JIRSS (2025)
Vol. 24, No. 2, pp 69-90
DOI: 10.22034/jirss.2026.2076260.1154

Influence Diagnostics for the Gamma-Pareto Regression: A DFFITS-Based Comparison of Residuals

Nasir Saleem[1]**, Atif Akbar**[1]**, Asad Abbas**[1]

[1]Department of Statistics, Bahauddin Zakariya University (BZU), Multan, Pakistan, 60800.

Received: 30/10/2025, Accepted: 31/01/2026, Published online: 31/01/2026

Abstract. The generalized linear models (GLMs) use Gamma-Pareto regression Model (G-PRM) to address the sensitivity of influential observations. Difference of Fits (DFFITS) is a popular technique for identifying influential observations. We apply DFFITS to the G-PRM with various residuals. We present illustrative real and simulated data. One class of adjusted Pearson residuals is more effective in detecting influential observations, considering small or large dispersion parameters. We calculate detection percentages to evaluate the proposed procedure's performance, replicating the process 10,000 times.

Keywords. Difference of fits, Gamma-Pareto regression, Residuals, Inverse Link function,Influence diagnostics.

MSC: 62-XX, 62Rxx, 62R10.

1 Introduction

The Gamma-Pareto distribution (G-PD) is a continuous probability distribution with parameters α, β, and γ, often used to model skewed data. The G-PRM is a regression model that extends the G-PD to incorporate covariates, allowing for the analysis of the relationship between the response variable and predictors. The G-PRM is useful for modeling data with heavy tails and positive support, such as income, loss data, or

Nasir Saleem (nasirsaleem160@gmail.com).
Corresponding author: Atif Akbar (atifakbar@bzu.edu.pk).
Asad Abbas (asadabbas0786@yahoo.com).

survival times. It provides a flexible framework for analyzing and predicting outcomes in various fields, including economics, finance, and insurance.

GLMs extend linear regression to non-normal responses by specifying a link function that relates the mean to predictors. In the study by Saleem and Akbar (2024) and Nelder and Wedderburn (1972), GLMs combined some statistical models, including the linear regression model (LRM), the G-PRM, the Poisson regression model (PRM), the beta regression model (BRM), etc. Influential observations refer to observations that lack adjacent values and are significantly different to other observations. Simply put, influential observations are abnormal data values. Potential residual plot, DFFITS and index plot are the statistical tools used to identify influential observations (see Hadi (1992)). In most cases, researchers propose deleting these values, which, however, causes the loss of information (Khan and Akbar (2019)).

In finance, it is used to study the heavy tails of stock prices, returns, and volatility (Hanum (2015), Hanum et al. (2016)). Failure time modelling and reliability engineering are crucial for parts with long life cycles. In hydrology, models are used to understand and control water resources, including rainfall, water flow distributions, and extreme floods. The science of materials, which is used to model fatigue life and material strength, which often have heavy-tailed distributions. Environmental science models pollutant concentrations, pollutant emissions and patterns, and health-care costs, disease severity, and survival period, all of which are often skewed and heavy tailed. Monitored and controlled through quality control manufacturing processes of heavy-tailed distributions of flaws or failures are monitored and controlled in complex networks which often exhibit heavy-tailed behavior. The G-PRM is useful in creating better analysis, model, and decisions due to its capability to correctly represent the behavior of extreme events and heavy-tailed phenomena in these areas.

The results of regression analysis can be significantly influenced by one observation. It can also lead to false covariance matrix and wrong estimates of coefficients. These observations are to be found and removed in order to get the right estimates in the regression models. These are probit and logistic regressions that have a binary response. It is relevant provided the response variable is constant-variance and positively skewed with a mean proportional to the variance. Some of the common applications of this model are survival analysis and reliability (Lawless (2003)).

The GLM is in practice a violation of the non-influential observation assumption (see Hardin and Hilbe (2012)). Cook (1977) first introduced influence diagnostics of LRM. Belsley et al. (1980) explored these impact diagnostics in many different perspectives. As it has been discussed, and argued by Preisser and Qaqish (1996), Pregibon (1981), Vanegas et al. (2013), Williams (1987), the influence diagnostics in the GLM remains the focus area. The Pearson residuals are common in the analysis of influential observations in influence diagnostics. Williams (1987) also employed the use of deviance residuals to indicate influence diagnostics.

Residuals play a crucial role in model diagnostics and fitting evaluation. GLMs provide several residual structures, including Pearson, deviance, and likelihood residuals, whereas LRM utilize raw residuals for model diagnostics. Pearson and deviance residuals, as well as likelihood residuals, are the most commonly used residuals in GLMs for influence diagnostics. The normality assumption cannot be met by these residuals, and they possess different probability distributions. Adjusted residuals are essential in order to fulfill this normalcy assumption. Adjusted deviance residuals offered by Pierce and Schafer (1986) and adjusted Pearson residuals proposed by Cordeiro (2004) based on Cox and Snell (1968) are the two main theories of adjusted residuals, which are currently in use. The objective of these theories is attainment of normalcy. However, the same results are realized when analyzing adjusted Pearson residuals of the beta regression (Anholeto et al. (2014)) and the exponential family of nonlinear models (see Simas and Cordeiro (2009)). Regression diagnostics is an important element in the development of an LRM. Regression diagnostics is one of the ways of identifying the odd observations (Amin et al. (2016); Amin et al. (2017); Amin et al. (2019)). Other articles that can be found on the identification of influential observations in GLMs include Amin et al. (2016), Amin et al. (2020) and Amin et al. (2021) identifying influential observations in the gamma regression model (GRM).

Zheng et al. (2014) wrote about the Shifted Gamma-Generalized Pareto Distribution model of crash estimation and crashes mapping the safety continuum. The new Log-Gamma-Pareto Distribution is created by Ashour et al. (2014). Alzaatreh and Ghosh (2016) proposed a new Gamma-Pareto (IV) distribution and explained the application. The article by De Andrade et al. (2017) has touched upon the definition of the gamma generalized Pareto distribution and its applications in survival analysis. Alzaghal (2020) used the exponentiated gamma-Pareto on the susceptibility of bladder cancer. Dar et al. (2020) talked about the weighted gamma-pareto distribution and its usage. Cordeiro (2004) suggested adjusted Pearson residuals, according to Cox and Snell (1968). These theories are aimed at achieving normalcy. Simas and Cordeiro (2009) found that the same findings are achieved by studying the adjusted Pearson residuals (APR) of exponential family of nonlinear models. Several methods have also been suggested in literature to diagnose significant observations or points in the LM including Chatterjee and Hadi (1988), Cook and Weisberg (1982), Atkinson (1985), and Cook (1986). Lee (1986) on the other hand provided a method of evaluating partial influence in the GLM. One way of determining the impact on the GLM regression coefficients was suggested by Thomas and Cook (1989).

This study pioneers applying DFFITS to G-PRM, extending GLM influence diagnostics to heavy-tailed data. We systematically compare standardized and adjusted Pearson, deviance, and likelihood residuals within G-PRM, assessing effectiveness across dispersion and sample size. Adjusted Pearson residuals show superior performance. Bridging a gap, this is the first study to examine DFFITS residuals in G-PRM, offering practical guidance for robust modeling.

2 Methodology

2.1 The Gamma-Pareto regression model

Alzaatreh et al. (2012) state that the G-PD pdf is provided by:

$$f(y;\alpha,\beta,\gamma) = \frac{\gamma}{\beta^{\alpha}\Gamma(\alpha)}\left(\log\left(\frac{y}{\gamma}\right)\right)^{\alpha-1}\left(\frac{y}{\gamma}\right)^{-\left(\frac{1}{\beta}+1\right)} \tag{2.1}$$

with $\alpha, \beta, \gamma > 0$ and $y \geq \gamma$.

The mean and variance of G-P distribution are, $E(a(y)) = \alpha\beta$, $V(a(y)) = \alpha\beta^2$ respectively and $\alpha(y)$ is a function of $f(y)$. According to Hanum et al. (2016), with parameters, Eq. (2.1) can be modified $\alpha = \frac{1}{\phi}$ and $\beta = \mu\phi$. The Gamma Pareto density for y under these conditions is given by,

$$f(y;\mu,\phi) = \frac{\gamma}{(\mu\phi)^{\frac{1}{\phi}}\Gamma\left(\frac{1}{\phi}\right)}\left(\log\left(\frac{y}{\gamma}\right)\right)^{\frac{1}{\phi}-1}\left(\frac{y}{\gamma}\right)^{-\left(\frac{1}{\mu\phi}+1\right)} \tag{2.2}$$

with $y \geq 0$, $\mu > 0$ and $\phi > 0$. Since the mean and variance of y are $E(y) = \mu$ and $V(y) = \phi V(\mu) = \phi\mu^2$.

The notation and parameterization for the G-PD and G-PRM. The G-PD has parameters α, β, and γ. For the G-PRM, we consider the following parameterization such as model parameter vector μ and ϕ where μ is the mean parameter (location), ϕ is the dispersion parameter (shape) and γ is the Pareto tail index (shape). The link function relates the mean μ to the linear predictor η and $g(\mu_i) = x_i^T\beta = \eta$. Where x is the covariate vector, and β is the regression coefficient vector. The mapping between parameterizations and the original parameters α, β, and γ are related to the re-parameterized version as follows $\alpha = \frac{1}{\phi}$ and $\beta = \mu\phi$.

For the ith observation, let $x_{i1}, x_{i2}, \ldots, x_{ip}$ represent the p non-stochastic regressors. Following that, the G-PRM for the response variable y mean is provided by

According to Hanum et al. (2016), Link function g in GLM is $g(\mu_i) = x_i^T\beta = \eta_i$ where $\mu_i = E(a(y_i))$, which are presented in Table 1.

Table 1: Different link function of G-PD

Link function	Form of link function	Reference
Inverse link function	$\mu = \frac{1}{X'\beta}$	
Identity link function	$\mu = X'\beta$	Hanum et al. (2016)
Log link function	$\mu = \log(X'\beta)$	
	$\mu = e^{X'\beta}$	

2.2 Parameters Estimation of the GLM Gamma-Pareto Regression Model

Finding the likelihood function's derivative with respect to β_j is the first step in estimating the parameter β_j using maximum likelihood and τ_i is the function of $f(y)$. By Eq. (2.2)

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \tau_i}\frac{\partial \tau_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right] \tag{2.3}$$

Now

$$\frac{\partial l_i}{\partial \tau_i} = a(y)b'(\tau) + c'(\tau) = \beta^{-2}\left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right)$$

$$\frac{\partial \tau_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \tau_i}} = \frac{1}{\frac{\partial \alpha\beta}{\partial \beta}} = \frac{1}{\alpha}$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}x_{ij}$$

Where $\frac{\partial \mu_i}{\partial \eta_i}$ is based on the GLM's link function. So, the score for β_j in GLM Gamma-Pareto is

The derivative of β_j is given by

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N}\alpha^{-1}\beta^{-2}\left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right)\frac{\partial \mu_i}{\partial \eta_i}x_{ij}. \tag{2.4}$$

Lastly, the jth score is presented

$$U_j = \sum_{i=1}^{N}\left[\text{var}\left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right)\right]^{-1}\left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right)\frac{\partial \mu_i}{\partial \eta_i}x_{ij}.$$

The variance U_j is

$$\text{var}(U_j) = \zeta_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{\left[\text{var}\left(\log\left(\frac{y_i}{\gamma}\right)\right)\right]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = X^T W X$$

where

$$W = \frac{1}{\left[\text{var}\left(\log\left(\frac{y_i}{\gamma}\right)\right)\right]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Since the estimators of β_j is not in closed form.

Iterative weighted least squares (IWLS) are a statistical technique used in regression analysis particularly in GLMs. IWLS is an extension of Weighted least squares (WLS). IWLS is used to find maximum likelihood estimates and standard deviations in GLMs.

IWLS can be used to address issues in G-PRM such as multicollinearity and influential observations. IWLS can help mitigate multicollinearity by weighting observations assigning weights to observations can reduce the impact of correlated predictors. To stabilize estimates, IWLS can provide more stable estimates by down-weighting observations with high leverage. IWLS can also address influential observations by down-weighting outliers assigning lower weights to observations with large residuals or leverage. In G-PRM, IWLS can be particularly useful due to the distribution's heavy-tailed nature. By addressing multicollinearity and influential observations, IWLS can improve model fit and predictive accuracy. The benefits using IWLS in G-PR can lead to improved model fit by addressing multicollinearity and influential observations. More accurate predictions by reducing the impact of influential observation and correlated predictors.

Let l_i be the log-likelihood of the response variable, then by equation (2.2)

$$l_i(\mu_i, \phi) = \sum_{i=1}^{n} \left\{ \frac{(y_i/\mu_i - \ln(\mu_i))}{-\phi} + \frac{1-\phi}{\phi} \ln(y_i) - \frac{\ln(\phi)}{\phi} - \ln\Gamma\left(\frac{1}{\phi}\right) \right\}. \quad (2.5)$$

Let $\hat{\beta}$, $\hat{\mu}$ and $\hat{\phi}$ be the maximum likelihood estimate (MLE) obtained by maximizing the log-likelihood equation (2.5) using the Newton-Raphson iterative algorithm. The MLE of β is the solution of the system of equation by setting first derivative of equation (2.5) that equals to zero, then we have

$$U(\beta) = \frac{\partial l_i}{\partial \beta} = \frac{1}{\phi}\left(y - \frac{1}{x\beta}\right) X = 0, \quad (2.6)$$

where $U(\beta)$ is the score vector with dimension $(p+1)\times 1$. Since solution of the system of equations in Equation (2.6) is nonlinear, so the iterative Fisher scoring procedure is used to estimate the unknown parameter. For iterative procedure of the G-PRM, initial values and full algorithm for the estimation of unknown parameter can be found in Hardin and Hilbe (2012). Let $\beta^{(k)}$ be the approximated MLE of β with k iterations.

$$\beta^{(k+1)} = \beta^{(k)} + \left\{ I(\beta^{(k)}) \right\}^{-1} U(\beta^{(k)}) \quad (2.7)$$

where $I(\beta^{(k)})$ is the Fisher information matrix with dimension $(p+1)\times(p+1)$ and both information and score vectors are evaluated at $\beta^{(k)}$. At convergence in deviance for Equation (2.7), the unknown parameter can be calculated as given below.

IWLS were proposed by Dobson and Barnett (2002) as a method for estimating β_j. It's the IWLS.

The iterative equation for generalized linear models (GLMs) is given by

$$X^T W X b^{(m)} = X^T W z$$

$$b^{(m)} = (X^T W X)^{-1}(X^T W z). \quad (2.8)$$

Using W and $\text{var}(U_j)$ for G-P, the iteration for β_j (where i is a number of observations $i = 1, 2, 3, \ldots, n$ and j is a number of parameters $j = 1, 2, 3, \ldots, p$ and $k \neq j$) is

$$X^T W X b^{(m)} = \sum_{k=1}^{p} \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\mathrm{var}\left(\log\left(\frac{y_i}{\gamma}\right)\right)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \frac{\left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right) x_{ij}}{\mathrm{var}\left(\log\left(\frac{y_i}{\gamma}\right)\right)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

$$z_i = \sum_{i=1}^{N} x_{ij} b_k^{(m-1)} + \left(\log\left(\frac{y_i}{\gamma}\right) - \mu_i\right) \frac{\partial \mu_i}{\partial \eta_i}.$$

There are currently numerous varieties of GLM residuals available in the literature (Hardin and Hilbe (2012)). However, we only used Pearson; the likelihood and deviance residuals are adjusted and standardized, respectively.

Pearson residuals are a type of standardized residual used in regression, especially for GLMs, that measure the difference between an observed value and a predicted value, scaled by the square root of the estimated variance, to help assess model fit by identifying influential observation or patterns.

The Pearson residuals in the G-PRM is given by

$$R_{pr} = \frac{y_i - \hat{\mu}}{\hat{\mu}}. \tag{2.9}$$

$$g(\mu_i) = \eta_i = \log(X^T \beta)$$

and fitted model, $\hat{\mu} = \eta_i = \log(X^T \beta)$.

The standardized Pearson residuals is present by using Eq. (2.9)

$$R_{spr} = \frac{R_{pr}}{\sqrt{\phi(1 - h_{ii})}}. \tag{2.10}$$

Since h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H} = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}$.

The adjusted Pearson residuals is defined by using Eq. (2.9)

$$R_{apr} = \frac{R_{pr} - \hat{E}(R_{pr})}{\sqrt{\hat{V}(R_{pr})}}. \tag{2.11}$$

$\hat{E}(R_{pr})$ and $\hat{V}(R_{pr})$ are the expected value and variance of R_{pr} respectively.

A normal distribution is closely followed by the adjusted Pearson residuals (Cordeiro (2004)).

The deviance residual for an individual observation in a GLM, such as the Gamma-Pareto Regression Model (G-PRM), quantifies that observation's contribution to the overall model deviance (a goodness-of-fit measure). It is the signed square root of the individual deviance component.

The deviance residuals for the G-PRM are given by

$$R_{di} = \mathrm{sign}(y_i - \hat{\mu}_i) \sqrt{|d_i|} \tag{2.12}$$

where

$$d_i = -2\left[\ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(\frac{y_i - \hat{\mu}_i}{p}\right)\right]$$

and $\text{sign}(y_i - \hat{\mu}_i)$ is the signum function, defined as

$$\text{sign}(y_i - \hat{\mu}_i) = \begin{cases} +1 & \text{if } y_i > \hat{\mu}_i \\ 0 & \text{if } y_i = \hat{\mu}_i \\ -1 & \text{if } y_i < \hat{\mu}_i \end{cases}.$$

Eq. (2.12) is used to present the standardized deviance residuals.

$$R_{\text{sdif}} = \frac{R_{di}}{\sqrt{1 - h_{ii}}} \tag{2.13}$$

Cox and Snell (1968) were the first to introduce adjusted residuals. The adjusted deviance residuals for both approaches are in accordance with Cordeiro (2004) and Pierce and Schafer (1986).

The adjusted deviance residuals are defined by using Eq. (2.12)

$$R_{\text{adj}} = \frac{R_{di} - E(R_{di})}{\sqrt{V(R_{di})}} \tag{2.14}$$

where $E(R_{di})$ and $V(R_{di})$ are the expected value and variance of R_{di}, respectively.

According to Pierce and Schafer (1986), the adjusted deviance residuals have a normal distribution.

In GLMs likelihood residuals, more commonly referred to as deviance residuals, are a standardized measure of the difference between the observed data and the values predicted by the model.

Similarly, the likelihood residuals (McCullagh and Nelder (1989)) for the G-PRM are given by

$$R_{\text{lik}} = \frac{l_i - \hat{l}_i}{\sqrt{N \cdot \text{Var}(l_i)}} \tag{2.15}$$

Eq. (2.15) is used to define the standardized likelihood residuals.

The standardized residual is given by

$$R_{str} = \frac{R_{it}}{\sqrt{\hat{\phi}(1 - h_{it})}} \tag{2.16}$$

Eq. (2.15) is used to define the adjusted likelihood residuals.

The adjusted likelihood residual is

$$R_{\text{alt}} = \frac{R_{it} - \hat{E}(R_{it})}{\sqrt{\hat{V}(R_{it})}} \tag{2.17}$$

where $\hat{E}(R_{it})$ and $\hat{V}(R_{it})$ are the expected value and variance of R_{it}, respectively.

3. Influence diagnostics, G-PRM

Atkinson (1981) enumerates that poor value on the LM influences model estimates and inferences. The reason is that these low figures may be outliers or effecting factors. In the explanatory variable, an extreme value will give a powerful observation whereas in the response variable, extreme value will give an outlier. Owing to the fact that the GLM applying the pearson, deviance, and likelihood residuals (standardized and adjusted) have not been dealt with yet, some of them are as discussed here to the G-PRM influence diagnostics. The reason behind this is that little attention has been given to the GLM influence diagnostics under varying GLM residuals. Pregibon (1981) was the initial author who studied residuals in the GLM. The tools of the GLM influence assessment are computed by use of different GLM residuals. A diagnostic measure that has received considerable attention in the literature is *DFFITS*, which represents the scaled difference between the fitted value with the *i*th observation removed and the fitted value of the full data.

$$\text{DFFITS}_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\phi}_i h_{ii}}}. \tag{2.18}$$

Eq. (2.18) can also be written as

$$\text{DFFITS}_i = \frac{w_{ii}^{\frac{1}{2}} x_i^T (y_i - \hat{y}_i)}{\sqrt{\hat{\phi}_i h_{ii}}}. \tag{2.19}$$

$$\text{DFFITS}_i = |t_i| \sqrt{\frac{h_{ii}}{1 - h_{ii}}}. \tag{2.20}$$

The DFFITS for standardized Pearson residuals used Eq. (2.10)

$$t_i = R_{iP} \sqrt{\frac{n-p-1}{n-p-(h_{ii})^2}} \tag{2.21}$$

The DFFITS for adjusted Pearson residuals used Eq. (2.11)

$$t_i = R_{iAP} \sqrt{\frac{n-p-1}{n-p-(h_{ii})^2}} \tag{2.22}$$

The DFFITS for standardized deviance residuals used Eq. (2.13)

$$t_i = R_{iD} \sqrt{\frac{n-p-1}{n-p-(h_{ii})^2}} \tag{2.23}$$

The DFFITS for adjusted deviance residuals used Eq. (2.14)

$$t_i = R_{iAD} \sqrt{\frac{n-p-1}{n-p-(h_{ii})^2}} \tag{2.24}$$

Table 2: Summary of Residuals

Sr. #	Residuals	Notation
1	Pearson Residuals	$R_{pr} = \frac{y_i - \hat{\mu}}{\sqrt{\hat{\mu}}}$
2	Standardized Pearson Residuals	$R_{spr} = \frac{R_{pr}}{\sqrt{\phi(1-h_{ii})}}$
3	Adjusted Pearson Residuals	$R_{apr} = \frac{R_{pr} - E(R_{pr})}{\sqrt{V(R_{pr})}}$
4	Deviance Residuals	$R_{dr} = \text{sign}(y_i - \hat{\mu}) \sqrt{\lvert d_i \rvert}$
5	Standardized Deviance Residuals	$R_{sdr} = \frac{R_{dr}}{\sqrt{\phi(1-h_{ii})}}$
6	Adjusted Deviance Residuals	$R_{adr} = \frac{R_{dr} - E(R_{dr})}{\sqrt{V(R_{dr})}}$
7	Likelihood Residuals	$R_{lr} = \text{sign}(y_i - \hat{y}_i) \sqrt{h_{ii}(R_{spr})^2 + (1 - h_{ii})(R_{sdr})^2}$
8	Standardized Likelihood Residuals	$R_{slr} = \frac{R_{lr}}{\sqrt{\phi(1-h_{ii})}}$
9	Adjusted Likelihood Residuals	$R_{alr} = \frac{R_{lr} - E(R_{lr})}{\sqrt{V(R_{lr})}}$

The DFFITS for standardized likelihood residuals used Eq. (2.16)

$$t_i = R_{iL} \sqrt{\frac{n - p - 1}{n - p - (h_{ii})^2}} \tag{2.25}$$

The DFFITS for adjusted likelihood residuals used Eq. (2.17)

$$t_i = R_{iAL} \sqrt{\frac{n - p - 1}{n - p - (h_{ii})^2}} \tag{2.26}$$

where the ith hat matrix H diagonal element for the G-PERM is $h_{ii} = \text{diag}(H)$, $H = X'(X'WX)^{-1}X'W$, McCullagh and Nelder (1989). These diagonal elements, also known as leverages, are used for influence diagnostics. The leverages are used as an indicator to impact further diagnostic procedure. If the DFFITS value exceeds one, an observation is considered influential, particularly for small datasets (Chatterjee and Hadi (1988)). For large datasets, an observation is deemed influential if its DFFITS value surpasses $2\sqrt{p/n}$ (Belsley et al. (1980)). For our study, we use a cut-off point of 2 for the diagnostic measure of DFFITS. DFFITS is used to quantify the effect of the ith influential observation on the fitted and estimated values. In a similar vein, alternative formats of standardized and adjusted G-PRM residuals can be used to identify influential observations. To compare the results with the traditional use of standardized and adjusted residuals, we use the same cut-off point for the DFFITS computation with standardized and adjusted G-PRM residuals (Table 2).

3 Numerical Results

3.1 The Monte Carlo simulation

This section aims to illustrate, through simulation, the effectiveness of the G-PRM's standardized and adjusted residuals for influence diagnostics. There are five key

Table 3: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 0.04$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{ltr}	R_{alr}
25	5	92.3	92.4	99.5	68.5	72.8	92	94.1	99.4
	10	87.2	87.1	97.4	83.5	83.7	87.5	90	98.9
	15	83.5	83.4	96.2	81	80.6	83.3	82.2	95.4
	20	76.1	76.6	93.9	72.9	72	76.4	75.9	93.7
	25	68.8	70	91.7	65.9	64.7	69.6	70.5	92.2
50	5	94	94	99.5	70	73.3	93.5	70.9	93.9
	10	86.3	86.5	98.3	83.6	83.7	86.7	85.8	90
	15	83.6	83.7	97	81	80.1	83.4	79.1	81.8
	20	76.2	76.4	93.9	72.9	72.1	76.8	73.4	76.4
	25	72.6	72.9	92.5	69.8	69.5	73.1	68.6	71.3
100	5	94.6	94.6	99.5	71.4	75.1	94.2	94	99.3
	10	88.3	88	98.5	83.2	83.5	88.4	89.4	99
	15	83.2	82.5	96	78.3	78	82.9	84.3	96.2
	20	77.6	78.1	95.9	73.6	72.8	77.7	77.1	94.6
	25	74.4	73.9	93.1	71.8	69.8	74.9	70.9	91.9
200	5	94.4	94.4	99.8	68.6	72.8	93.5	71.6	92.6
	10	88.4	88.4	98.5	85.9	86	89.1	84.7	89.7
	15	84.3	84.5	96.2	81.3	80.1	84.1	80.9	84.6
	20	78.8	79.5	95.6	76	74.7	78.6	74.6	77.4
	25	73.9	74.7	93	72.2	70.7	74.2	67.6	71.6

points that make up the independent variables. We consider the following Monte Carlo scheme to compare the G-PRM residuals' performance with DFFITS. Hanum et al. (2016)'s algorithm was utilized to create a response variable with a gamma Pareto distribution. The data generation is as follows: $y_i \sim G - P(\alpha, \beta, \gamma)$, where $\hat{\mu} = E(y_i) = (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})^{-1}$, $i = 1, 2, \ldots, n$ is mean function and ϕ is dispersion parameter $\hat{\phi} = (1/n - p - 1) \sum_{i=1}^{n} X_i^2$. The choice of dispersion values $\phi = 0.04, 0.11, 0.17, 0.33, 0.67, 2, 5, 10$ and the wide range (0.04 to 10) are used for several reasons. Assessing model performance by using a wide range of dispersion values, the study evaluates the performance of the G-PRM model and its residuals under different levels of data dispersion. Investigating a wide range helps understand how dispersion affects the detection of influential observations, which is crucial for influence diagnostics. A cut-off points of $2\sqrt{\frac{p+1}{n}}$ is employed for the diagnostic measure of DFFITS in our study. For the true parameters, we choose the following arbitrary values as $\beta_0 = 0.05$, $\beta_1 = 0.0025$, $\beta_2 = 0.005$ and $\beta_3 = 0.0001$ (Amin et al. (2016), Amin et al. (2017)) and y is minimum value of response variable. In this case, the design matrix $\mathbf{X}$ has no influential points of sample sizes $n = 25, 50, 100$ and 200 generated as $X_i \sim N(-1, 1)$, $i = 1, 2, \ldots, n$; and $j = 1, 2, 3$, and then we make 5^{th}, 10^{th}, 15^{th}, 20^{th}, 25^{th} points in the $\mathbf{X}$ as $X_{ij} = \alpha_0 + X_{ij}$, $i = 5, 10, 15, 20$ and 25, and $j = 1, 2, 3$, where $\alpha_0 = \bar{X}_j + 100$. For the estimation of G-PRM, the link function used is inverse link function. The detection percentage has been calculated to observe the performance of proposed procedure. The process is replicated 10,000 times.

Tables 3 to 10 present the performance of DFFITS in G-PRM, showcasing its effectiveness in detecting influential observations. The results indicate that the dispersion parameter ϕ has a substantial impact on the detection percentage for all methods, except for deviance residuals. Notably, the detection rate increases as ϕ increases, highlighting

Table 4: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 0.11$

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	93.1	93.1	99.8	66.9	72.8	93	92.2	92.2	99.7
	10	87.5	87.7	97.9	82.2	82.9	86.9	88.9	88.8	98.2
	15	82.4	82.7	96.4	79.8	78.8	81.9	82.1	82.2	96.3
	20	79	79.3	95.7	76.5	76.2	79.4	77.2	77.5	94.5
	25	71.6	73.3	93.6	69.7	68.2	72.6	72.4	73.6	92.5
50	5	93.4	93.4	99.6	68.5	74.1	92.6	67.3	70.3	91.3
	10	87.7	88.1	98	84.6	84.9	88.2	84.7	85	89
	15	83	83.2	96.8	80.4	79.8	82.7	79.4	78.8	82.1
	20	75.9	76.5	94.1	72.8	71.8	76	74.8	73.4	77.4
	25	71.8	73.4	93.4	69.2	67.3	72.5	70	68.5	72.9
100	5	94.4	94.5	99.5	70.3	75.3	93.9	94	94.1	99.6
	10	89.5	89.1	99	85.1	85.1	89.5	87.9	87.6	97.8
	15	82.6	83	96.3	79.7	79	82.9	82.5	82.3	96.3
	20	76.4	76.4	95.1	73.3	71.8	76.7	78.5	77.9	94.4
	25	73	73.9	92.7	70.5	70.5	73.7	70.8	71.9	92.1
200	5	93.9	93.9	99.7	70	73.2	93.4	70.6	74.9	93.6
	10	90.8	90.7	98.7	85.8	86.2	90.8	82.6	83.5	88.9
	15	84.3	84.2	97.3	80.8	79.8	84.3	79.2	78.5	82.8
	20	75.5	76.3	92.9	73.3	71.9	76.1	76.2	75.4	78.6
	25	73.1	73	91.7	71.3	69.8	73.2	69.3	67.6	71.4

Table 5: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 0.17$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	92.7	92.7	99.6	68.6	71.4	92.9	94.2	94.2	99.8
	10	87.4	87.4	97.8	84	84.1	87.5	88.6	88.7	98.9
	15	83	82.7	96.4	79.8	78.6	83.4	81.1	81.3	96.1
	20	77.7	78.1	94.5	74.8	74.4	78	78.2	78.4	94.3
	25	71.1	71.7	90.6	68.6	67.2	71.7	71.5	72.7	92.9
50	5	94.6	94.6	99.6	71.6	73.4	93.9	67.6	74.2	94
	10	87.9	87.8	98.6	83.7	83.4	88.2	85.3	85.4	89.3
	15	84.4	84.1	96.3	81.8	80.9	84.7	78.7	77.5	81
	20	75.4	75.5	93.4	72	70.3	75.6	75.5	74.8	78.6
	25	68.9	70.4	91.5	66	65.4	69.9	68.1	68.1	72.1
100	5	94.4	94.4	99.7	68.1	73.4	93.8	93	93	99.5
	10	86.6	87	98.6	83.3	82.8	87	89.5	89.5	98.2
	15	84.3	84.2	96.9	79.5	79.1	84	83.9	83.7	96.3
	20	80.3	80.6	95.9	77.4	76.6	80.4	76.8	78	93.2
	25	73.5	74.5	91.8	70.6	70.1	74.3	73.1	73.8	93
200	5	94	94	99.8	69.8	72.9	94.3	69.7	74.9	92.9
	10	87.9	87.7	97.5	84.1	84	88.1	85.4	85.2	89.3
	15	83.7	84	96.4	80.8	80.1	83.5	79.3	78.6	83.4
	20	79.7	80.2	95.6	77.1	75.9	79.8	73.9	72.5	76.8
	25	70.4	72.6	92.5	68.8	67.5	71.2	70.9	69.2	73.4

the importance of considering dispersion levels when evaluating model diagnostics. Interestingly, sample size also plays a significant role in detecting influential observations. As the sample size increases, the detection percentage rises, and at very large sample sizes, all influential observations are identified. This trend is consistent across all residual types, suggesting that larger sample sizes can improve the accuracy of influence diagnostics. A closer examination of the results reveals that the behavior of each residual is essentially the same, with adjusted Pearson residuals performing better for small and large sample sizes. However, for large sample sizes, all residuals perform almost equally well, indicating that the choice of residual type becomes less critical as the sample size increases. These findings emphasize the importance of considering both dispersion levels and sample size when evaluating the performance of influence

Table 6: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 0.33$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	94.7	94.7	100	68.6	73.4	95.1	94.3	94.3	99.4
	10	88	88.1	97.2	83.4	84	88	87.8	87.9	98.3
	15	82.7	82.9	96.6	80.2	80	82.8	83.1	83.2	96.4
	20	78.7	79.1	96.6	76.6	74.8	78.8	77.2	78.3	95.5
	25	68.7	69.9	90.6	66.6	65.9	69.8	71.5	73.2	92.5
50	5	94.4	94.4	99.5	67.9	72.3	94.3	68.2	71.8	93.7
	10	89.5	89.5	99	85.8	86.5	89.7	84.8	84.3	87.8
	15	80.3	80.3	95.6	76.4	75.4	80.5	80.3	79.7	82.7
	20	78.4	78.3	95.7	75.6	74.3	78.9	74.8	73.3	77.4
	25	70.4	71.8	91.6	67.2	66.4	71.1	68.9	67.7	72.4
100	5	92	92	99.2	68.7	73.1	91.9	93.3	93.3	99.4
	10	88.6	88.7	98.1	85.4	84.7	88.6	88.8	88.7	98.4
	15	83.6	83.7	96.4	81.1	80.5	83.9	82.7	83	96.5
	20	78.4	78.9	94.1	75.7	74.5	78.4	76.9	77	94.6
	25	70.8	72.6	91.7	68.5	67.1	71.8	72.2	73.8	92.6
200	5	95	94.8	99.8	69.9	75.2	93.4	68.4	72.6	93.4
	10	89.1	89	98.1	84.7	85.2	89.2	85.6	85.5	88.6
	15	81.3	81.2	96.1	77.5	76.6	81.5	80.4	80.2	83
	20	79	78.9	95.1	75.9	75.1	79.2	74.2	72.9	77
	25	72.8	74.1	92.9	70.1	69.1	73.7	70.1	69.7	73.4

Table 7: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 0.67$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	94.1	94	99.7	69.6	74.4	93.8	93.6	93.6	99.5
	10	88.3	88.2	97.4	83.3	83.2	88.4	89.6	89.6	98.8
	15	84	83.6	97.1	80.9	79.6	84.4	82.3	82.8	96
	20	79.7	79.5	94.9	76.5	75.9	79.8	77.6	78.1	94.3
	25	68.6	70.2	90.9	66.2	66	69.4	71.2	73.3	90.9
50	5	94.5	94.5	100	70.6	75.8	93	66.9	72.7	93.1
	10	88.6	88.5	98	84.1	84	88.4	87.3	86.8	89.5
	15	82.5	82.9	96.9	79.3	78.3	82.3	80.1	79.6	82.8
	20	77.6	78.5	94	75.3	74.3	77.6	74.5	73.6	77.9
	25	71.5	72.2	92.2	69.3	68	72.3	69.1	68.5	72.4
100	5	92.9	92.9	99.7	69.4	73.2	92.8	92.9	93	99.4
	10	88.9	88.8	97.9	85	85	88.8	87	86.9	97.8
	15	82.4	82.5	96.9	77.9	77.9	81.7	84.3	84.1	96.3
	20	79.1	80	94.4	76.7	75.8	79.2	77.8	78.4	95.6
	25	72.3	73.5	92.5	70.1	69.3	73.4	71.7	73.7	92.8
200	5	95.9	95.9	100	70.5	75.2	94.8	72.6	74.3	91.9
	10	87.9	88.2	98	84.8	84.9	87.5	83.4	83.8	87.4
	15	81.4	81.3	95.5	77.8	76.7	81.6	80.9	80.1	84.1
	20	77.2	77.2	94.4	74	73.5	77.3	74.5	73.4	77.9
	25	70.5	72.6	92.6	67.6	67	71	69.4	68.1	72.7

diagnostics in G-PRM. The results suggest that adjusted Pearson residuals are a good choice for detecting influential observations, particularly in small sample sizes or when dispersion levels are high. Figure 1 displays the simulation results (for $n = 25, 50, 100$ and 200) graphically.

Table 8: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 2$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	93.7	93.9	99.2	69.6	74.3	92.6	93.8	93.8	99.9
	10	89.1	89.2	98.7	84.5	84.6	88.6	88.8	89	97.8
	15	86.3	86	97.5	83.1	82.2	86.3	82.2	82.7	96.4
	20	77.4	77.4	93.8	75	73.1	77.7	78	78.3	94.2
	25	72	73.2	94	69.5	68.7	72.3	72.4	72.8	92.9
50	5	94.5	94.4	99.8	70.4	75.3	93.9	68.4	72.1	94.2
	10	88.1	88.3	98.3	84.6	85.2	88.1	84.1	84.4	88.5
	15	83.6	83.8	96.1	80.2	79.5	83.6	79.4	78.7	82
	20	77.5	78.2	95	75.6	74.4	77.9	75.3	74.5	78
	25	70.1	71.3	92.1	68.2	66.9	71.1	69.9	68	73
100	5	93.8	93.8	99.6	69.4	73.6	92.5	68.4	72.1	94.2
	10	88.5	88.4	98.3	84.6	85.1	88.8	95.3	95.3	99.7
	15	81	81.1	95.6	77.3	75.8	81	87.5	87.5	97.4
	20	77.3	77.2	94.8	74.7	72.8	77.4	85.3	85.3	97.1
	25	69.8	70.4	91.3	67.3	65.9	70.7	79.7	79.1	95
200	5	94.1	94.1	99.4	69.9	74.4	93.1	70.4	71.6	91
	10	90	90.2	97.7	85.6	85.6	90.3	73.9	77.2	94.9
	15	82.6	82.9	96.2	79.6	78.6	82.3	82.4	82.9	87.6
	20	78.6	78.4	93.4	75.8	75.3	78.7	83	82.2	85.4
	25	73.1	74.6	92.5	70.3	68.8	74	76.7	76	80

Table 9: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 5$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	93.7	93.8	99.4	69.2	73.6	93.8	93.9	93.9	99.7
	10	87.8	87.9	97.9	83.9	84	87.8	88.9	89.1	97.7
	15	83	82.7	96.5	79.2	78.2	83.2	82.4	83	96.3
	20	76.6	76.3	94.5	73.8	72	76.6	75.1	75.3	93.9
	25	73.5	74.3	92.6	71.7	70.5	74	71.7	73.1	91.4
50	5	95.2	95.2	99.8	71.5	74.2	95	72.5	76.5	93.3
	10	88.3	88.2	98.4	85.4	84.9	88.2	83.8	84.1	88.2
	15	83.4	83.3	96.1	80.3	79.2	83.2	79.7	78.7	82.4
	20	76.2	77.1	93.7	73.8	72.5	76.5	72.5	71.1	75.5
	25	70.7	72.2	93	68.2	68.2	71.5	69.4	68.5	72.2
100	5	94.1	94.1	99.7	70.6	72.7	93.9	94.6	94.6	99.6
	10	89.7	89.8	97.7	86.3	86.9	90.7	88.2	88.3	98
	15	83	82.7	96.1	80.2	79.4	83.5	85.1	85.2	96.9
	20	78.3	79.6	95.2	75.6	74.4	78.4	76.2	76.6	94.6
	25	71.2	72.7	91.5	68.5	67.2	72.1	71.7	72.8	92.5
200	5	94.9	94.9	99.8	69.4	75.5	94.6	71.3	74.8	93.8
	10	87.4	87.4	98.6	84.2	83.8	87.7	83.3	83.6	88
	15	83	82.9	96.6	80.2	79.4	83	82	81.3	84.9
	20	79.8	80.5	95.5	76.2	76.2	79.8	72.6	71.7	76.2
	25	74.2	75	92.9	71.8	70.7	74.8	69.9	68.4	72.2

Table 10: Estimated influential observations detection rate (%) of the G-PRM with different residuals when dispersion $\phi = 10$.

n	Influential points (IP)	R_{pr}	R_{spr}	R_{apr}	R_{dr}	R_{sdr}	R_{adr}	R_{lr}	R_{slr}	R_{alr}
25	5	93.4	93.4	99.7	69	73	92.8	94.6	94.7	99.8
	10	88.8	88.8	98.3	84.1	84.1	88.3	87.7	87.5	98
	15	83.6	84	96.3	80.3	78.9	83.7	85.1	85.2	97.2
	20	76.7	77.2	95	73.8	72.9	77	76.1	76.9	94.9
	25	71.3	71.8	92.6	68.5	67.4	71.9	73.3	74.1	93
50	5	92.6	92.6	99.5	69.7	74.2	92.5	65.9	72.6	93.9
	10	88.5	88.7	98.9	85	84.4	89	84.4	84.3	87.7
	15	83.6	83.7	96.3	79.5	78.5	83.2	82.3	82.3	84.7
	20	78.4	79.2	95.7	75.9	74.6	78.7	73.7	72.9	76.3
	25	73.6	75.4	92.8	70.4	69.7	74	71.8	70.6	73.8
100	5	93.2	93.1	99.9	69.8	73.8	91.9	94.1	94.1	99.9
	10	89.7	89.7	98	86.2	86.4	89.6	87.8	87.8	97.6
	15	83.4	83.8	95.9	79.7	79.1	83.6	83.3	83.4	97.4
	20	76.5	77.8	94.6	73.7	72.6	76.7	77.7	77.7	94.6
	25	70.6	70.8	91.1	67.9	65.8	71.1	71	72.2	92.4
200	5	94.2	94.2	99.8	69.9	73	94	67.7	74	93.1
	10	88	87.9	97.8	83.9	84.1	88.4	84.4	84.4	87.5
	15	81	81.8	96.3	77.3	76.4	81	79	78	83.3
	20	76.8	77	94.4	74.6	73.9	77.6	75.4	74.5	78.1
	25	70.9	71.5	92.1	68.6	67.2	71.6	68.4	67.1	71.6

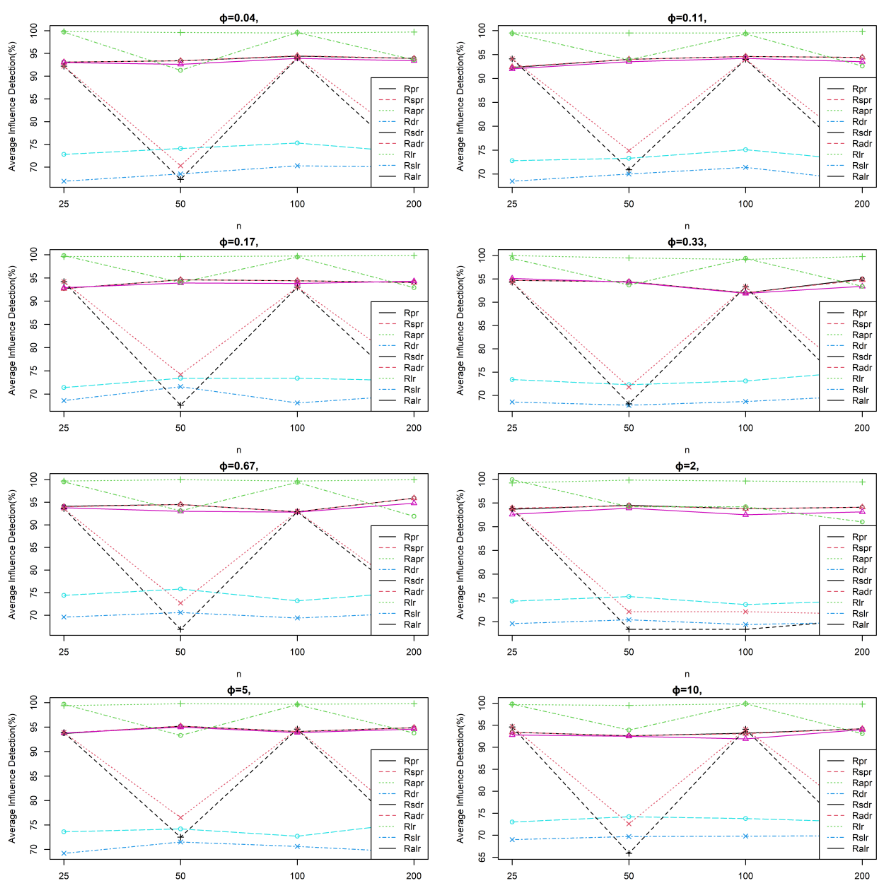

Figure 1: Simulation results for different sample sizes.

4 Application: ARDENNES data

This time we assessed the effectiveness of the suggested method with the help of the practical application. To this end, we used ARDENNES data set retrieved by Amin et al. (2022) and Barnard et al. (1995). The primary objective of this data set was to find the first etch biopsy, or the starting point of a removed layer of incisor enamel (Y), based on two explanatory variables of the data obtained on 55 children. The first among these explanatory variables was the etched depth (X_1) that was obtained by measuring the amount of calcium extracted during the etch biopsy. The second explanatory variable was the age of children (X_2) that was recorded as years and months, but converted to the decimal system. The dataset doesn't fit a normal distribution due to the positive skewness of the response variable. To identify the optimal regression model, we tested the probability distribution of the response variable using Anderson-Darling, Cramer-von Mises, and Pearson Chi-square tests (Zhang (2001); Evans et al. (2008)). Table 11 shows that the ARDENNES dataset fits the Gamma-Pareto distribution (G-PD) well. We analyzed the performance of G-PRM residuals and used G-PRM to identify influential points. Based on the distribution fitting test, the G-PD fits the dataset well, with results presented in Table 11.

The Gamma-Pareto distribution yields the largest p-values across all three tests (AD, CVM, PCS), meaning the data do not reject the null hypothesis of a good fit. In contrast, the Gamma model shows moderate fit, while the Weibull (and Pareto) exhibit lower p-values, indicating poorer agreement with the ARDENNES data. The G-PD's flexibility (combining Gamma body with Pareto tail) captures the data's skewness and heavy-tail behavior better than the single-shape Gamma or Weibull models.

The fitted G-PRM for inverse link function using real data is given by,

$$\hat{Y}_i = (0.6700[0.092, S] - 88.689x_1[0.0007, S] + 57.982x_2[0.0002, S])^{-1}$$

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.

Table 11: Goodness of Fit Distribution Tests for ARDENNES Data.

Goodness of fit test (GFT)	Probability Distribution						
	Gamma	Pareto	Gamma-Pareto	Weibull	Weibull-Pareto	Normal	Normal-Pareto
Anderson-Darling (AD)	0.3553	0.8110	0.5832	0.5883	0.4897	1.0945	3.0023
	(0.4699)	(0.0790)	(0.6145)	(0.1279)	(0.3451)	(0.0066)	(0.6340)
Cramer-von Mises (CVM)	0.0539	0.9823	0.5941	0.0725	0.1231	0.1405	0.7231
	(0.4617)	(0.0012)	(0.7980)	(0.2542)	(0.0023)	(0.0311)	(0.0003)
Pearson chi-square (PCS)	5.5454	11.430	15.967	7.7273	4.889	10.636	2.7350
	(0.5937)	(0.5612)	(0.8070)	(0.3572)	(0.4432)	(0.1553)	(0.0654)
Gamma-Pareto Distribution (G-PD)							

In Table 13, numerical evaluation of G-PRM with and without influential observations is presented. For G-PRM, higher value of MSE and of R^2 is observed when influential observations found in the data. But the reduction in MSE and R^2 is observed

Table 12: Detect influential points (IP) with Gamma-pareto regression model (G-PRM). Residuals

IP detection methods	G-PRM residuals		
	Pearson	Deviance	Likelihood
Index plots residuals	48, 52	48	12, 23, 48, 52
Index plots standardized residuals	5, 23, 48, 52	5, 23, 48	1, 4, 5, 6, 7, 9, 10, 13, 16, 23, 28, 29, 48, 52
Index plots adjusted residuals	5, 23, 48, 52	5, 23, 48	1, 4, 5, 6, 7, 9, 10, 12, 23, 28, 29, 30, 37, 38, 39, 41
Influential points (IP), Gamma-Pareto regression model (G-PRM)	43, 48, 52		

Table 13: Influential observations' effect in G-PRM

With influential observations (G-PRM)		Without influential observations (G-PRM)	
MSE	R^2	MSE	R^2
18.453	0.890	18.453	0.890

when influential observations were removed from the data. The Anderson-Darling (AD), Cramer-von Mises (CVM), and Pearson chi-square tests are used to test the probability distribution of the dependent variable to find the best regression model (for more details, see Zhang (2001)). Table 11 shows that the ARDENNES data fits the G-PD quite well. As a consequence, we use the G-PRM to find influential points and compare the performance of the G-PRM residuals. Residuals, Standardized and adjusted version of the G-PRM residuals are separated in order to determine the influential points based on DFFITS.

These influential points detection methods using the G-PRM standardized and adjusted residuals are tabulated in Table 12. Table 12 shows the 1st, 4th, 5th, 6th, 7th, 9th, 10th, 12th, 13th, 16th, 23th, 28th, 29th, 30th, 37th, 38th, 39th, 41th, 43th, 48th, 52th points are influential for given data.

The graphical results of Table 12 are displayed in Figures 2 and 3.

5 Conclusion

In conclusion, our study demonstrates the effectiveness of the DFFITS statistic in identifying influential observations in the G-PRM using different residuals. Based on our simulation study and real data analysis, we recommend using the adjusted Pearson residual class for detecting influential observations, as it performs best across various scenarios. Detection rates improve with larger sample sizes, and the dispersion parameter significantly impacts the identification of influential observations. Specifically, as the dispersion parameter increases, the detection percentage also increases.

However, our study has some limitations. We only investigated the inverse link

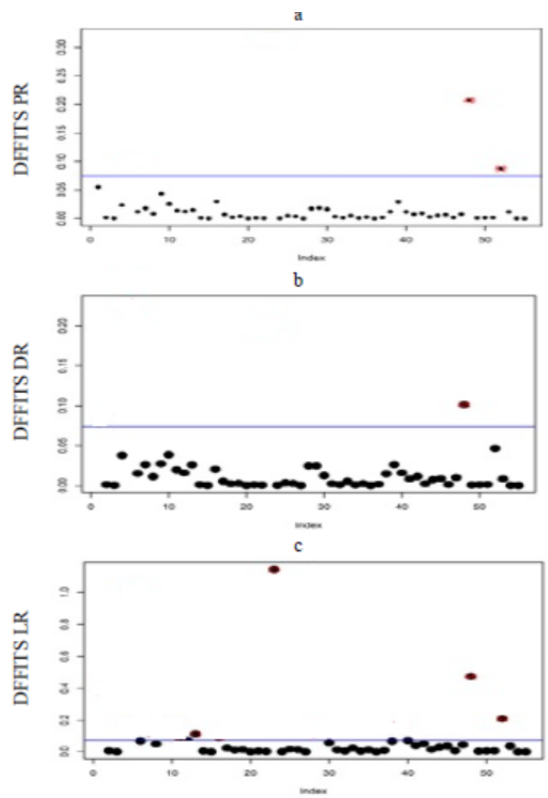

Figure 2: Graphical results for influential point detection (part 1).

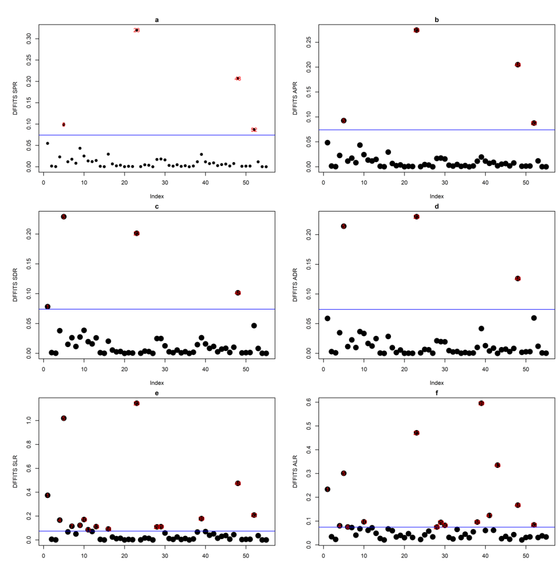

Figure 3: Graphical results for influential point detection (part 2).

function, and future research should explore other link functions to assess their impact on detection rates. Additionally, our simulation study considered a range of dispersion parameters ($\phi = 0.04$ to 10), and further research is needed to evaluate the performance of DFFITS for other ranges of ϕ.

For practical applications, we provide the following guidance. For small sample sizes ($n < 25$), use adjusted Pearson residuals, especially when ϕ is small (e.g., $\phi = 0.04$). For large sample sizes ($n \geq 200$), all residuals converge, and detection rates are high, regardless of the residual type. When ϕ is large (e.g., $\phi = 10$), detection rates are higher, and adjusted Pearson residuals are recommended.

Funding

No funding resource is available for this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Nasir Saleem https://orcid.org/0000-0001-6556-4737
Atif Akbar http://orcid.org/0000-0003-3227-9731

References

Alzaatreh A, Famoye F, Lee C. Gamma-Pareto distribution and its applications. Journal of Modern Applied Statistical Methods. 2012;11(1):7.

Alzaatreh A, Ghosh I. A study of the Gamma-Pareto (IV) distribution and its applications. Communications in Statistics-Theory and Methods. 2016;45(3):636–654.

Alzaghal A. The exponentiated gamma-Pareto distribution with application to bladder cancer susceptibility. J Stat Appl Probab Lett. 2020;7:39–56.

Amin M, Afzal S, Akram MN, Muse AH, Tolba AH, Abushal TA. Outlier detection in gamma regression using Pearson residuals: Simulation and an application. AIMS MATHEMATICS. 2022;7(8):15331–15347.

Amin M, Amanullah M, Aslam M. Empirical evaluation of the inverse Gaussian regression residuals for the assessment of influential points. Journal of Chemometrics. 2016;30(7):394–404.

Amin M, Amanullah M, Aslam M, Qasim M. Influence diagnostics in gamma ridge regression model. Journal of Statistical Computation and Simulation. 2019;89(3):536–556.

Amin M, Amanullah M, Cordeiro GM. Influence diagnostics in the gamma regression model with adjusted deviance residuals. Communications in Statistics-Simulation and Computation. 2017;46(9):6959–6973.

Amin M, Faisal M, Akram MN. Influence diagnostics in the inverse Gaussian ridge regression model: Applications in chemometrics. Journal of Chemometrics. 2021;35(6):e3342.

Amin M, Qasim M, Amanullah M, Afzal S. Performance of some ridge estimators for the gamma regression model. Statistical papers. 2020;61:997–1026.

Anholeto T, Sandoval MC, Botter DA. Adjusted Pearson residuals in beta regression models. Journal of Statistical Computation and Simulation. 2014;84(5):999–1014.

Ashour SK, Said DR, Fahim MAA. The Log-Gamma-Pareto Distribution. 2014;.

Atkinson AC. Two graphical displays for outlying and influential observations in regression. Biometrika. 1981;68(1):13–20.

Atkinson AC. Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis; 1985.

Barnard TE, Booksh KS, Brereton RG, Coomans DH, Deming SN, Hayashi Y, et al. Chemometrics in environmental chemistry-statistical methods, vol. 2. Springer; 1995.

Belsley DA, Kuh E, Welsch R. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley; 1980.

Chatterjee S, Hadi AS. Sensitivity Analysis in Linear Regression. Wiley: New York; 1988.

Cook RD. Detection of influential observation in linear regression. Technometrics. 1977;19(1):15–18.

Cook RD. Assessment of local influence. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1986;48(2):133–155.

Cook RD, Weisberg S. Residuals and influence in regression; 1982.

Cordeiro GM. On Pearson's residuals in generalized linear models. Statistics & Probability Letters. 2004;66(3):213–219.

Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society: Series B (Methodological). 1968;30(2):248–265.

Dar AA, Ahmed A, Reshi JA. weighted gamma-pareto distribution and its application. Pakistan Journal of Statistics. 2020;36(4).

De Andrade TA, Fernandez LMZ, Silva FG, Cordeiro GM. The gamma generalized pareto distribution with applications in survival analysis. 2017;.

Dobson AJ, Barnett AG. An introduction to generalized linear models. CRC press; 2002.

Evans DL, Drew JH, Leemis LM. The distribution of the Kolmogorov–Smirnov, Cramer–von Mises, and Anderson–Darling test statistics for exponential populations with estimated parameters. In: Computational probability applications.; 2008. p. 165–190.

Hadi AS. A new measure of overall potential influence in linear regression. Computational Statistics & Data Analysis. 1992;14(1):1–27.

Hanum H. Modeling Extreme Rainfall with Gamma. Applied Mathematical Sciences. 2015;9(121):6029–6039.

Hanum H, Wigena AH, Djuraidah A, Mangku IW. Developing generalized linear model of Gamma-Pareto Distribution. Far east Mathematical Sciences Journal. 2016;.

Hardin JW, Hilbe JM. Generalized Linear Models and Extensions. 3rd ed. USA: Stata Press Publication; 2012.

Khan JA, Akbar A. Empirical performance of nonparametric regression over LRM and IGRM addressing influential observations. Journal of Chemometrics. 2019;33(7):e3143.

Lawless JF. Statistical Models and Methods for Lifetime Data. New York: Wiley; 2003.

Lee AH. Assessing partial influence in generalized linear models. Biometrics. 1986;44:71–77.

McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman & Hall; 1989.

Nelder J, Wedderburn R. Generalized Linear Models. Journal of the Royal Statistical Society, Series A (General). 1972;135(3):370–384.

Pierce DA, Schafer DW. Residuals in generalized linear models. Journal of the American Statistical Association. 1986;81(396):977–986.

Pregibon D. Logistic regression diagnostics. The Annals of Statistics. 1981;9(4):705–724.

Preisser JS, Qaqish BF. Deletion diagnostics for generalised estimating equations. Biometrika. 1996;83(3):551–562.

Saleem N, Akbar A. Empirical evaluation of the gamma-Pareto regression residuals for influence diagnostics. Journal of Statistical Computation and Simulation. 2024;94(16):3573–3598.

Simas AB, Cordeiro GM. Adjusted Pearson residuals in exponential family nonlinear models. Journal of Statistical Computation and Simulation. 2009;79(4):411–425.

Thomas W, Cook RD. Assessing influence on regression coefficients in generalized linear models. Biometrika. 1989;76(4):741–749.

Vanegas LH, Rondón LM, Cordeiro GM. Diagnostic tools in generalized Weibull linear regression models. Journal of Statistical Computation and Simulation. 2013;83(12):2315–2338.

Williams D. Generalized linear model diagnostics using the deviance and single case deletions. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1987;36(2):181–191.

Zhang J. Powerful goodness-of-fit and multi-sample tests. PhD thesis, Toronto: York University; 2001.

Zheng L, Ismail K, Meng X. Shifted Gamma-Generalized Pareto Distribution model to map the safety continuum and estimate crashes. Safety Science. 2014;64:155–162.