754	0.246	0.010	0.741	0.259	0.009
Z	0.114	0.886	0.003	0.170	0.830	0.001	0.199	0.801	0.001	0.230	0.771	0.001	0.242	0.758	0.000
Z*	0.349	0.651	0.029	0.692	0.309	0.030	0.770	0.230	0.031	0.839	0.161	0.031	0.860	0.140	0.028
G	0.437	0.563	0.021	0.718	0.282	0.014	0.771	0.229	0.012	0.799	0.202	0.011	0.819	0.181	0.009
TSPR	0.319	0.681	0.011	0.631	0.369	0.012	0.733	0.267	0.013	0.798	0.202	0.014	0.805	0.195	0.013
TJPR	0.445	0.555	0.019	0.694	0.307	0.015	0.766	0.234	0.014	0.818	0.182	0.015	0.820	0.180	0.014
TAPR	0.324	0.676	0.024	0.640	0.361	0.017	0.741	0.259	0.015	0.806	0.194	0.015	0.809	0.191	0.014
GASPR	0.638	0.362	0.156	0.871	0.129	0.093	0.907	0.093	0.086	0.931	0.070	0.074	0.935	0.065	0.073
GAJPR	0.649	0.351	0.172	0.873	0.127	0.098	0.907	0.093	0.088	0.931	0.069	0.075	0.935	0.065	0.074
GAAPR	0.595	0.405	0.097	0.875	0.125	0.072	0.910	0.090	0.063	0.939	0.062	0.060	0.942	0.058	0.059

Table 4. pout, pmask, and pswamp for 5% contaminated data ($\phi = 2$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.017	0.983	0.000	0.234	0.767	0.002	0.364	0.636	0.003	0.413	0.587	0.004	0.432	0.568	0.005
JPRs	0.520	0.480	0.054	0.901	0.099	0.039	0.951	0.049	0.032	0.957	0.043	0.030	0.960	0.040	0.027
APRs	0.452	0.548	0.059	0.892	0.109	0.063	0.966	0.034	0.053	0.982	0.018	0.051	0.988	0.012	0.046
Z	0.118	0.882	0.004	0.400	0.601	0.005	0.604	0.396	0.007	0.748	0.252	0.012	0.799	0.201	0.013
Z*	0.438	0.562	0.062	0.929	0.071	0.138	0.985	0.015	0.172	0.992	0.009	0.191	0.994	0.006	0.197
G	0.456	0.544	0.025	0.843	0.158	0.016	0.911	0.089	0.013	0.916	0.084	0.013	0.922	0.078	0.010
TSPR	0.137	0.863	0.007	0.576	0.424	0.015	0.798	0.202	0.020	0.886	0.115	0.026	0.923	0.077	0.027
TJPR	0.254	0.746	0.014	0.640	0.361	0.018	0.821	0.179	0.022	0.897	0.103	0.027	0.929	0.071	0.028
TAPR	0.252	0.748	0.025	0.580	0.420	0.025	0.792	0.208	0.027	0.885	0.115	0.030	0.932	0.068	0.031
GASPR	0.544	0.456	0.145	0.938	0.063	0.086	0.984	0.016	0.082	0.988	0.013	0.076	0.991	0.009	0.070
GAJPR	0.570	0.430	0.167	0.940	0.060	0.087	0.984	0.016	0.083	0.988	0.013	0.077	0.991	0.009	0.070
GAAPR	0.587	0.413	0.169	0.952	0.049	0.133	0.987	0.013	0.119	0.994	0.006	0.117	0.996	0.004	0.112

while GASPR and GAJPR exhibit the highest pswamp values (greater misidentification of inliers as outliers). Considering all three criteria, GAJPR performs best overall in Table 2 (high pout and low pmask), despite relatively high pswamp; JPRs, GASPR, and GAAPR follow, and SPRs, APRs, and Z are least reliable.

Table 3 ($\phi = 0.67$) GAAPR generally achieves the highest pout (strongest detection); at $n = 20$, GAJPR slightly exceeds GAAPR. JPRs, GASPR, and GAJPR follow with lower pout; Z*, G, TSPR, TJPR, and TAPR are moderate, and SPRs, APRs, and Z have the lowest pout. GAAPR records the lowest pmask (best at preventing masking); JPRs, GASPR, and GAJPR have higher pmask values, though GASPR matches GAAPR at $n = 60$ and 100. Z*, G, TSPR, TJPR, and TAPR are moderate on pmask, while SPRs, APRs, and Z show the highest pmask. On pswamp, Z again shows the lowest values (near 0); at $n = 20$, SPRs has a lower pswamp than Z. APRs, Z*, G, TSPR, TJPR, and TAPR are moderate on pswamp, whereas GASPR, GAJPR, and GAAPR show higher pswamp values. On the

whole, GAAPR leads Table 3 on the combined criteria (high pout, low pmask) despite elevated pswamp; JPRs, GASPR, and GAJPR follow, and SPRs, APRs, and Z are less reliable.

Table 4 ($\phi = 2$) GAAPR consistently shows the highest pout across sample sizes, confirming its strong outlier detection ability. At small n (e.g., $n = 20$), JPRs, GASPR, and GAJPR follow GAAPR. APRs, Z*, G, TSPR, TJPR, and TAPR are moderate, while SPRs and Z exhibit the weakest detection. GAAPR also attains the lowest pmask at all n (best masking control), followed by JPRs, GASPR, and GAJPR; SPRs and Z show the highest pmask values. On pswamp, SPRs shows the lowest values (near 0), followed by Z; JPRs, APRs, G, TSPR, TJPR, and TAPR are moderate, while Z*, GASPR, GAJPR, and GAAPR show the highest pswamp values. In Table 4, GAAPR performs best overall (high pout, low pmask) despite higher pswamp; JPRs, APRs, Z*, GASPR, and GAJPR follow, and SPRs and Z are least effective for detection.

Across Tables 2-4 (5% contamination), GAJPR and GAAPR consistently rank among the top methods by combining high pout and low pmask, i.e., they detect true outliers well and minimize masking, though both tend to have relatively high pswamp (more false positives). JPRs and GASPR are strong alternatives with good detection and moderate pswamp. Z*, G, TSPR, TJPR, and TAPR deliver balanced, moderate performance with low to moderate pswamp. SPRs, APRs, and Z are the least effective overall because they detect fewer true outliers and exhibit higher masking, despite showing low pswamp in many settings.

For the 10% contaminated case, we simulate datasets with sample sizes $n = 20$, 40, 60, 80, and 100, and dispersion $\phi = 0.33$, 0.67, and 2. The pout, pmask, and pswamp are reported in Tables 5-7, respectively.

Table 5. pout, pmask, and pswamp for 10% contaminated data ($\phi = 0.33$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.057	0.943	0.000	0.168	0.833	0.001	0.188	0.812	0.001	0.192	0.809	0.002	0.193	0.807	0.001
JPRs	0.475	0.525	0.045	0.605	0.395	0.022	0.626	0.374	0.017	0.645	0.355	0.016	0.638	0.362	0.016
APRs	0.072	0.928	0.007	0.162	0.838	0.002	0.170	0.830	0.001	0.186	0.814	0.002	0.178	0.822	0.001
Z	0.034	0.966	0.001	0.043	0.958	0.000	0.042	0.958	0.000	0.042	0.958	0.000	0.044	0.956	0.000
Z*	0.177	0.824	0.015	0.267	0.733	0.006	0.290	0.710	0.005	0.306	0.694	0.004	0.288	0.712	0.003
G	0.250	0.751	0.011	0.351	0.650	0.006	0.379	0.621	0.005	0.384	0.616	0.005	0.379	0.621	0.004
TSPR	0.288	0.712	0.014	0.502	0.498	0.012	0.550	0.450	0.011	0.574	0.426	0.011	0.577	0.423	0.011
TJPR	0.369	0.631	0.022	0.531	0.469	0.014	0.568	0.433	0.012	0.586	0.414	0.012	0.587	0.413	0.011
TAPR	0.227	0.773	0.023	0.464	0.536	0.014	0.538	0.463	0.012	0.564	0.436	0.011	0.572	0.428	0.011
GASPR	0.517	0.483	0.138	0.644	0.357	0.092	0.668	0.332	0.086	0.690	0.310	0.077	0.686	0.314	0.080
GAJPR	0.540	0.460	0.152	0.645	0.355	0.096	0.669	0.331	0.088	0.691	0.309	0.079	0.687	0.314	0.082
GAAPR	0.384	0.617	0.044	0.589	0.411	0.024	0.619	0.381	0.019	0.648	0.353	0.018	0.640	0.360	0.017

Table 6. pout, pmask, and pswamp for 10% contaminated data ($\phi = 0.67$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.040	0.960	0.001	0.140	0.861	0.002	0.163	0.838	0.002	0.175	0.825	0.003	0.178	0.822	0.003
JPRs	0.448	0.553	0.047	0.629	0.371	0.022	0.659	0.341	0.019	0.678	0.322	0.018	0.692	0.308	0.017
APRs	0.164	0.836	0.014	0.366	0.635	0.009	0.412	0.588	0.008	0.439	0.561	0.008	0.450	0.550	0.008
Z	0.049	0.952	0.001	0.083	0.917	0.000	0.115	0.885	0.000	0.116	0.884	0.000	0.121	0.879	0.000
Z*	0.271	0.729	0.021	0.467	0.534	0.017	0.540	0.460	0.016	0.571	0.429	0.013	0.587	0.413	0.010
G	0.307	0.694	0.012	0.450	0.550	0.006	0.472	0.528	0.004	0.477	0.523	0.003	0.475	0.525	0.002
TSPR	0.219	0.781	0.012	0.480	0.521	0.013	0.547	0.454	0.013	0.582	0.418	0.013	0.606	0.394	0.013
TJPR	0.303	0.697	0.020	0.518	0.483	0.014	0.569	0.431	0.014	0.599	0.402	0.014	0.620	0.380	0.014
TAPR	0.207	0.793	0.021	0.460	0.540	0.016	0.536	0.465	0.014	0.581	0.419	0.014	0.605	0.395	0.014
GASPR	0.494	0.506	0.144	0.692	0.308	0.078	0.736	0.264	0.062	0.756	0.244	0.056	0.772	0.228	0.052
GAJPR	0.510	0.491	0.160	0.694	0.306	0.082	0.736	0.264	0.063	0.756	0.244	0.057	0.772	0.228	0.053
GAAPR	0.436	0.564	0.079	0.690	0.310	0.054	0.742	0.258	0.046	0.761	0.239	0.041	0.781	0.219	0.038

Table 5 ($\phi = 0.33$) GAJPR attains the highest pout, indicating the strongest outlier detection; JPRs, GASPR, and GAAPR follow with somewhat lower pout. Z*, G, TSPR, TJPR, and TAPR show moderate pout, while SPRs, APRs, and Z have the lowest pout (weaker detection). GAJPR also records the lowest pmask (best at avoiding masking); at $n = 100$ GASPR matches GAJPR's pmask. Z*, G, TSPR, TJPR, and TAPR are moderate on pmask,

Table 7. pout, pmask, and pswamp for 10% contaminated data ($\phi = 2$)

Method	$n = 20$			$n = 40$			$n = 60$			$n = 80$			$n = 100$		
	pout	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*	*pout*	*pmask*	*pswamp*
SPRs	0.016	0.984	0.001	0.113	0.887	0.002	0.161	0.839	0.003	0.180	0.820	0.004	0.188	0.813	0.004
JPRs	0.373	0.627	0.049	0.679	0.321	0.025	0.733	0.267	0.021	0.734	0.266	0.020	0.735	0.265	0.019
APRs	0.296	0.704	0.042	0.691	0.310	0.045	0.792	0.209	0.039	0.831	0.170	0.034	0.843	0.157	0.029
Z	0.101	0.900	0.002	0.217	0.784	0.001	0.350	0.650	0.001	0.462	0.538	0.002	0.521	0.479	0.001
Z*	0.349	0.651	0.048	0.780	0.220	0.108	0.889	0.111	0.141	0.912	0.088	0.156	0.923	0.077	0.161
G	0.349	0.652	0.012	0.520	0.481	0.003	0.558	0.442	0.001	0.564	0.437	0.001	0.567	0.433	0.000
TSPR	0.115	0.885	0.009	0.415	0.585	0.012	0.592	0.408	0.017	0.709	0.292	0.021	0.748	0.253	0.021
TJPR	0.181	0.820	0.015	0.473	0.527	0.014	0.622	0.379	0.018	0.721	0.279	0.021	0.760	0.240	0.022
TAPR	0.195	0.805	0.022	0.436	0.565	0.020	0.601	0.400	0.023	0.715	0.285	0.025	0.754	0.246	0.024
GASPR	0.420	0.580	0.147	0.791	0.209	0.058	0.881	0.119	0.051	0.901	0.099	0.044	0.900	0.101	0.040
GAJPR	0.449	0.552	0.168	0.793	0.207	0.060	0.881	0.119	0.051	0.902	0.099	0.044	0.900	0.100	0.040
GAAPR	0.480	0.520	0.149	0.814	0.186	0.115	0.901	0.099	0.093	0.919	0.081	0.081	0.929	0.071	0.075

whereas SPRs, APRs, and Z show the highest pmask (more masking). On pswamp, Z shows the lowest values (best control of false positives); at $n = 20$ SPRs is even lower. JPRs, TSPR, TJPR, TAPR, and GAAPR register moderate pswamp, while GASPR and GAJPR exhibit the highest pswamp (greater misidentification of inliers as outliers). Considering all three measures, GAJPR performs best overall in Table 5 (high pout, low pmask) despite relatively high pswamp; JPRs, GASPR, and GAAPR follow, and SPRs, APRs, Z are least reliable.

Table 6 ($\phi = 0.67$) GAAPR generally achieves the highest pout (strong detection), although GAJPR exceeds GAAPR at $n = 20$ and 60. JPRs, GASPR, and GAJPR follow, while APRs, Z*, G, TSPR, TJPR, and TAPR are moderate, and SPRs and Z have the lowest pout. GAAPR records the lowest pmask (best masking control); JPRs, GASPR, and GAJPR have higher pmask, and at several sample sizes GASPR matches GAJPR's pmask. Z*, G, TSPR, TJPR, and TAPR remain moderate on pmask, while SPRs and Z show the highest pmask. For pswamp, Z has the lowest values (near 0); at $n = 20$ SPRs matches Z. SPRs and G then show somewhat higher pswamp, many methods (JPRs, APRs, Z*, TSPR, TJPR, TAPR, GAAPR) are moderate, and GASPR and GAJPR exhibit the highest pswamp. Overall, GAAPR leads Table 6 on combined criteria (high pout, low pmask) despite elevated pswamp; JPRs, GASPR, and GAJPR follow, and SPRs, APRs, Z are less reliable.

Table 7 ($\phi = 2$) GAAPR consistently shows the highest pout across sample sizes, confirming the strongest outlier-detection ability; at $n = 20$ GASPR and GAJPR follow closely. JPRs, APRs, Z, Z*, G, TSPR, TJPR, and TAPR display moderate pout, while SPRs has the lowest pout. GAAPR also attains the lowest pmask at all n (best masking control); GASPR and GAJPR typically follow, and SPRs and Z show the highest pmask (most masking). On pswamp, Z yields the lowest values (best false-positive control), with SPRs slightly higher; JPRs, APRs, Z*, G, TSPR, TJPR, and TAPR are moderate, while GASPR, GAJPR, and GAAPR show the highest pswamp. In Table 7, GAAPR performs best overall (high pout, low pmask) despite higher pswamp; JPRs, APRs, Z*, GASPR, and GAJPR follow, and SPRs/Z are least effective.

Therefore, Tables 5-7 (10% contamination), GAAPR and GAJPR consistently rank among the top methods by combining high pout and low pmask, i.e., they detect true outliers well and minimize masking though both tend to have relatively high pswamp (more false positives). JPRs and GASPR are strong alternatives with good detection and moderate pswamp. Z*, G, TSPR, TJPR, and TAPR deliver balanced, moderate performance with relatively low pswamp. SPRs, APRs, and Z are the least effective overall because they detect fewer true outliers and exhibit greater masking, despite often having low pswamp.

5. Results on the real data

The real-data analysis uses a coronary heart disease dataset containing patients with and without CHD. The response variable is low-density lipoprotein (LDL), and the four independent variables are (i) the amount of tobacco used per year, (ii) adiposity, (iii) alcohol consumption, and (iv) age. Although the dataset includes 160 patients with coronary heart disease [5], CHD status is not treated as an a priori outlier label. Instead, the proposed methods are

applied to the full dataset to identify influential observations and assess their performance in a real-data setting. The corresponding values of pout, pmask, and pswamp are shown in Table 8.

Table 8. pout, pmask, and pswamp in coronary heart disease data

Method	$pout$	$pmask$	$pswamp$
SPRs	0.025	0.975	0.007
JPRs	0.088	0.913	0.056
APRs	0.000	1.000	0.003
Z	0.019	0.981	0.003
Z*	0.088	0.913	0.023
G	0.075	0.925	0.020
TSPR	0.050	0.950	0.023
TJPR	0.050	0.950	0.023
TAPR	0.038	0.963	0.026
GASPR	0.194	0.806	0.182
GAJPR	0.194	0.806	0.182
GAAPR	0.025	0.975	0.013

From Table 8, all methods have low pout values. GASPR and GAJPR show higher pout values than the other methods, indicating they identify more outliers. APRs fail to identify any outliers (pout = 0). In terms of pmask, all methods have values greater than 0.8; GASPR and GAJPR have the lowest pmask values, indicating they are better at avoiding misclassification of outliers as inliers. APRs exhibit complete masking of outliers (pmask = 1). Regarding pswamp, APRs and Z show the lowest values (mostly 0 or close to 0), demonstrating their ability to minimize misclassification of inliers as outliers. In contrast, GASPR and GAJPR have the highest pswamp values (up to 0.182), indicating greater misclassification of inliers as outliers. Considering all three performance criteria, GASPR and GAJPR identify more outliers than the other methods but also misidentify more inliers as outliers. In the CHD dataset, however, the flagged observations should be interpreted as model-based unusual cases rather than clinically abnormal patients. Since the proposed methods are designed to detect outliers under the gamma regression model, not to classify CHD status, it is expected that only a subset of CHD patients is flagged. Many CHD cases may still conform to the fitted covariate-response relationship and therefore are not detected as outliers.

6. Conclusion

This study develops six outlier detection methods for gamma regression (TSPR, TJPR, TAPR, GASPR, GAJPR, and GAAPR) using Pearson residuals, Tukey's boxplot, and group absolute residuals, and compares them to six existing methods (SPRs, JPRs, APRs, Z, Z*, and G). Performance is evaluated by simulation (uncontaminated data and contaminated data at 5% and 10%) and on a coronary heart disease dataset, using three metrics: pout (probability of detecting all true outliers), pmask (probability of masking true outliers, i.e., misclassifying outliers as inliers), and pswamp (probability of misclassifying inliers as outliers). Simulation results show that GAJPR and GAAPR are the most effective overall, achieving high pout and low pmask, though both exhibit relatively high pswamp. On the real dataset, GASPR and GAJPR perform best, with GAJPR emerging as the single most effective method across simulation and empirical analyses. Moreover, JPRs, TSPR, TJPR, TAPR, GASPR, GAJPR, and GAAPR are recommended when the priority is to avoid missing potential outliers; by contrast, SPRs, APRs, Z, Z*, and G are preferable when minimizing false detections (i.e., ensuring identified outliers are truly outliers). The proposed methods are useful as screening procedures for potential outliers, particularly when sensitivity is prioritized. However, their relatively higher false-positive rate (pswamp) suggests that flagged observations should be further examined before being treated as true outliers.

7. Discussion

Across our simulation scenarios and the real data application, the group absolute residuals-based hybrid methods that combine Pearson residuals scaling with Tukey's boxplot, especially GAJPR and GAAPR, consistently demonstrated the strongest detection performance. These methods achieved the highest pout (probability of detecting all true outliers) and the lowest pmask (probability of masking true outliers as inliers), meaning they are most effective at identifying contaminants that would otherwise bias gamma regression results. JPRs and GASPR provided competitive performance with more moderate false-positive rates, while SPRs, APRs, and the Z-family tended to be conservative: they produced low pswamp (few inliers misclassified as outliers) but often missed true outliers (low pout, high pmask). On the real coronary heart disease dataset, GASPR and GAJPR performed best empirically, with GAJPR emerging as the most reliable across both simulated and empirical evaluations.

The strong performance of the group absolute residuals-based is intuitive, which is read as grouped absolute residuals rather than a search algorithm. By grouping absolute Pearson-scaled residuals and applying Tukey's boxplot thresholding within or across groups, these procedures adapt thresholds to local residual structure and to the gamma regression's mean–variance relationship. Grouping allows the method to tailor sensitivity where residual dispersion varies with the fitted mean (or covariates), making it more likely to detect outliers that are large relative to their local dispersion or that interact with leverage. The principal trade-off is a tendency toward higher pswamp: because grouping and adaptive thresholds broaden the criteria for flagging observations to avoid masking, more inliers can be falsely flagged. The relative rankings are broadly stable across contamination rates (5% and 10%), dispersion levels, and sample sizes: the group absolute residuals-based methods remain top performers in most settings. Small samples variability occasionally changes fine-grained ordering between GAJPR and GAAPR, and larger samples generally stabilize rankings as residual estimates become more reliable. Importantly, the simulation settings we used (percentage and type of contamination, covariate distributions, link function) influence performance; alternative contamination mechanisms (e.g., clustered or covariate-dependent outliers) could alter method advantages, especially if grouping choices do not align well with the contamination structure.

For practitioners, these findings imply a clear, context-dependent workflow. If the analyst's priority is sensitivity, avoiding missed outliers that might bias inference, use GAJPR or GAAPR as primary detectors, but follow up flagged cases with substantive checks: verify data-entry and measurement provenance, examine leverage and influence diagnostics, and perform sensitivity analyses (re-fit models with and without flagged points). If specificity, minimizing false alarms, is paramount (for example, where follow-up costs are high), prefer conservative methods (SPRs, APRs, Z, Z*, and G).

In summary, grouped-absolute-residuals methods, particularly GAJPR and GAAPR, offer powerful, adaptive detection for gamma regression outliers by tailoring thresholds to local residual behavior, thereby maximizing detection and minimizing masking at the cost of increased false positives; consequently, choice of method should reflect the analyst's tolerance for missed outliers versus false alarms, and the group absolute residuals-based screening should be validated before acting on flagged observations.

Author Contributions

WS.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, validation, and writing– original draft, review, and editing. LP.: data curation, formal analysis, software, visualization.

Acknowledgment

This research was financially supported by the Research, Innovation, and Academic Services Fund, Faculty of Science, Khon Kaen University, fiscal year 2026.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. E. Altun, M. C. Korkmaz, M. EI-Morshedy, and M. S. Eliwa, *The extended gamma distribution with regression model and applications*, AIMS Mathematics, vol. 6, no. 3, pp. 2418–2439, 2020. https://doi.org/10.3934/math.2021147
2. M. Amin, S. Afzal, M. N. Akram, A. H. Muse, A. H. Tolba, and T. A. Abushal, *Outlier detection in gamma regression using Pearson residuals: simulation and an application*, AIMS Mathematics, vol. 7, no. 8, pp. 15331–15347, 2022. https://doi.org/10.3934/math.2022840
3. M. Amin, M. Amanullah, M. Aslam, and M. Qasim, *Influence diagnostics in gamma ridge regression model*, Journal of Statistical Computation and Simulation, vol. 89, no. 3, pp. 536–556, 2018. https://doi.org/10.1080/00949655.2018.1558226
4. C. O. Arimie, E. O. Biu, and M. A. Ijomah, *Outlier detection and effects on modeling*, Open Access Library Journal, vol. 7, no. 9, e6619, 2020. https://doi.org/10.4236/oalib.1106619
5. B. Basener, *Coronary heart disease*, Retrieved 5 February 2025, from https://www.kaggle.com/datasets/billbasener/coronary-heart-disease
6. M. C. Bossio, and E. C. Cuervo, *Gamma regression models with the gammareg R package*, Comunicaciones en Estadística, vol. 8, no. 2, pp. 211–223, 2015.
7. T. Chaitongdee, W. Srisodaphol, O. D. Rahmashari, B. Rattanawong, and K. Prakhammin, *Enhanced outlier detection in linear-circular regression using circular distance and mean resultant length*, Statistics, Optimization & Information Computing, vol. 11, no. 4, pp. 936–948, 2025. https://doi.org/10.19139/soic-2310-5070-1681
8. E. D. Dan, and O. A. Ijeoma, *Statistical analysis/methods of detecting outliers in a univariate data in a regression analysis model*, International Journal of Education and Research, vol. 1, no. 5, pp. 1–24, 2013.
9. R. N. Das, and J. Kim, *GLM and joint GLM techniques in hydrogeology: an illustration*, International Journal of Hydrology Science and Technology, vol. 2, no. 2, pp. 185–201, 2012. https://doi.org/10.1504/IJHST.2012.047408
10. J. W. Hardin, and J. M. Hilbe, *Generalized linear models and extensions (4th ed.)*, Stata Press Publication, 2018.
11. A. F. Lukman, I. Dawoud, B. M. G. Kibria, Z. Y. Algamal, and B. Aladeitan, *A new ridge-type estimator for the gamma regression model*, Scientifica, vol. 2021, no. 1, pp. 1–8, 2021. https://doi.org/10.1155/2021/5545356
12. Y. Murakami, T. Okamura, K. Nakamura, K. Miura, and 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, vol. 3, no. 3, e002234, 2013. https://doi.org/10.1136/bmjopen-2012-002234
13. S. S. S. A. Mutalib, S. Z. Satari, and W. N. S. W. Yusoff, *A new single linkage robust clustering outlier detection procedures for multivariate data*, Sains Malaysiana, vol. 52, no. 8, pp. 2431–2451, 2023. https://doi.org/10.17576/jsm-2023-5208-19
14. O. D. Rahmashari, and W. Srisodaphol, *Advanced outlier detection methods for enhancing beta regression robustness*, Decision Analytics Journal, vol. 14, Article 100557, 2025. https://doi.org/10.1016/j.dajour.2025.100557
15. J. S. Ruíz, O. A. M. López, G. H. Ramírez, and J. C. Hiriart, *Generalized linear mixed models with applications in agriculture and biology*, Cham: Springer Nature, 2023.
16. S. Santoyo, *A brief overview of outlier detection techniques*, Retrieved 10 October 2025, from http://engdashboard.blogspot.com/2017/09/a-brief-overview-of-outlier-detection.html.
17. E. Serdahl, *An introduction to graphical analysis of residual scores and outlier detection in bivariate least squares regression analysis*, The annual meeting of the Southwest Educational Research Association, New Orleans, 1996.
18. M. Stroganova, *How to Read and Make Box Plot: A Complete Guide + Best Practices*, Retrieved 20 October 2025, from https://mlsamurai.medium.com/how-to-read-and-make-box-plot-a-complete-guide-best-practices-92b233e59c3b.
19. N. S. Zulkipli, S. Z. Satari, and 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*, Pakistan journal of statistics and Operation research, vol. 18, no. 3, pp. 561–573, 2022. http://dx.doi.org/10.18187/pjsor.v18i3.3982