Osprey Optimization for Influence Diagnostics in Regression Models

Luay Adil Abduljabbar and Sabah Manfi Ridha
Department of statistics, College of Administration and Economy, University of Baghdad, 10071 Baghdad, Iraq
{luay.abd2201, drsabah }@coadec.uobaghdad.edu.iq

Keywords: Cooks Distance, DFFITS, Gamma Regression Model, Osprey Optimization Algorithm.

Abstract: Diagnosing of influence is an originality that is necessary to help identify influential observations, which affect inference, particularly estimation of the model. Diagnostics such as Cooks Distance and DFFITS are well-known classical diagnostic tools that may turn out to be inappropriate or not in complexities of models or at varying category dispersion of data. In the present article, a new attempt to extend the field of influence diagnostics to the Gamma regression models (GRM) is attempted to apply the metaheuristic algorithm referred to as the Osprey Optimization Algorithm (OOA). In order to compare the GRM detecting power of Cook s Distance and DFFITS as well as the OOA, we perform a far-reaching simulation study of the two sample sizes and dispersion parameters. The simulation outcomes confirm that the Cook Distance as well as the DFFITS are effective yet OOA-enhanced diagnostic scheme is superior to identify influential cases especially in conditions of high dispersion and in-limited to medium sample. When described through the prism of the compared analysis, one may claim OOA is more comprehensive in terms of the detection mechanism.

1 INTRODUCTIONS

One of the statistical methods is regression that recognizes the relationship perfectly between the response variable and the explanation one. Regression can help us to decipher and examine the impact of one or more predictors on the response variable. The field of its application is great in commercial, industrial, medical, etc. sphere. The predictive and the forecasting analysis of the variables of interest could be aided by the regression analysis [1].

The gamma regression model (GRM) is a wide variety of generalized linear models (GLM) which are applied to continuous, positively valued, strictly positive response variables which tend to be right-skewed [2]-[4]. In contrast to traditional linear regression that expects constant variance and normality of distributions, gamma regression is quite suitable to be applied to those data where the variance is growing in tandem with the mean as is a common property of gamma-distributed outputs [4]. Under this model, the dependent variable is gamma distributed and it is represented by two parameters including a shape parameter and a scale parameter. The correlation of the mean and the variance is affirmed by the fact that the variance is a multiple of the mean squared to ensure effectiveness of the model to counter heteroscedasticity which occurs when the data is skewed [5]-[7].

Influence diagnostics in regression analysis refers to efforts to determine specific values of independent data variables, known as influential observations, which have unusually large influence on the estimated model parameters or the fit in general [8], [9]. Such observations may have significant influence on parameter estimates, standard errors, predictions and statistical inferences and may result into misleading findings, unless identified and correctly addressed [10].

The major aim of influence diagnostics is to assess the changes in the main elements of regression model due to the removal or adjustment of each observation [11]-[13]. The measures commonly calculated in techniques include Cooks distance, DFBETAS (changes in parameter estimates), DFFITS (changes in fitted values), leverage values and changes in deviance or Pearson residuals [14]. These measures assist in signifying information that may force model behavior unduly.

Influence diagnostics were originally developed with application to classical linear regression but have been seen applied to GLMs, such as Poisson, logistic, and GRM and models of overdispersion or multicollinearity. In particular, the detection of influential observations is critical in complex models under which the assumptions are more sensitive to the outliers or leverage [15].

A higher-level mathematical framework or strategy known as a metaheuristic algorithm is employed to address optimization issues that are challenging, intricate, or impractical to resolve with conventional optimization methods. Usually, natural processes like simulated annealing, swarm intelligence, or evolutionary processes serve as the inspiration for these algorithms [16]. Although they seek to locate a passably decent answer in a reasonable period, metaheuristic algorithms do not promise an ideal solution. They are frequently used in situations when more conventional optimization techniques are unworkable or wasteful, particularly in situations involving huge search spaces or non-linear connections.

In regression modelling, influential observations may have a greater influence than warranted on parameter estimates, standard errors and overall model fit and may result in biased or unstable Inference. Variants of the classical influence diagnostic procedures, cooks distance, DFBETAS, leverage measures and residual-based statistics are based on analytic or deletion-based calculations, and can (as models grow in complexity, particularly with high-dimensional data, multicollinearity, overdispersion, or nonlinearity) be computationally costly or less useful. General-purpose optimization algorithms motivated by natural process inspirations (e.g., genetic algorithms, particle swarm optimization, simulated annealing), so-called meta-heuristic methods, may provide some potential benefits in the context of influence diagnostics because they efficiently handle large parameter spaces, can identify influential points using optimization criteria, do not require closed-form solutions, and do not require case-by-case deletions, since they can be equally applicable in creating and building targets. Lemonte (2009) [17] studied methods for diagnosing influential findings of the Birnbaum Saunders nonlinear regression model which uses three diagnostics, including the case-deletion method which measures the impact of eliminating an observation on the parameter estimates. Arimiyaw Zakaria, Nathaniel Kwamina Howard and Bismark Kwao Nkansah (2014) [18] dealt with the issue of identification of influential observation in linear regression analysis and compared various classical diagnostic measures which includes Distance of Cook Studentized Deleted Residuals., Muhammad Amin, Muhammad Amanullah, and Gauss M. Cordeiro (2016) [19] developed modified deviance residuals for the gamma regression model to diagnose influential observations, subsequently comparing them with classical methods such as the Pierce and Schafer residuals. Amin, Amanullah, Aslam, and Qasim (2018) [20] they suggest diagnostic methods for finding influential values in the standard (GRM) and in the (GRM) under multicollinearity. The authors Aamna Khan, Muhammad Amanullah, Muhammad Amin, Randa Alharbi, Abdisalam Hassan Muse, and M. S. Mohamed (2021) [21] presented in their study the classical diagnostic methods for Poisson regression in the presence of multicollinearity and influential observations. Saleem, Akbar, Laeeq, Ahmad, and Hanum (2025) [22] conducted a comparison of link functions (logarithmic, inverse, and identity) in the Gamma-Pareto regression model using classical diagnostic measures such as Cook's Distance and DFFITS.

2 METHODOLOGY

This study helps to fill this gap by suggesting and comparing meta-heuristic methods as resilient inference diagnostics strategies, with an aim of increasing the detection yield in settings where prior approaches fail because of model and data complexity, dimensionality, or computational overhead. Osprey optimization algorithm is employed as an influence diagnostic in gamma regression model, applying metaheuristic algorithms to statistical diagnostics and provides practitioners with a powerful tool for improving the reliability of Gamma regression analyses.

3 GAMMA REGRESSION MODEL

Model Gamma regression model Positively skewed. there are many cases in which data emerge in epidemiological, social, and economic studies. This type of data consists of nonnegative values. A popular distribution that fits such type of distribution is gamma distribution. data. GRM is applied in order to model the correlation between the skewed response variable which is positive. and potentially regressors [23].

Let y_i denotes the response variable and is distributed as a gamma. where is the shape parameter ν and is the scale parameter of the distribution γ, i.e.

$y_i \sim Gamma(\nu, \gamma)$, then the probability density function is such that:

$$f(y_i) = \frac{\gamma}{\Gamma\nu}(\gamma y_i)^{\nu-1} e^{-\gamma y_i}, \quad y_i \geq 0, \quad (1)$$

with $E(y) = \nu/\gamma = \theta$ and $var(y) = \nu/\gamma^2 = \theta^2/\nu$. Given that $\gamma = \nu/\theta$, (1) can re-parameterized with the shape (ν) parameters and the mean (θ), as a function of the mean and the shape parameters and is written as follows:

$$f(y_i) = \text{EXP}\left\{\frac{y_i(-1/\theta) - log(-1/\theta)}{1/\nu} + c(y_i, \nu)\right\}, \quad (2)$$

where the canonical link function is, the dispersion parameter is $-1/\theta$, the dispersion parameter is $\phi = 1/\nu$ and $c(y_i, \nu) = \nu\, log(\nu) + \nu\, log(y_i) - log(y_i) - log(\Gamma(\nu))$.

GRM in terms of the. canonical link function (reciprocal), $\theta_i = -1/x_i^T\beta$ that is given as a linear combination of covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$. The log link function, $\theta_i = exp(x_i^T\beta)$, The reciprocal link function is not included but the log link function, is used instead. ensures that $\theta_i > 0$.

The most used way of estimating the coefficients of GRM is to employ the. maximum likelihood method of (2). Considering the assumption of the observations being; independent and $\theta_i = -1/x_i^T\beta$, the log-likelihood function is

$$\ell(\beta) = \sum_{i=1}^{n}\left\{\frac{y_i x_i^T\beta - log(x_i^T\beta)}{1/\nu} + c(y_i, \nu)\right\}, \quad (3)$$

the ML estimator is then obtained by computing the first derivative of the (3) and setting it equal to zero, as:

$$\frac{\partial\ell(\beta)}{\partial\beta} = \frac{1}{\nu}\sum_{i=1}^{n}\left[y_i - \frac{1}{x_i^T\beta}\right]x_i = 0. \quad (4)$$

Regrettably, First derivative cannot be solved analytically due to the presence of (4) is nonlinear $\boldsymbol{\beta}$. The least squares algorithm or Fisher-scoring algorithm can be iteratively weighted least squares (IWLS) algorithm used to get the ML estimators of the gamma regression parameters. In each iteration, the parameters are updated by:

$$\beta^{(r+1)} = \beta^{(r)} + I^{-1}(\beta^{(r)})S(\beta^{(r)}), \quad (5)$$

where $S(\beta) = \partial\ell(\beta)/\partial\beta$ and $I^{-1}(\beta) = \left(-E(\partial^2\ell(\beta)/\partial\beta\partial\beta^T)\right)^{-1}$. The last process of the estimated coefficients is where is a defined as

$$\hat{\beta}_{GR} = (X^T\widehat{W}X)^{-1}X^T\widehat{W}\hat{u}, \quad (6)$$

where $\widehat{W} = \text{diag}(\hat{\theta}_i^2)$ and $\hat{u}$ is a vector where i^{th} element equals to $\hat{u}_i = \hat{\theta}_i + ((y_i - \hat{\theta}_i)/\hat{\theta}_i^2)$.

The ML estimator is asymptotically normally distributed with a covariance matrix that corresponds to the inverse of the Hessian matrix

$$cov(\hat{\beta}_{\text{GR}}) = \left[-E\left(\frac{\partial^2\ell(\beta)}{\partial\beta_i\,\partial\beta_k}\right)\right]^{-1} = \nu^{-1}(X^T\widehat{W}X)^{-1}. \quad (7)$$

4 THE PROPOSED METHOD

The suggested approach Influence diagnostic methods are vital. in GRM to identify observations with disproportionate influence of model estimates, predictions, or overall fit. The hat matrix H is essential in the diagnostics of influences. regression models. GRM Hat matrix is:

$$\text{H} = \widehat{\text{W}}^{1/2}\text{X}(\text{X}^\text{T}\widehat{\text{X}}\text{X})^{-1}\text{X}^\text{T}\widehat{\text{W}}^{1/2}\,. \quad (8)$$

The diagonal elements of the hat matrix are called leverages denoted as hii = diag (H), and hii are the i^{th} diagonal entry of the hat matrix.

Two influence measures in the GRM can be used Cook's Distance (CD) which is measuring the change in the entire parameter estimate vector if a particular observation is deleted. In this, large values indicate influential observations that substantially affect regression coefficients. [2] DFFITS (Difference in Fits) which is measuring the change in the fitted value for an observation when that observation is omitted. The CD and DFFITS for GRM can be defined, respectively, as [24]:

$$CD_i = \frac{1}{r}\acute{\chi}_i{}^2\frac{hii}{1-hii} \;;\, i = 1,2, \ldots, n, \quad (9)$$

$$\text{DFFITS} = |t_i|\sqrt{\frac{hii}{1-hii}}. \quad (10)$$

where $\acute{\chi}_i{}^2 = \frac{\chi_i}{\hat{\phi}(1-hii)}$, is the standardized Pearson residual and *h* is the leverage , $t_i = \acute{\chi}_i\sqrt{\frac{n-r}{n-p-\acute{\chi}_i{}^2}}$ is the jackknife Pearson residual, and $\hat{\phi}$ is the estimated dispersion parameter which is calculated by $\hat{\phi} = (1/(n-p))\sum_{i=1}^{n}\left((y_i - \hat{\theta}_i)^2/\hat{\theta}_i^2\right)$. The largest value of (11) specifies that the i^{th} observation is the influential observation and the DFFITS declared the i^{th} observation as the influential if $DFFITS > 2\sqrt{\frac{r}{n}}$.

The main objective of using meta-heuristic algorithms to perform influence diagnostics in the GRM was the optimization of influential data points detection and evaluation. These algorithms offer a global search power which has the capability of searching complex and nonlinear high dimensionality areas effectively like those experienced in gamma regression impact diagnostics.

Meta-heuristic algorithms can be used to optimize either criteria or objective functions associated with the influence diagnostics measures including Cook s distance, DFFITS, residual or likelihood displacement in gamma regression. The aim is to detect a set of observations or parameter perturbations that influence model fitting or parameter estimates as much as possible. Some of the algorithms such as the secretary bird optimization algorithm propose solutions by iterating the candidate solution to find influential cases. This algorithm is able to search combinations of observations to identify the ones with high influence as well as balanced exploration with go exploitation so as to avoid local optima and replicated search the data space.

Osprey optimization algorithm (OOA) is a novel. metaheuristic algorithm bio-inspired that has the hunting as a source of inspiration. behavior of the osprey bird. Ospreys are said to possess an excellent capability. in locating, fishing, and transporting the fish to an ideal place in order to be consumed. The phenomena are mathematical examples of such and have real-life simulated examples in OOA and used to solve complex optimization problems with a good trade-off between exploration and exploitation phases [25]. The steps of OOA are as follow:

4.1 First: Inspiration

Known as the fish hawk, river hawk, and sea hawk, the osprey is a diurnal, fish-eating bird of prey that is found everywhere in the world. Ospreys are about 50–66 cm long, 0.9–2.1 kg in weight, and have a wing. Almost all the diet of an osprey is fish, which accounts for about 99%, as reported [26]. Ospreys can see underwater targets because of their superior vision. In flights between 10 and 40 meters above water, the osprey locates the position of fish under water. Then it moves towards the fish, dams its feet in the water, and dives under th`e surface to grab the fish [27]. By observing the osprey’s clever ways of catching and transporting fish to its feeding location, we discover cognitive natural mechanisms that have the potential to aid in the formulation of a unique optimization. algorithm. Thus, the mathematical models of these intelligent behaviors in ospreyes inform the guide to them. implementation of the new OOA algorithm.

4.2 Second: Mathematical modeling

With the help of a<|human|>Second: Mathematical modeling: With the assistance of a population-based structure, the OOA enables the search powers of its members to have a convergent point on an. suitable solution by repetition in the space of problems. As an tie of population of OOA, individually each osprey provides values of problem variables based on. the position of its position in the search space. Every osprey is a candidate solution therefore, acts as a candidate solution, which is characterized by a vector that describes its position in the search space. Ospreys as a whole form the OOA population, and their locations can be mathematically represented as a matrix according to (11). The first placement of ospreys in the search space at the beginning of OOA is done through random initialization using (12).

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N\times m} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,m} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,m} \end{bmatrix}_{N\times m} \quad (11)$$

$$x_{i,j} = la_j + r_{i,j} \times (ua_j - la_j), i = 1,2,\ldots,N, j = 1,2,\ldots,m \quad (12)$$

where X is the location of the ospreys is the individual, $X_i$is the ithindividual (candidate solution), $x_{i,j}$is the jthdimension (problem variable) for that osprey, $N$is the number of total of ospreys, $m$is the total number of variables problem, $r_{i,j}$is selected randomly number between 0 & 1, $la_j$and ua_jdenote the lower and upper limits of the jthvariable in the problem.

As the proposed solutions to the problem, the individual ospreys are each of them. Vector representation based on (13) can demonstrate the evaluated results of the objective function for the problem.

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N\times 1} . \quad (13)$$

In which F is the objective function values $F_i$is a vector and is what is received of the objective function. value for the ithosprey. Primary basis of objectives when it comes to the values obtained is the objective function. appraising the quality of the candidate solutions. Therefore, the most favorable outcome of the goal. function is linked with the best candidate solution (the best member) and the worst.

result of the objective function is linked with the worst candidate solution (the worst). member. The positions of the ospreys in the search space are updated every relplcation.

4.3 Exploration

Because of their great eyes, ospreys are able to find fish in the water. The former half of the population. OOA update has been meant to imitate the natural behavior of the Ospreys when hunting. The simulation of an the search for fish by means of ospreys brings significant alterations to the whereabout of the ospreys in the search space, and thus enhancing the capacity of OOA to search the optimal area and not to become trapped in local optima. The ospreys of each osprey of the OOA design determine the locations of the ospreys that have better objective. function underwater fishes. The fish of each osprey is determined by the use of the (14).

$$FP_i = \{X_k | k \in \{1,2,\ldots,N\} \wedge F_k < F_i\} \cup \{X_{best}\} \quad (14)$$

where FP_i is the collection of fish positions for the ith osprey and X_{best} is the best osprey.

From the simulation of the movement of the osprey towards the fish, a new position of the corresponding. osprey is computed using (17). This new position, so long as it adds to the value of the objective. function, takes the place of the old position of the osprey (15).

$$x_{i,j}^{P1} = x_{i,j} + r_{i,j} \times (SF_{i,j} - I_{i,j}.x_{i,j}),$$

$$x_{i,j}^{P1} = \begin{cases} x_{i,j}^{P1}, \; la_j \le x_{i,j}^{P1} \le ua_j \\ la_j, \; x_{i,j}^{P1} \le la_j \\ ua_j, \; x_{i,j}^{P1} > ua_j \end{cases} \quad (15)$$

$$Xi = \begin{cases} X_i^{P1}, & if \; F_i^{P1} < F_i \\ X_i \;, & else \end{cases} \quad (16)$$

Where X_i^{P1} is the new position of the ith osprey in the first stage of OOA, $x_{i,j}^{P1}$is its jth dimension, $F_i^{P1}$is its objective function value and SF_i is the selected fish for ith osprey. The value of $SF_{i,j}$is its jthdimension, while $r_{i,j}$are random numbers between 0 ana 1, $I_{i,j}$ being random numbers from the set {1,2}.

The model for updating the population in the second phase of OOA is formulated by modeling the manner. ospreys behave. The osprey has illustrated in its modelling of the act of transporting the fish to a suitable location. position in the search space is adjusted slightly the search space, which causes the OOA to perform better exploitation in local search and to move closer to better solutions discovered so far.

$$x_{i,j}^{P2} = x_{i,j} + \frac{iaj + r \times (uah - laj)}{t}, \quad (17)$$

$$i = 1,2,\ldots,N, \; j = 1,2,\ldots,m, \; t = 1,\ldots,T$$

$$x_{i,j}^{P2} = \begin{cases} x_{i,j}^{P2}, \; la_j \le x_{i,j}^{P2} \le ua_j \\ la_j, \; x_{i,j}^{P2} < la_j \\ ua_j, \; x_{i,j}^{P2} > ua_j \end{cases}$$

$$Xi = \begin{cases} X_i^{P2}, & if \; F_i^{P2} < F_i \\ X_i \;, & else \end{cases} \quad (18)$$

Where X_i^{P2} is the new position of the ith of the second stage of OOA, $x_{i,j}^{P2}$is its jth dimension, $F_i^{P2}$is objective value function, where $r_{i,j}$are randomly numbers between 0 & 1. The replcation counter is t and Tcorresponds to the iterations total.

4.4 Repetitions Process

The OOA has an iterative methodology, and the initial iteration is done by updating each osprey. position based on the data of the first and second stages. Subsequently, the best the candidate solution is updated on the consideration of which one gives the lowest objective function value. Lastly, once the algorithm has been fully implemented, OOA is the best candidate to choose. solution Solution that is stored in iterations as a solution to the problem.

In OOA, each member is coded as 0 (the training instance is considered as influential) or 1 (the training instance is not considered as influential). A representation of the purpose of OOA is shown in Figure 1.

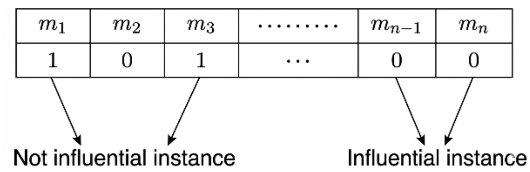

Figure 1: An illustration of purpose OOA.

The proposed approach will efficiently help to find and eliminate the influential instances with less estimation error performance. The parameter configurations for our proposed approach are presented as follows.

1) The number of osprey is set to 25 and number of iterations is t=500_{max}.

2) The stations of each secretary birds are randomly chosen randomly by uniform distribution with [0, 1].
3) The fitness function is defined as (11).
4) The positions are updated with the help of the equations (13)-(20).
5) Steps 3 and 4 are repeated until a t_{max} is reached.

5 SIMULATION RESULTS

In this section, a Monte Carlo simulation is done. testing is carried out to analyse the performance of our proposed method. The response variable of n generated variable of observations is formed $y_i \sim Gamma(\theta_i, \nu)$, where $\nu \in \{0.75, 2\}$ and $\theta_i = exp(x_i^T \beta)$, $\beta = (\beta_1, \ldots, \beta_p)$ with $\sum_{j=1}^{p} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = \ldots = \beta_p$. In addition:

1) Three explanatory variables are included as p = 1,3,7.
2) Four sample size are included as n=30, 50, 100, 200.
3) The explanatory variable are generated as $Xij \sim N(0,1)$ with influential points (IP) as 5^{th}, 10^{th}, 15^{th} , 20^{th} , 25^{th}in the X as $xij = ao + xij$, i=5,10, 15, 20, 25, j=1,2,3,4, where $ao = \bar{x}j + 100$.
4) The simulation is replicated 1000 times to detect the generated influential observations in percentages.

The simulation results in terms of the influential observations detections (in parentage %) for all the combinations are summarized in Tables 1-6. Form these tables, several points are observed:

1) For every combination of dataset size and number of influential points, the SBOA method consistently detects the highest percentage of influential observations. Detection rates with SBOA mostly fall around 90-95%, substantially higher than CD and DFFITS.
2) As the number of influential observations increases (from 5 to 25), detection percentages tend to increase slightly for all methods, indicating improved detection when more influential points are present.
3) Detection percentages generally improve or remain stable as the dataset size grows from 30 to 200, where SBOA maintains high detection rates (above 90% across settings), suggesting good scalability and robustness.
4) Cook's Distance slightly outperforms DFFITS in most scenarios, especially for larger numbers of influential observations, but both lags noticeably behind SBOA.
5) The advantage of SBOA persists consistently across small, medium, and larger sample sizes, showing its effectiveness in influence diagnostics.
6) Across all sample sizes and numbers of influential observations, SBOA detects a slightly higher proportion of influential points when dispersion is 0.75 compared to 2. The difference varies but is consistently positive, typically around 1-2 percentage points higher.
7) Regardless of dispersion, SBOA demonstrates high accuracy in detecting influential observations, with percentages mostly above 89% even in small samples and lower numbers of influential points.
8) Detection rates consistently improve as the number of predictors decreases. When p=1 SBOA yields the highest detection percentages, followed by p=3, and with p=7 having the lowest percentages. This suggests that SBOA performs better in simpler models with fewer predictors, likely because fewer variables reduce complexity and noise, enabling clearer identification of influential points.

Table 1: Influential observations detections (%) for the GRM when p=1 and v=0.75.

n	IP	CD	DFFITS	OOA
30	5	69.74	67.55	89.74
	10	70.38	68.73	90.21
	15	72.84	70.41	90.54
	20	69.58	69.22	89.9
	25	75.14	72.82	91.87
50	5	70.79	68.6	90.79
	10	71.43	69.78	91.26
	15	73.89	71.46	91.59
	20	70.63	70.27	90.95
	25	76.19	73.87	92.92
100	5	70.98	68.79	90.98
	10	71.62	69.97	91.45
	15	74.08	71.65	91.78
	20	70.82	70.46	91.14
	25	76.38	74.06	93.11
200	5	71.71	69.52	91.71
	10	72.35	70.7	92.18
	15	74.81	72.38	92.51
	20	71.55	71.19	91.87
	25	77.11	74.79	93.84

Table 2: Influential observations detections (%) for the GRM when p=1 and v=2.

n	IP	CD	DFFITS	OOA
30	5	68.42	66.23	88.42
	10	69.06	67.41	88.89
	15	71.52	69.09	89.22
	20	68.26	67.9	88.58
	25	73.82	71.5	90.55
50	5	69.47	67.28	89.47
	10	70.11	68.46	89.94
	15	72.57	70.14	90.27
	20	69.31	68.95	89.63
	25	74.87	72.55	91.6
100	5	69.66	67.47	89.66
	10	70.3	68.65	90.13
	15	72.76	70.33	90.46
	20	69.5	69.14	89.82
	25	75.06	72.74	91.79
200	5	70.39	68.2	90.39
	10	71.03	69.38	90.86
	15	73.49	71.06	91.19
	20	70.23	69.87	90.55
	25	75.92	73.04	92.97

Table 3: Influential observations detections (%) for the GRM when p=3 and v=0.75.

n	IP	CD	DFFITS	OOA
30	5	68.42	66.23	88.42
	10	69.06	67.41	88.89
	15	71.52	69.09	89.22
	20	68.26	67.9	88.58
	25	73.82	71.5	90.55
50	5	69.47	67.28	89.47
	10	70.11	68.46	89.94
	15	72.57	70.14	90.27
	20	69.31	68.95	89.63
	25	74.87	72.55	91.6
100	5	69.66	67.47	89.66
	10	70.3	68.65	90.13
	15	72.76	70.33	90.46
	20	69.5	69.14	89.82
	25	75.06	72.74	91.79
200	5	70.39	68.2	90.39
	10	71.03	69.38	90.86
	15	73.49	71.06	91.19
	20	70.23	69.87	90.55
	25	75.79	73.47	92.52

Table 4: Influential observations detections (%) for the GRM when p=3 and v=2.

n	IP	CD	DFFITS	OOA
30	5	67.01	64.82	87.01
	10	67.65	66	87.48
	15	70.11	67.68	87.81
	20	66.85	66.49	87.17
	25	72.41	70.09	89.14
50	5	68.06	65.87	88.06
	10	68.7	67.05	88.53
	15	71.16	68.73	88.86
	20	67.9	67.54	88.22
	25	73.46	71.14	90.19
100	5	68.25	66.06	88.25
	10	68.89	67.24	88.72
	15	71.35	68.92	89.05
	20	68.09	67.73	88.41
	25	73.65	71.33	90.38
200	5	68.98	66.79	88.98
	10	69.62	67.97	89.45
	15	72.08	69.65	89.78
	20	68.82	68.46	89.14
	25	74.51	71.63	91.56

Table 5: Influential observations detections (%) for the GRM when p=7 and v=0.75.

n	IP	CD	DFFITS	OOA
30	5	66.74	64.55	86.74
	10	67.38	65.73	87.21
	15	69.84	67.41	87.54
	20	66.58	66.22	86.9
	25	72.14	69.82	88.87
50	5	67.79	65.6	87.79
	10	68.43	66.78	88.26
	15	70.89	68.46	88.59
	20	67.63	67.27	87.95
	25	73.19	70.87	89.92
100	5	67.98	65.79	87.98
	10	68.62	66.97	88.45
	15	71.08	68.65	88.78
	20	67.82	67.46	88.14
	25	73.38	71.06	90.11
200	5	68.71	66.52	88.71
	10	69.35	67.7	89.18
	15	71.81	69.38	89.51
	20	68.55	68.19	88.87
	25	74.11	71.79	90.84

Table 6: Influential observations detections (%) for the GRM when p=7 and v=2.

N	IP	CD	DFFITS	OOA
30	5	65.3	63.11	85.3
	10	65.94	64.29	85.77
	15	68.4	65.97	86.1
	20	65.14	64.78	85.46
	25	70.7	68.38	87.43
50	5	66.35	64.16	86.35
	10	66.99	65.34	86.82
	15	69.45	67.02	87.15
	20	66.19	65.83	86.51
	25	71.75	69.43	88.48
100	5	66.54	64.35	86.54
	10	67.18	65.53	87.01
	15	69.64	67.21	87.34
	20	66.38	66.02	86.7
	25	71.94	69.62	88.67
200	5	67.27	65.08	87.27
	10	67.91	66.26	87.74
	15	70.37	67.94	88.07
	20	67.11	66.75	87.43
	25	72.8	69.92	89.85

4 CONCLUSIONS

This study of the diagnostic of influence in GRM through the OOA, and the conventional measures, like Cook Distance and DFFITS provided significant information on the effectiveness of the measures of influence on diagnostic in regression. The simulation findings indicate that Cook Distance and DFFITS are still valuable classical statistics to identify influential values, however, that applying OOA in the higher dispersion and small sample size setting is beneficial to realize higher diagnostic accuracy. OOA is adaptive, thus can more economically and more accurately identify (potentially relatively large) influential data points, so it can achieve lower frequencies of false positives and false negatives than classical methods. The informative potential of this comparative analysis is as follows: the OOA has the potential to evolve into a strong optimization-based goodness-of-influence diagnostics measure through ongoing and growing extension of traditional tools and thereby improve reliability and accuracy of the diagnostics of the influence that takes place in the GRM. The research could be undertaken in the future to evaluate how OOA can be implemented in a more complex model of regression and further extend to the modeling and estimation of use.

REFERENCES

[1] E. Nunez, E. W. Steyerberg, and J. Nunez, “Regression modeling strategies,” Revista Española de Cardiología (English Edition), vol. 64, no. 6, pp. 501-507, 2011.

[2] Z. Y. Algamal, “Developing a ridge estimator for the gamma regression model,” Journal of Chemometrics, vol. 32, no. 10, p. e3054, 2018.

[3] E. M. Ortega, V. G. Cancho, and G. A. Paula, “Generalized log-gamma regression models with cure fraction,” Lifetime Data Analysis, vol. 15, no. 1, pp. 79-106, 2009.

[4] Y. Asar and Z. Algamal, “A new two-parameter estimator for the gamma regression model,” Statistics, Optimization & Information Computing, vol. 10, no. 3, pp. 750-761, 2022.

[5] I. Dawoud, “A new improved estimator for the gamma regression model,” Communications in Statistics - Simulation and Computation, pp. 1-12, 2025.

[6] E. Dunder, S. Gumustekin, and M. A. Cengiz, “Variable selection in gamma regression models via artificial bee colony algorithm,” Journal of Applied Statistics, vol. 45, no. 1, pp. 8-16, 2018.

[7] N. A. Al-Thanoon, O. S. Qasim, and Z. Y. Algamal, “Variable selection in Gamma regression model using binary gray Wolf optimization algorithm,” in Journal of Physics: Conference Series, IOP Publishing, 2020.

[8] P. W. Bernhardt, “Model validation and influence diagnostics for regression models with missing covariates,” Statistics in Medicine, vol. 37, no. 8, pp. 1325-1342, 2018.

[9] A. Tapia, et al., “Influence diagnostics in mixed effects logistic regression models,” Test, vol. 28, no. 3, pp. 920-942, 2019.

[10] S. Liu and A. H. Welsh, “Regression diagnostics,” in International Encyclopedia of Statistical Science, Springer, 2025, pp. 2151-2156.

[11] B. Rajaratnam, et al., “Influence diagnostics for high-dimensional lasso regression,” Journal of Computational and Graphical Statistics, vol. 28, no. 4, pp. 877-890, 2019.

[12] J. V. Oliveira Jr, F. Cribari-Neto, and J. S. Nobre, “Influence diagnostics and model validation for the generalized extreme-value nonlinear regression model,” Journal of Statistical Computation and Simulation, vol. 90, no. 3, pp. 515-549, 2020.

[13] J. N. da Cruz, E. M. Ortega, and G. M. Cordeiro, “The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis,” Journal of Statistical Computation and Simulation, vol. 86, no. 8, pp. 1516-1538, 2016.

[14] Z. Y. Algamal, “Diagnostic in poisson regression models,” Electronic Journal of Applied Statistical Analysis, vol. 5, no. 2, pp. 178-186, 2012.

[15] M. Amin, M. Amanullah, and G. M. Cordeiro, “Influence diagnostics in the gamma regression model with adjusted deviance residuals,” Communications in Statistics - Simulation and Computation, vol. 46, no. 9, pp. 6959-6973, 2017.

[16] N. A. Al-Thanoon, Z. Y. Algamal, and O. S. Qasim, “Feature selection based on a crow search algorithm for big data classification,” Chemometrics and Intelligent Laboratory Systems, vol. 212, 2021.

[17] A. J. Lemonte and A. G. Patriota, “Influence diagnostics in Birnbaum-Saunders nonlinear regression models,” Journal of Applied Statistics, vol. 38, no. 5, pp. 871-884, 2011.

[18] 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, 2019.

[19] A. Khan, M. Amanullah, M. Amin, R. Alharbi, A. H. Muse, and M. S. Mohamed, “Influence diagnostic methods in the Poisson regression model with the Liu estimator,” Computational Intelligence and Neuroscience, vol. 2021, no. 1, p. 4407328, 2021.

[20] N. Saleem, A. Akbar, S. Laeeq, and S. Ahmad, “Influential observations detection in the gamma-pareto regression model under different link functions with standardized and adjusted deviance residuals: simulation and application,” Sigma: Journal of Engineering and Natural Sciences, vol. 43, no. 3, 2025.

[21] M. Amin, M. Amanullah, and G. M. Cordeiro, “Influence diagnostics in the gamma regression model with adjusted deviance residuals,” Communications in Statistics - Simulation and Computation, vol. 46, no. 9, pp. 6959-6973, 2017.

[22] A. Zakaria, N. K. Howard, and B. K. Nkansah, “On the detection of influential outliers in linear regression analysis,” 2014.

[23] E. Uusipaikka, Confidence intervals in generalized regression models. Chapman & Hall/CRC Press, 2009.

[24] M. Amin, M. Amanullah, and G. M. Cordeiro, “Influence diagnostics in the Gamma regression model with adjusted deviance residuals,” Communications in Statistics - Simulation and Computation, vol. 46, no. 9, pp. 6959-6973, 2017.

[25] M. Dehghani and P. Trojovský, “Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems,” Frontiers in Mechanical Engineering, vol. 8, p. 1126450, 2023.

[26] R. A. Grove, C. J. Henny, and J. L. Kaiser, “Osprey: worldwide sentinel species for assessing and monitoring environmental contamination in rivers, lakes, reservoirs, and estuaries,” Journal of Toxicology and Environmental Health, Part B, vol. 12, no. 1, pp. 25-44, 2009.

[27] A. F. Poole, R. O. Bierregaard, and M. S. Martell, Osprey: Pandion Haliaetus. Birds of North America, Incorporated, 2002.