STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, December 2025, pp 3299–3309.
Published online in International Academic Press (www.IAPress.org)

Influence Diagnostics in Gamma Regression Model Using Secretary Bird Optimization Algorithm

Luay Adil Abduljabbar*, Sabah Manfi Ridha

College of Administration and Economy, University of Baghdad, Baghdad, Iraq

Abstract The diagnosing of influence is essential to assist in detecting influential observations, which influences the inference, especially estimation of the model. Classical diagnostic tools like Cooks Distance and DFFITS are well established and it is possible that less appropriate in model complexities or under different dispersion conditions of data. In the current paper, a novel effort to advance the area of influence diagnostics to the Gamma regression models (GRM) is proposed to utilize the metaheuristic approach as called the Secretary Bird Optimization Algorithm (SBOA). To compare the GRM detection ability of TC and MRE of Cook s Distance and DFFITS and the SBOA based SMOX approach, we run an extensive simulation study across sample sizes and dispersion parameters. The results of the simulations prove that the Cook Distance and the DFFITS are reliable but SBOA-ameliorated diagnostic scheme perform better to detect influential cases particularly in high dispersion scenarios and a limited to moderate samples. Viewed through compared analysis, it can be said SBOA offers a more thorough detection mechanism.

Keywords Cooks distance, DFFITS; gamma regression model, secretary bird optimization algorithm, influential observations

AMS 2010 subject classifications 62J12, 62J20

DOI: 10.19139/soic-2310-5070-2899

1. Introduction

Regression is one of the statistical techniques which identifies the connection flawlessly involving the response variable and the explanation one. Regression assists us to interpret and analyze how one or more predictors affect the response variable. Its uses are numerous in the commercial, industrial, medical, etc. field. The regression analysis can assist in the predictive and forecasting analysis of the variables of interest [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, 3, 4, 21, 22, 23]. 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, 24, 25]. Under this model, the dependent variable is considered to be 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, 6, 7, 26, 27, 28].

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,

*Correspondence to: Luay Adil Abduljabbar (Email: luay.abd2201@coadec.uobaghdad.edu.iq).

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

standard errors, predictions and statistical inferences and may result into misleading findings, unless identified and correctly addressed [10, 29, 30].

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, 12, 13, 31]. 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, 32, 33]. 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, 34, 35].

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, 36, 37]. Although they seek to locate a passably decent answer in a reasonable period of time, 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 [38, 39, 40].

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. Secretary bird optimization algorithm is employed as an influence diagnostic in GRM.

2. Gamma regression model

Positively skewed data often arise in epidemiology, social, and economic studies. This type of data consists of nonnegative values. Gamma distribution is a well-known distribution that fits to such type of data. GRM is used to model the relationship between the positively skewed response variable and potentially regressors [17, 41, 42, 43, 44, 45].

Let y_i be the response variable and follows a gamma distribution with nonnegative shape parameter ν and nonnegative scale parameter γ, i.e. $y_i \sim Gamma(\nu, \gamma)$, then the probability density function is defined as

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

with $E(y) = \nu/\gamma = \theta$ and $var(y) = \nu/\gamma^2 = \theta^2/\nu$. Given that $\gamma = \nu/\theta$, Eq. (3) can re-parameterized as a function of the mean (θ) and the shape (ν) parameters and written depending on the exponential function as

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

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

Gamma regression model is usually modeled using the canonical link function (reciprocal), $\theta_i = -1/\mathbf{x}_i^T\boldsymbol{\beta}$ which is expressed as a linear combination of covariates$\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$. The log link function, $\theta_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$, is alternatively used rather than the reciprocal link function because it ensures that $\theta_i > 0$.

The most common method of estimating the coefficients of GRM is to use the maximum likelihood method of Eq. (4). Given the assumption that the observations are independent and $\theta_i = -1/\mathbf{x}_i^T\boldsymbol{\beta}$, the log-likelihood function is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\frac{y_i\mathbf{x}_i^T\boldsymbol{\beta} - \log(\mathbf{x}_i^T\boldsymbol{\beta})}{1/\nu} + c(y_i, \nu)\right\}, \tag{3}$$

the ML estimator is then obtained by computing the first derivative of the Eq. (5) and setting it equal to zero, as

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{1}{\nu}\sum_{i=1}^{n}\left[y_i - \frac{1}{\mathbf{x}_i^T\boldsymbol{\beta}}\right]\mathbf{x}_i = 0. \tag{4}$$

Unfortunately, the first derivative cannot be solved analytically because Eq. (6) is nonlinear in $\boldsymbol{\beta}$. The iteratively weighted least squares (IWLS) algorithm or Fisher-scoring algorithm can be used to obtain the ML estimators of the gamma regression parameters. In each iteration, the parameters are updated by

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

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

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

where $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\theta}_i^2)$ and $\hat{\mathbf{u}}$ is a vector where ith 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

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

3. The proposed method

Influence diagnostic methods are critical in GRM for detecting observations that have a disproportionate impact on model estimates, predictions, or overall fit. The hat matrix, H, is fundamental in influence diagnostics for regression models. The hat matrix for GRM is

$$H = \hat{W}^{1/2}X(X^T\hat{X}X)^{-1}X^T\hat{W}^{1/2} \tag{8}$$

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: (1) 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 [18]:

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

$$\mathrm{DFFITS} = |t_i|\sqrt{\frac{hii}{1-hii}} \tag{10}$$

where $\acute{\chi}_i^2 = \frac{\chi_i}{\widehat{\phi}(1-hii)}$, 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 Eq. (9) specifies that the ith 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 exploitation in order to evade local optima and exhaustively traverse the data space.

Secretary Bird Optimization Algorithm (SBOA) is new population-based metaheuristic algorithm which is developed based on the survival behavior of secretary bird in its natural habitat [19]. The algorithm is used to adapt two primary survival modes of these birds that include hunting snakes (exploration) and predator evasion (exploitation) to drive exploration and exploitation successfully to optimize complex problems. Secretary birds survive by persistently hunting snakes and escaping predators. This dual behavior forms the basis of SBOA's two-phase search mechanism: (1) Exploration Phase (Hunting Behavior): Models the bird's search for prey, encouraging global search and diversity. (2) Exploitation Phase (Escaping Behavior): Models the bird's evasion of predators, focusing on refining and exploiting promising solutions.

In the initial implementation of the SBOA, the random position initialization of Secretary Birds in the search space as

$$X_{ij} = lb_j + r \times (ub_j - lb_j), \quad i = 1, 2, ..., N, j = 1, 2, ..., Dim \tag{11}$$

The position of the i^{th}secretary bird $X_i$is determined by the random number $r$between 0 and 1 while lb_jandub_j represent the lower and upper bounds. Each secretary bird provides values for problem variables which allows evaluation of the objective function. The objective function values are combined into a vector after the evaluation process.

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N \times 1} \tag{12}$$

The vector Fcontains objective function values while $F_i$represents the objective function value achieved by the i^{th} secretary bird

The strategy of hunting by the secretary birds as a mode of trying to get to the snakes as outlined in the exploration Phase has three main phases namely; the searching phase, the feeding phase as well as the attacking phase. Biological statistical data of the hunting stages of the secretary bird and the time spent on the hunting process have been considered to divide the process into three equal shares. These intervals correspond to searching for prey $t < \frac{1}{3}T$, consuming prey $\frac{1}{3}T < t < \frac{2}{3}T$ and attacking prey $\frac{2}{3}T < t < T$.

The mathematical modeling of secretary bird position updates during the Searching for Prey stage is

$$While\ t < \frac{1}{3}T, x_{i,j}^{newP1} = x_{i,j} + (x_{random\_1} - x_{random\_2}) \times R_1 \tag{13}$$

$$Xi = \begin{cases} X_i^{new,P1}, & if\ F_i^{new,P1} < F_i \\ X_i\ , & else \end{cases} \tag{14}$$

When a secretary bird locates a snake it follows a unique hunting pattern. The secretary bird follows a different hunting pattern than other raptors because it uses its quick foot movements to circle the snake [20]. The mathematical model for updating the secretary bird's position in the Consuming Prey stage is [19].

$$RB = randn(1, Dim) \tag{15}$$

$$While\ \ \frac{1}{3}T < t < \frac{2}{3}T,\ x_{i,j}^{newP1} = x_{best} + \exp((t/T) \wedge 4) \times (RB - 0.5) \times (x_{best} - x_{i,j}) \tag{16}$$

$$Xi = \begin{cases} X_i^{new,P1}, & if\ F_i^{new,P1} < F_i \\ X_i\ , & else \end{cases} \tag{17}$$

where $randn(1, Dim)$represents a randomly generated array of dimension $1 \times Dim$from a standard normal distribution (mean = 0, standard deviation =1), and x_{best}represents the current best value.

The mathematical modeling for updating the secretary bird's position in the Attacking Prey stage is

$$While\ \ t > \frac{2}{3}T,\ x_{i,j}^{newP1} = x_{best} + ((1 - t/T) \wedge (2 \times t/T)) \times x_{i,j} \times RL \tag{18}$$

$$Xi = \begin{cases} X_i^{new,P1}, & if\ F_i^{new,P1} < F_i \\ X_i\ , & else \end{cases} \tag{19}$$

The weighted Levy flight named RL serves to improve the optimization accuracy of the algorithm.

$$RL = 0.5 \times Levy(Dim) \tag{20}$$

The$Levy(Dim)$function represents the Levy flight distribution function in this context. It is calculated as follows:

$$Levy(D) = s \times \frac{u \times \sigma}{|v|^{\frac{1}{\eta}}} \tag{21}$$

The constant $s$maintains a value of 0.01 while η maintains a value of 1.5. Both $u$and $v$represent random numbers within the [0, 1] interval. The mathematical expression for σappears below

$$\sigma = \left(\frac{\Gamma(1+\eta) \times \sin(\frac{\pi\eta}{2})}{\Gamma(\frac{1+\eta}{2}) \times \eta \times 2(\frac{\eta-1}{2})} \right)^{\frac{1}{\eta}} \tag{22}$$

The gamma functionΓ appears in this expression with ηset to 1.5.

In the exploitation stage, secretary birds use different avoidance techniques to defend themselves and their food sources when they encounter predators. The SBOA design assumes that either of these two conditions will happen with equal likelihood:

1. C_1 refer to camouflage by environment.
2. C_2 refer to fly or run away.

The initial response of secretary birds to predator detection involves searching for appropriate camouflage areas. The birds will choose to flee or run quickly when they cannot locate a suitable and safe camouflage area. We introduce a dynamic perturbation factor denoted as $(1 - t/T)^2$ in this context. The changing parameter element allows the algorithm to maneuver effectively between finding new solutions (exploration) and relying on existing solutions (exploitation). The adjustment of these factors enables users to boost exploration intensity or maximize exploitation effectiveness during different stages of the process. The mathematical modeling of secretary birds' evasion strategies through Eq. (23) leads to an updated condition which can be expressed through Eq. (24).

$$x_{i,j}^{new,P2} = \begin{cases} C_1 :\ x_{best} + (2 \times RB - 1) \times (1 - \frac{t}{T})^2 \times x_{i,j},\ if\ r\ and < r_i \\ C2 : x_{i,j} + R_2 \times (x_{random} - K \times x_{i,j}), \qquad else \end{cases} \tag{23}$$

$$Xi = \begin{cases} X_i^{new,P2}, & if\ F_i^{new,P2} < F_i \\ X_i\ , & else \end{cases} \tag{24}$$

The calculation involves $r = 0.5$and $R_2$for random array generation from normal distribution and $x_{random}$for random candidate solution and $K$for random integer selection through Eq. (25).

$$K = round(1 + rand(1, 1)) \tag{25}$$

In SBOA, 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 SBOA is shown in Figure 1.

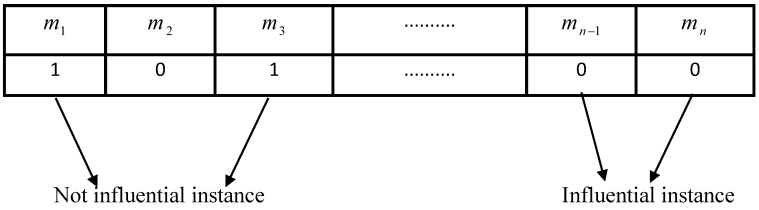

Figure 1. An illustration of purpose SBOA.

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 secretary birds is set to 30 and the number of iterations is $\mathrm{t_{max}}= 1000$.
2. The positions of each secretary birds are randomly determined by uniform distribution with the range [0, 1].
3. The fitness function is defined as Eq. (9).
4. The positions are updated using Eqs. (13)-(23).
5. Steps 3 and 4 are repeated until a $t_{\max}$ is reached.

4. Simulation results

In this section, a Monte Carlo simulation experiment is used to examine the performance of our proposed method. The response variable of n observations from GRM is generated as $y_i \sim Gamma(\theta_i, \nu)$, where $\nu \in \{0.75, 2\}$ and $\theta_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ with $\sum_{j=1}^{p} \beta_j^2 = 1$ and $\beta_1 = \beta_2 = ... = \beta_p$. 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 $ao\ =\ \overline{x}j\ + 100$.
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.

Table 1. Influential observations detections (%) for the GRM when p=1 and $\nu = 0.75$

n	IP	CD	DFFITS	SBOA
30	5	70.58	68.39	**90.58**
	10	71.22	69.57	**91.05**
	15	73.68	71.25	**91.38**
	20	70.42	70.06	**90.74**
	25	75.98	73.66	**92.71**
50	5	71.63	69.44	**91.63**
	10	72.27	70.62	**92.1**
	15	74.73	72.3	**92.43**
	20	71.47	71.11	**91.79**
	25	77.03	74.71	**93.76**
100	5	71.82	69.63	**91.82**
	10	72.46	70.81	**92.29**
	15	74.92	72.49	**92.62**
	20	71.66	71.3	**91.98**
	25	77.22	74.9	**93.95**
200	5	72.55	70.36	**92.55**
	10	73.19	71.54	**93.02**
	15	75.65	73.22	**93.35**
	20	72.39	72.03	**92.71**
	25	77.95	75.63	**94.68**

Table 2. Influential observations detections (%) for the GRM when p=1 and $\nu = 2$

n	IP	CD	DFFITS	SBOA
30	5	69.26	67.07	**89.26**
	10	69.9	68.25	**89.73**
	15	72.36	69.93	**90.06**
	20	69.1	68.74	**89.42**
	25	74.66	72.34	**91.39**
50	5	70.31	68.12	**90.31**
	10	70.95	69.3	**90.78**
	15	73.41	70.98	**91.11**
	20	70.15	69.79	**90.47**
	25	75.71	73.39	**92.44**
100	5	70.5	68.31	**90.5**
	10	71.14	69.49	**90.97**
	15	73.6	71.17	**91.3**
	20	70.34	69.98	**90.66**
	25	75.9	73.58	**92.63**
200	5	71.23	69.04	**91.23**
	10	71.87	70.22	**91.7**
	15	74.33	71.9	**92.03**
	20	71.07	70.71	**91.39**
	25	76.76	73.88	**93.81**

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

Table 3. Influential observations detections (%) for the GRM when p=3 and $\nu = 0.75$

n	IP	CD	DFFITS	SBOA
30	5	69.26	67.07	**89.26**
	10	69.9	68.25	**89.73**
	15	72.36	69.93	**90.06**
	20	69.1	68.74	**89.42**
	25	74.66	72.34	**91.39**
50	5	70.31	68.12	**90.31**
	10	70.95	69.3	**90.78**
	15	73.41	70.98	**91.11**
	20	70.15	69.79	**90.47**
	25	75.71	73.39	**92.44**
100	5	70.5	68.31	**90.5**
	10	71.14	69.49	**90.97**
	15	73.6	71.17	**91.3**
	20	70.34	69.98	**90.66**
	25	75.9	73.58	**92.63**
200	5	71.23	69.04	**91.23**
	10	71.87	70.22	**91.7**
	15	74.33	71.9	**92.03**
	20	71.07	70.71	**91.39**
	25	76.63	74.31	**93.36**

Table 4. Influential observations detections (%) for the GRM when p=3 and $\nu = 2$

n	IP	CD	DFFITS	SBOA
30	5	67.85	65.66	**87.85**
	10	68.49	66.84	**88.32**
	15	70.95	68.52	**88.65**
	20	67.69	67.33	**88.01**
	25	73.25	70.93	**89.98**
50	5	68.9	66.71	**88.9**
	10	69.54	67.89	**89.37**
	15	72	69.57	**89.7**
	20	68.74	68.38	**89.06**
	25	74.3	71.98	**91.03**
100	5	69.09	66.9	**89.09**
	10	69.73	68.08	**89.56**
	15	72.19	69.76	**89.89**
	20	68.93	68.57	**89.25**
	25	74.49	72.17	**91.22**
200	5	69.82	67.63	**89.82**
	10	70.46	68.81	**90.29**
	15	72.92	70.49	**90.62**
	20	69.66	69.3	**89.98**
	25	75.35	72.47	**92.4**

that SBOA performs better in simpler models with fewer predictors, likely because fewer variables reduce complexity and noise, enabling clearer identification of influential points.

5. Conclusion

The research on the diagnostic of influence in GRM using SBOA, as well as conventional measures, such as Cook Distance and DFFITS revealed considerable information about the effectiveness of diagnostic measures of influence in regression. The simulation results show that both Cook Distance and DFFITS remain useful classical statistics to detect influential observations, but that using SBOA within the higher dispersion and small sample size scenarios is useful to achieve greater diagnostic precision. Because SBOA is adaptive, it is able to detect

Table 5. Influential observations detections (%) for the GRM when p=7 and $\nu = 0.75$

n	IP	CD	DFFITS	SBOA
30	5	67.58	65.39	**87.58**
	10	68.22	66.57	**88.05**
	15	70.68	68.25	**88.38**
	20	67.42	67.06	**87.74**
	25	72.98	70.66	**89.71**
50	5	68.63	66.44	**88.63**
	10	69.27	67.62	**89.1**
	15	71.73	69.3	**89.43**
	20	68.47	68.11	**88.79**
	25	74.03	71.71	**90.76**
100	5	68.82	66.63	**88.82**
	10	69.46	67.81	**89.29**
	15	71.92	69.49	**89.62**
	20	68.66	68.3	**88.98**
	25	74.22	71.9	**90.95**
200	5	69.55	67.36	**89.55**
	10	70.19	68.54	**90.02**
	15	72.65	70.22	**90.35**
	20	69.39	69.03	**89.71**
	25	74.95	72.63	**91.68**

Table 6. Influential observations detections (%) for the GRM when p=7 and $\nu = 2$

n	IP	CD	DFFITS	SBOA
30	5	66.14	63.95	**86.14**
	10	66.78	65.13	**86.61**
	15	69.24	66.81	**86.94**
	20	65.98	65.62	**86.3**
	25	71.54	69.22	**88.27**
50	5	67.19	65	**87.19**
	10	67.83	66.18	**87.66**
	15	70.29	67.86	**87.99**
	20	67.03	66.67	**87.35**
	25	72.59	70.27	**89.32**
100	5	67.38	65.19	**87.38**
	10	68.02	66.37	**87.85**
	15	70.48	68.05	**88.18**
	20	67.22	66.86	**87.54**
	25	72.78	70.46	**89.51**
200	5	68.11	65.92	**88.11**
	10	68.75	67.1	**88.58**
	15	71.21	68.78	**88.91**
	20	67.95	67.59	**88.27**
	25	73.64	70.76	**90.69**

(potentially relatively large) influential data points more efficiently and precisely, which makes false positives and negatives less likely than with classical techniques. The informative value of this comparative analysis is in the ability of SBOA to develop into a powerful optimization-based measure of goodness of influence diagnostics as it advances and expands traditional methods, thus enhancing reliability and accuracy of the diagnostics of the influence occurring in GRM. In future, the investigation can be conducted to examine how SBOA can apply to another complex model of regression and expand its use in modeling and estimation.

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*, 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, pp. 2151–2156, 2025.
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. E. Uusipaikka, *Confidence intervals in generalized regression models*, Chapman & Hall/CRC Press, NW, 2009.
18. 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.
19. Y. Fu, et al., *Secretary bird optimization algorithm: a new metaheuristic for solving global optimization problems*, Artificial Intelligence Review, vol. 57, no. 5, p. 123, 2024.
20. S.D. Hofmeyr, C.T. Symes, and L.G. Underhill, *Secretarybird Sagittarius serpentarius population trends and ecology: insights from South African citizen science data*, PLoS One, vol. 9, no. 5, p. e96772, 2014.
21. Algamal, Z. Y., & Asar, Y. *Liu-type estimator for the gamma regression model* , Communications in Statistics-Simulation and Computation, vol. 49, no . 8, p. 2035-2048, 2020.
22. Algamal, Z. Y., & Lee, M. H. *A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives* , SAR and QSAR in Environmental Research, vol. 28, no. 1 , p. 75-90, 2017.
23. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. *Improving whale optimization algorithm for feature selection with a time-varying transfer function* , Numerical Algebra, Control and Optimization, vol. 11, no. 1 , p. 87-98 , 2020.
24. Algamal, Z. Y., Qasim, M. K., & Ali, H. T. M. *A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine* , SAR and QSAR in Environmental Research, vol. 28 , no. 5, p. 415-426 , 2017.
25. Algamal, Z. Y., Lee, M. H., & Al-Fakih, A. M. *High-dimensional quantitative structure–activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression* , Journal of Chemometrics, vol. 30 , no.2 m p. 50-57 , 2016.
26. Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. *High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty* , Journal of Chemometrics, vol. 31 , p.6, p. e2889 , 2017.
27. Algamal, Z. Y., Qasim, M. K., Lee, M. H., & Ali, H. T. M. *High-dimensional QSAR/QSPR classification modeling based on improving pigeon optimization algorithm* , Chemometrics and Intelligent Laboratory Systems, vol. 206, p. 104170 , 2020.
28. Ismael, O. M., Qasim, O. S., & Algamal, Z. Y. *Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning* , Journal of Chemometrics, vol 34 , no. 11, e3311, 2020.
29. Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., & Taha, I. M. *A new two-parameter estimator for beta regression model: method, simulation, and application* , Frontiers in Applied Mathematics and Statistics, vol. 7,p. 780322 ,2022.
30. Algamal, Z. Y., & Abonazel, M. R. *Developing a Liu-type estimator in beta regression model* , Concurrency and Computation: Practice and Experience, vol.34 ,no. 5 , p. e6685.
31. Algamal, Z., & Ali, H. M. *An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression* , Electronic Journal of Applied Statistical Analysis, vol . 10 , no . 1 , 242-256 , 2017.

32. Salih, A. M., Algamal, Z., & Khaleel, M. A. *A new ridge-type estimator for the gamma regression model* , Iraqi Journal for Computer Science and Mathematics, vol . 5 , no .1 , p. 85-98.
33. Alkhateeb, A., & Algamal, Z. *). Jackknifed Liu-type estimator in Poisson regression model* , Journal of the Iranian Statistical Society,Vol .11 ,no . 1, p. 21-37, 2022.
34. Mahmood, S. W., Basheer, G. T., & Algamal, Z. Y. *Quantitative Structure–Activity Relationship Modeling Based on Improving Kernel Ridge Regression* , Journal of Chemometrics, vol. 39, no. 5, p. e70027, 2025.
35. Mahmood, S. W., Basheer, G. T., & Algamal, Z. Y. *Improving kernel ridge regression for medical data classification based on meta-heuristic algorithms* , Kuwait Journal of Science, vol. 52, no. 3 , p. 100408 , 2025
36. Algamal, Z. Y. *Shrinkage parameter selection via modified cross-validation approach for ridge regression model*, Communications in Statistics-Simulation and Computation, vol. 49, no. 7 , p. 1922-1930, 2020
37. Algamal, Z. Y., Alhamzawi, R., & Ali, H. T. M. *Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression*, Computers in biology and medicine, vol. 97, p. 145-152, 2018
38. Algamal, Z. Y., & Lee, M. H. *A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression*, Journal of Chemometrics, vol. 31, no.(10), e2915, 2017
39. Qasim, M. K., Algamal, Z. Y., & Ali, H. M. *A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine*, SAR and QSAR in Environmental Research, vol. 29, no.(7), p.517-527, 2018
40. Algamal, Z. Y., & Lee, M. H. *Applying penalized binary logistic regression with correlation based elastic net for variables selection*, Journal of Modern Applied Statistical Methods, vol. 14, no.(1), p.15, 2015
41. Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. *High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm*, SAR and QSAR in Environmental Research, vol. 27, no.(9), p.703-719, 2016
42. Al-Taweel, Y., & Algamal, Z. *Some almost unbiased ridge regression estimators for the zero-inflated negative binomial regression model*, Periodicals of Engineering and Natural Sciences, vol. 8, no.(1), p.248-255, 2020
43. Ewees, A. A., Algamal, Z. Y., Abualigah, L., Al-Qaness, M. A., Yousri, D., Ghoniem, R. M., & Abd Elaziz, M. *A cox proportional-hazards model based on an improved aquila optimizer with whale optimization algorithm operators*, Mathematics, vol. 10, no.(8), p.1273, 2022
44. Shamany, R., Alobaidi, N. N., & Algamal, Z. Y. *A new two-parameter estimator for the inverse Gaussian regression model with application in chemometrics*, Electronic Journal of Applied Statistical Analysis, vol. 12, no.(2), p.453-464, 2019
45. Awwad, F. A., Odeniyi, K. A., Dawoud, I., Algamal, Z. Y., Abonazel, M. R., Kibria, B. G., & Eldin, E. T. *New two-parameter estimators for the logistic regression model with multicollinearity*, WSEAS Trans. Math, vol. 21, p.403-414, 2022