Mathematics

AIMS Mathematics, 9(2): 4057–4075.
DOI: 10.3934/math.2024198
Received: 15 August 2023
Revised: 25 December 2023
Accepted: 08 January 2024
Published: 12 January 2024

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

Research article

Diagnostic power of some graphical methods in geometric regression model addressing cervical cancer data

Zawar Hussain[1,*], Atif Akbar[2], Mohammed M. A. Almazah[3], A. Y. Al-Rezami[4] and Fuad S. Al-Duais[4]

[1] Govt Millat Graduate College Multan, Pakistan, 60800
[2] Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan, 60800
[3] Department of Mathematics, College of Sciences and Arts (Muhyil), King Khalid University, Muhyil, 61421, Saudi Arabia
[4] Mathematics Department, College of Humanities and Science, Prince Sattam Bin Abdulaziz University, Al-Kharj, 16278, Saudi Arabia

*** Correspondence:** Email: zawar.hussain55@yahoo.com.

Abstract: In the framework of generalized linear models (GLM), this paper explores the design and applicability of partial residual (PRES), augmented partial residual (APRES), and conditional expectation and residuals (CERES) plots for visualizing an outlier's diagnostics as a function of selected variables. Here, a geometric regression as a GLM is thoroughly described. Additionally, plots for PRES, APRES, and CERES have been built. Due to how the response variable and the associated link function interact with various covariates, the effectiveness of these plots for creating an appealing visual impression may vary. On the cervical cancer data, specific methodologies are used to identify trends for effective modelling. When compared to other approaches, the power of the tests for various plots demonstrates that PRES, CERES (L) and CERES (K) have the greatest endurance for the outlier's diagnostics. On the basis of the power of residual plots, the use is recommended for outlier diagnostics in presence of conventional tests.

Keywords: visual impression; geometric regression; diagnostics; predictors transformations; cervical cancer
Mathematics Subject Classification: 00A71

1. Introduction

It is important to emphasize the applications of statistical methods in both the natural sciences and other fields. A regression analysis is one of the most well-known strategies for measuring relationships between various elements in handling relatively complicated relationships that are frequently observed in the social sciences, health and life sciences, polling, bioscience, analytical and biochemistry, economics and finance, etc. Exponential family-based responses are used in generalized linear models (GLM), a particular type of regression model. Similar to different regression analysis theories, this work encircles specific assumptions that must be addressed before it can be used. These include an outlier's homoscedasticity and multicollinearity [1–4].

A statistical examination of data is used to practically explain all phenomena in everyday life. The analysis and visual assessment of the relevant data is aided by statistical graphics. They are typically used to establish relationships between combined data sets for various variables, to improve comparisons, to refine models, to store and retrieve data summaries, to report results effectively, and to simplify complex graphics information more effectively. The strength of statistical graphics comes from their ability to swiftly and effectively suggest enormous amounts of information, thus allowing people to immediately understand concepts that would not be apparent from a list of values [5–7].

Different plots are used to identify specific issues with the fitted model; partial residual plots (PRES) are among the most helpful plots [8]. In addition to the typical residual plot, partial residual plots are employed for diagnostic purposes in multiple regression analyses [9]. A useful way to identify an outlier and significant observations, curvature, and many other issues brought on by non-random data patterns is through partial residual plots [10].

According to Davison and Tsai's explanation in [11], PRES plots are employed when a proper modification is required while utilizing the nonlinear regression model. The partial residual graphs for weighted regression models were created by Hines and Carter [12]. In an effort to further the work of Larsen and McCleary [8], Wang [13] created a variable plot, added a PRES plot, and built a variable plot. Then, the multiple regression model was used to modify these plots for regression diagnostics; this study was expanded to include nonlinearity detection in the generalized linear model [14].

As shown by Almazah et al. [15], in a geometric regression, which is a generalization of the Poisson regression, the constraint that the mean is equal to the variance provided by the Poisson model is relaxed. A geometric regression is defined as a negative binomial regression with the dispersion parameter set to one. When looking at the traditional geometric, Poisson, and negative binomial regression models for count data, Makcutek [16] showed how the traditional models were extended by the obstacle and zero-inflated models.

The GLM was created by Jahan et al. [17] for the geometric distribution. The natural link function was used for one of the generalized regression models, and the log link function was used for the other. They carried out parameter estimates along with testing procedures. Additionally, they compared the outcomes of these regression models and concluded that the log-link, function-fitted model had a smaller Akaike's information criterion and deviance than the natural-link, function-fitted regression model. They discovered that the log-link function used in GLM for the geometric distribution produced better results than the natural link function.

The model for count data was created by utilizing the geometric and Poisson distribution, Pradhan and Kundu [18] created a model for count data. They employed frequentist and Bayesian criteria to choose these distributions. They studied whether the Bayesian technique is the best criterion to choose

a model and is pretty even for a modest sample size. Moreover, they employed a maximum likelihood estimate for the purpose of discrimination.

The geometric distribution, which Al-Balushi and Islam [19] addressed, is a member of the family of discrete distributions and is concerned with the quantity of trails required in any case to either succeed or occur for the first time. However, the GLM's application to the geometric distribution received little consideration. In their study, an attempt was made to model the data from the count using a geometric regression. It was shown that the geometric generalized regression model was appropriate for analyzing discrete data on the first prenatal visit's timing that showed under-dispersion, and the outcomes contrasted those of the negative binomial and Poisson regression model. They concluded that count data sets with potential over- or under-dispersions could be accommodated by the geometric regression model, which is a flexible approach, as well as the possibility of using the model for modelling count data as an alternative to the prevalent Poisson and negative binomial regression models.

PRES plots were frequently employed by Saulnier et al. [20] to assess the correlation between urine metabolites and structural lesions in diabetics; they concluded that low concentrations of a specific group of urine metabolites were related to kidney structural lesions in individuals with type 2 diabetes. These were utilized by Wouters et al. [22] in the South Western Atlantic to study the spatial disparity of the functional trait diversity of polychaete assemblages throughout a broad latitudinal gradient. Xie et al. [21] employed PRES plots to determine several metrics for plant communities and the functional characteristics of dominant plant species for various ecosystem services in green roofs.

PRES plots for generalized linear models under canonical relations were proposed by Landwehr et al. [23] to aid in the visualization of unknown functions and to determine the need for the transformation of regression predictors with a binary response. Cook and Croos-Dabrera [24] examined the link function and stochastic behavior of the predictors in the class of GLMs, as well as the use of PRES plots to display the perception of curvature in binary logistic regression as a function of predictors. In their inverse Gaussian regression model, Imran and Akbar [25] used PRES plots for regression diagnostics. Recently, a pattern for binomial regression was devised using data from chemical species plots treated with hindered internal rotation (HIR). This pattern included residual (RES), PRES, augmented partial residual (APRES), conditional expectation and residuals for kernel functions (CERES (K)), and conditional expectation and residuals for LOESS (CERES (L)), which were observed by Hussain and Akbar [26] by estimating the test’s power.

The construction of PRES, APRES, CERES (L), and CERES (K) for geometric regression models are all addressed in the current study. Using these residuals, plots are created for regression diagnostics. For each graphical technique, the test's power is also estimated for a range of assumptions and levels of significance. The simulation research and the cervical cancer patient data set are utilized to apply the aforementioned techniques.

The article is organized with the following structure. We outline the creation of residual plots in Section 2. Section 3 includes a real-world data example with graphs for specific residuals. Sections 4 and 5 explore the empirical evaluation and diagnostic power using simulated data, and section 6 discusses the findings.

2. Construction of residual plots in geometric regression

The method used to create PRES in GLMs is presented by Landwehr et al. [23], Cook and Croos-Dabrera [24], Imran and Akbar [25], and Hussain and Akbar [26]. As specified by the GLM,

$$g(\mu_i) = x'_i\boldsymbol{\beta} = \eta_i, \tag{2.1}$$

where x_i is the *i*th row of $\mathbf{x}$ and $\boldsymbol{\beta}$ is a vector parameter.

For the progression of regression issues involving a univariate response y and a $p \times 1$ a set of uncorrelated predictors $\mathbf{x}$, the constant predictor is not included. For a sample of independent observations with an identical distribution $(y_i, \mathbf{x}'_i), i = 1, 2, \ldots, N$, on the random vector$(y, \mathbf{x}')$, we use the probability function of the conditional distribution of $y|\mathbf{x}$ to be as follows:

$$f_{y|\mathbf{x}}(y|\theta, \varphi) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y, \varphi)\right\}, \tag{2.2}$$

where $a(.),\ b(.)$ and $c(.)$ are identified smooth functions, θ is the unidentified scalar parameter that depends on $\mathbf{x}$, and φ is an unknown variation parameter. In the regression function, $E(y|\mathbf{x}) = \partial\mu|\partial\theta = \mu(\mathbf{x})$ and the variance function is $\{\partial^2\mu|\partial\theta^2\}v(\varphi)$.

Following Cook and Croos-Dabrera [24], we partition $\mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2)$, where $\mathbf{x}_j$ is, $p_j \times 1, j = 1, 2$. The structure should be adequate to effectively convey the following regression function:

$$\eta(\mathbf{x}) = h\big(\mu(\mathbf{x})\big) = \alpha_0 + \boldsymbol{\alpha}'_1\mathbf{x}_1 + g(\mathbf{x}_2),$$

If the term $g(\mathbf{x}_2)$ when evaluating the parametric form, then

$$\eta(\mathbf{x}) = h\big(\mu(\mathbf{x})\big) = \alpha_0 + \boldsymbol{\alpha}'_1\mathbf{x_1} + \boldsymbol{\alpha}'_2\mathbf{x_2}. \tag{2.3}$$

Therefore, $h(.)$ is the monotonic and differentiable user-identified link function and $(\alpha_0, \boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2)'$ is vector of unknown parameters consisting of $(p_1 + 1) \times 1$ vector. The term $\mu(\mathbf{x}) = h^{-1}(\eta(\mathbf{x}))$ is either a function of $\mathbf{x}$ or a function of η, depending on interest and concerns.

The geometric distribution is given by the following [17]:

$$f(x_i, p) = p(1-p)^{x-1}, x = 1, 2, 3, \ldots$$

The mean and variance are as follows:

$$E(X) = \mu = \frac{1}{p} \text{ and } \mathrm{Var(X)} = \frac{\mathrm{q}}{\mathrm{p}^2} = \mu(1-\mu)$$

Therefore, the geometric probability mass function must be reparametrized in order for us to achieve our regression model for the mean of the geometric distribution $E(Y) = \mu = \frac{1}{p}$ and hence $p = \frac{1}{\mu}$.

Thus, it is evident that $E(Y) = \mu$ and $\mathrm{var(Y)} = \mu(1-\mu)$. Then, the geometric PMF can be written in the new parameterization as follows:

$$f(Y_i = y_i|\mu_i) = \frac{1}{\mu_i}(1 - \frac{1}{\mu_i})^{y_{i-1}}, i = 1,2, \ldots, n;\ \mu_i > 1 \tag{2.4}$$

The exponential family (2.2) allows the PMF of geometric distribution (2.4) to be expressed as follows:

$$= \exp\left[(y-1)log\frac{\mu_{i-1}}{\mu_i} - log\mu_i\right] \tag{2.5}$$

where

$$\theta = log\frac{\mu_i - 1}{\mu_i}, b(\theta) = log\mu_i \; and \; C(y) = 1.$$

Using (2.1) and (2.5), the link function of the geometric distribution can be described as follows:

$$\eta = \theta = h(\mu) = ln\frac{\mu-1}{\mu}, \tag{2.6}$$

where

$$\mu(\eta) = \frac{1}{1-\exp(\eta)}, \tag{2.7}$$

The geometric regression's log likelihood function [27] is as follows:

$$L(\boldsymbol{\beta}) = \sum_i^n\left[(y_i - 1)\ln\left(e^{\mathbf{x'}_i\boldsymbol{\beta}} - 1\right) - y_i\ln\left(e^{\mathbf{x'}_i\boldsymbol{\beta}}\right)\right], \tag{2.8}$$

The maximum likelihood estimator (MLE) for $\boldsymbol{\beta}$ can be found by resolving the undermentioned system of equations. Because the system's solution is non-linear, the approach for estimating the unknown parameters is the iterative weighted least squares.

Next, we provide the fitted geometric regression model using the following:

$$\eta_f\left(\mathbf{x}|\hat{\mathbf{b}}'\right) = h\left(\hat{\mu}_f\right) = [1 - \exp\left(\hat{b}_0 + \hat{\mathbf{b}}'_1\mathbf{x}_1 + \hat{\mathbf{b}}'_2\mathbf{x}_2\right)]^{-1}, \tag{2.9}$$

where $\hat{\mathbf{b}}' = (\hat{b}_0, \hat{\mathbf{b}}'_1, \hat{\mathbf{b}}'_2)$ and the subscript 'f'' mean on η_f and $\hat{\mu}_f$ denote the fitted model.

The coefficient estimates $\mathbf{b}_j, j = 0, 1, 2$ are obtained by minimizing the following convex objective function:

$$\hat{\mathbf{b}}' = \left(\hat{b}_0, \hat{\mathbf{b}}'_1, \hat{\mathbf{b}}'_2\right) = argmin\, L_N\left(\hat{\mathbf{b}}'\right), \tag{2.10}$$

where,

$$L_N\left(\hat{\mathbf{b}}\right) = \frac{1}{N}\sum_{i=1}^{N} L\left(\eta_f\left(\mathbf{x}_i|\hat{\mathbf{b}}, y\right)\right),$$

$$= \tfrac{1}{N}\textstyle\sum_i^n L\left(\left[1 - \exp\left(\hat{b}_0 + \hat{\mathbf{b}}'_1\mathbf{x}_1 + \hat{\mathbf{b}}'_2\mathbf{x}_2\right)\right]^{-1}, y_i\right),$$

where L (.,.) is an objective function chosen by the user that is presumptively convex with regard to its first argument. The usage of ordinary least squares and the maximum likelihood under (2.3) and (2.9) with a canonical link, where $\theta = \eta$, and specific robust estimates are at the very least included in this

class, which is not overly restricted. For the geometric regression with the connection specified in (2.6), the objective function regarding the maximum likelihood is as follows:

$$L(\eta_f(\mathbf{x}_i|\hat{\mathbf{b}}, y)) = \left\{\log\left(\frac{1}{1-\exp(\eta_f)}\right) - y\eta_f\right\}. \tag{2.11}$$

The objective function class corresponding to (2.10) is generalized to form the class of convex objective functions, $L(\eta_f, y) = L(y - \eta_f)$, used by Cook [28] for additive error models (2.7).

A partial residual Pr2 for $\mathbf{x}_2$ is obtained using (2.7) and (2.8) via (2.9), and is given by the following:

$$\text{PRES2} = (y - \hat{\mu}_f)h'(\hat{\mu}_f) + \hat{\mathbf{b}}'_2\mathbf{x}_2, \tag{2.12}$$

where, $h'(.)$ is the first derivative of $h(.)$ w.r.t 'μ' and $\hat{\mathbf{b}}$ can be obtained by (2.10).

The term $\hat{\mu}_f(\mathbf{x}) = h^{-1}(\eta_f(\mathbf{x}|\hat{\mathbf{b}}, y))$ is the regression function of μ_f evaluated at $\hat{\mathbf{b}}$.

The geometric link function's first derivative, which is given in (2.6), is as follows:

$$h'(\hat{\mu}_f) = -\frac{\mu}{(\mu-1)^3}$$

As a result, the fitted model utilizing a link for the geometric regression is as follows:

$$\hat{\mu}_f = [1 - \exp(\hat{b}_0 + \hat{\mathbf{b}}'_1\mathbf{x}_1 + \hat{\mathbf{b}}'_2\mathbf{x}_2)]^{-1}$$

where $\hat{b}_0, \hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2$ are the regression estimators, $\hat{\mu}_f$ denotes the fitted model, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are the predictors. The PRES for $\mathbf{x}_2$ is as follows:

$$\text{PRES2} = -(y - \hat{\mu}_f)\frac{\mu}{(\mu-1)^3} + \hat{\mathbf{b}}'_2\mathbf{x}_2.$$

Similar to this, the PRES for the model with p explanatory variables can be represented as follows:

$$\text{PRES}i = -(y - \hat{\mu}_f)\frac{\mu}{(\mu-1)^3} + \hat{\mathbf{b}}'_i\mathbf{x}_i, i = 1, 2, \ldots, p, \tag{2.13}$$

where p is the explanatory variable and the fitted model is as follows:

$$\hat{\mu}_f = [1 - \exp(\hat{b}_0 + \hat{\mathbf{b}}'_1\mathbf{x}_1 + \hat{\mathbf{b}}'_2\mathbf{x}_2 + \cdots + \hat{\mathbf{b}}'_p\mathbf{x}_p)].^{-1}$$

Due to the response residual, the PRES found in (2.13) can be denoted as RPRES. The PRES of the response is now given as follows:

$$\text{RPRES}i = -(y - \hat{\mu}_f)\frac{\mu}{(\mu-1)^3} + g(\mathbf{x}_i), i = 1, 2, \ldots, p. \tag{2.14}$$

where $g(\mathbf{x}_i) = \hat{\mathbf{b}}'_i\mathbf{x}_i$ for the PRES.

Equation (2.14) can be used to obtain the RAPRES, RCERES (L), and RCERES (K).

Using (2.14), one may obtain the CERES and APRES. When we substitute a quadratic and/or interaction factor for $g(\mathbf{x}_i)$ in (2.14), an APRES is obtained. Similarly, the conditional expectation

$(\mathbf{x}_i|\mathbf{x}_j)$ is used instead of $g(\mathbf{x}_i)$ in (2.14) to obtain a CERES. The $g(\mathbf{x}_i)$ is sometimes non-parametrically estimated by using LOESS and the kernel function [24,26,28,29].

Example. Data set for cervical cancer patients.

We took advantage of Irawan's [30] data on cervical cancer cases in Indonesia from the Sepuluh Nopember Institute of Technology. The purpose of this study was to identify the patient's frequency service, which are dispersed geometry parameters that affect the cervical cancer patients' survival. The total number of cases reported from the data collected from the cervical cancer patient survey is 198. The variable *Y* (cervical cancer patient) represents a response, and the explanatory variables are $\mathbf{x}_1$(age), $\mathbf{x}_2$(chemotherapy), $\mathbf{x}_3$(complications), $\mathbf{x}_4$(anemia), and $\mathbf{x}_5$(operation).

The regression model with a geometric fitting is provided by the following:

$$\hat{\mu}_f = [1 - exp(\hat{b}_0 + \hat{\mathbf{b}}'_1\mathbf{x}_1 + \hat{\mathbf{b}}'_2\mathbf{x}_2 + \hat{\mathbf{b}}'_3\mathbf{x}_3 + \hat{\mathbf{b}}'_4\mathbf{x}_4 + \hat{\mathbf{b}}'_5\mathbf{x}_5)]^{-1}.$$

By utilizing the iterative weighted least square method of the estimate, the necessary calculations are provided as follows in Table 1:

Table 1. Goodness of fit test.

Distribution	Null Deviance	d.f	Residual Deviance	d.f
Geometric	92.256	197	20.238	192
Negative Binomial	353.364	197	83.165	192
Poisson	353.375	197	83.168	192

By observing the null and residual deviances in Table 1, the geometric regression shows a strong result when compared to other popular distributions because the null and residual deviances have minimal values among the other distributions; therefore, the geometric regression is more suitable than the Poisson regression and the negative binomial regression for this example of data from cervical cancer patients [30].

Table 2 provides both descriptive statistics for the data and summary statistics for all the geometric regression model's metrics. The P-values for X_2 and X_4 are less than 0.05, which indicates a significance; however, the P-values for X_1, X_3, and X_5 are higher than 0.05, which indicates non-significance.

Table 2. Geometric regression analysis of variance.

Variable	Coefficients	SE(β)	t-value	Pr(>|t|)	VIF
Intercept	−0.0172	0.1835	0.9250	0.9250	
X_1	0.0018	0.0034	0.6061	0.6060	2.5067
X_2	0.3749	0.0163	22.948	0.0000	5.7547
X_3	0.1151	0.0795	1.4462	0.1500	4.5623
X_4	0.3241	0.0661	4.9070	0.0001	7.9876
X_5	−0.0102	0.0957	-0.1071	0.9150	1.3487
AIC = 802.1; BIC = 821.8; R^2 (%) = 98.80%; Adj-R^2(%) = 98.77%					

Now, we provide the diagnostics of the aforementioned issues using certain official testing methodologies to support our conclusions (As shown in Table 3).

Table 3. Diagnostic tests for outliers in regression.

Test Statistics	Statistic	P-Value
Grubb's Test	2.4599	0.0370
Anderson-Darling Test	0.7471	0.0021
F-test (Overall)	129.25	0.0000

The data in the above table show that the Grubb's test is an outlier and the Anderson-Darling test, which has previously been used by Hussain and Akbar [26], is a non-normality. There are numerous formal tests available for diagnostics, and we are aware that each of these tests is only utilized for a single diagnostic because they are all predicated on certain regularity constraints and are more computationally costly. In light of the discussion above, we can draw the conclusion that a single residual plot can be utilized for various diagnostics and is more efficient than traditional tests.

An outlier can be identified utilizing the response PRES plots (Figure 1a), which shows that there are red values that are much higher than all other values. Non-normality is anticipated because none of the points fall along the trend line and exhibit an erratic pattern on the residual plot. The issues shown in Figure 1a can also be detected in the remaining plots, as shown in Figure 1b–1e.

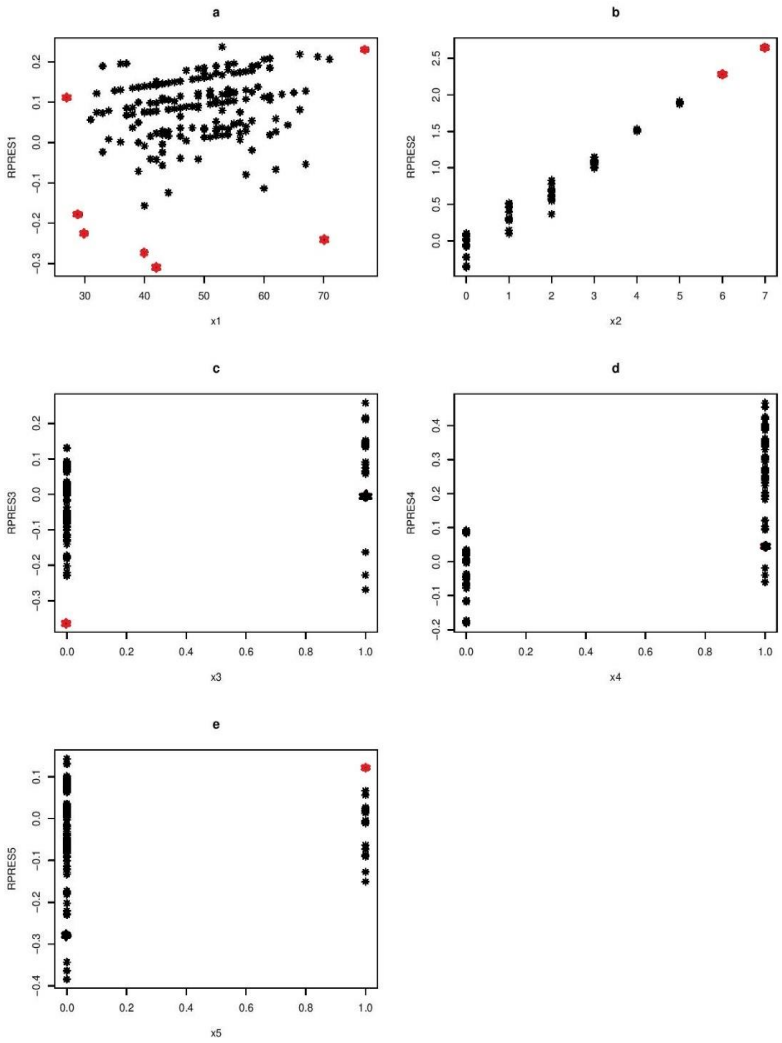

Figure 1. PRES Plots using response residuals with geometric fits.

An outlier can be found using the response APRES plots (Figure 2a), which show that there are red values that are significantly bigger than all other values. Non-normality is expected since no point on the residual plot has an abnormal pattern or falls along the trend line. The issues shown in Figure 2a are also evident in the subsequent plots, as Figures 2b–2e show.

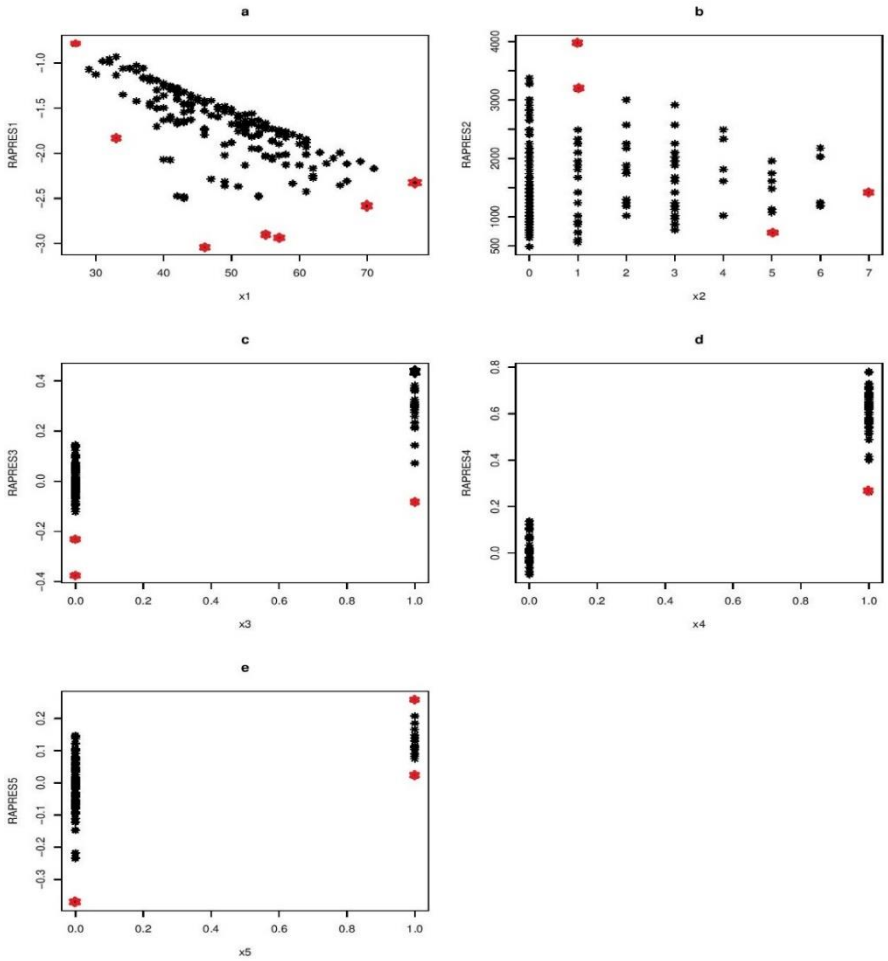

Figure 2. APRES Plots using response residuals with geometric fits.

Finding an outlier can be done with the use of the response CERES (L) plots (Figure 3a), which show that there are red values that are much higher than every other value. Non-normality is anticipated since there isn't a single point on the residual plot that exhibits an unusual pattern or falls along the trend line. The issues shown in Figure 3a also exist in the other plots, as can be seen from Figures 3b–3e.

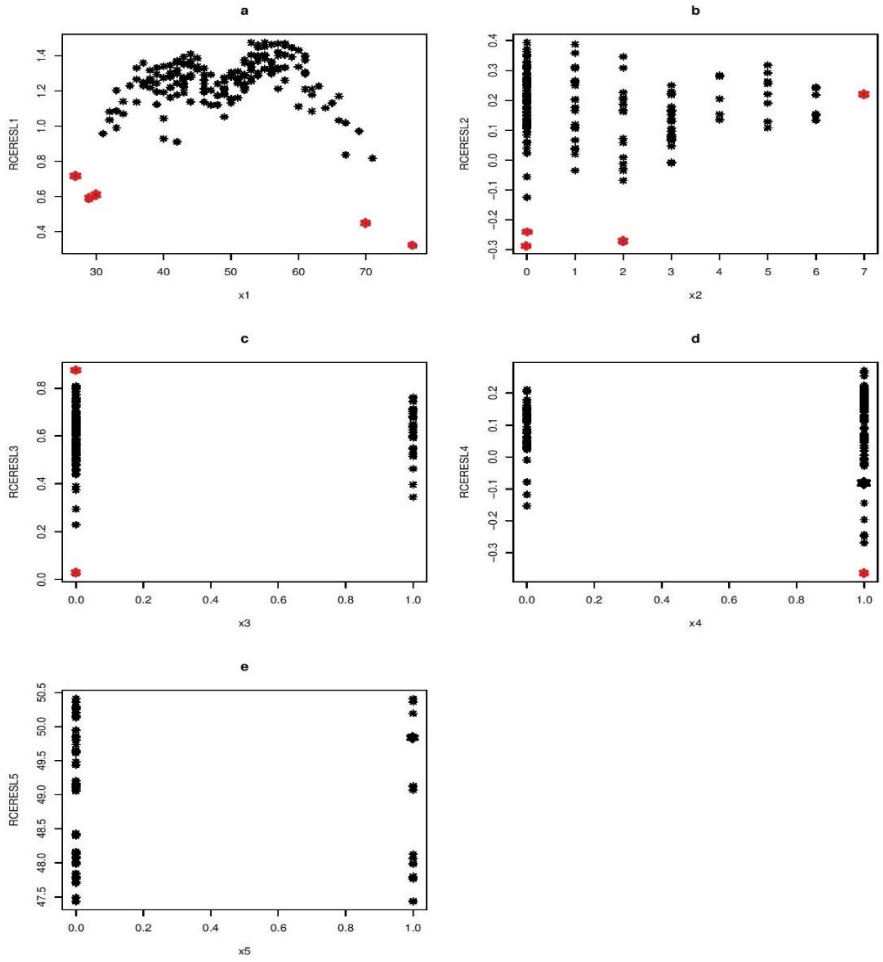

Figure 3. CERES (L) Plots using response residuals with geometric fits.

The response CERES(K) plots (Figure 4a), can be used to identify an outlier, which indicate that there are red values that are significantly greater than all other values. Non-normality is expected since none of the points on the residual plot show an irregular pattern or fall along the trend line. As Figure 4b–4e demonstrate, the problems depicted in Figure 4a are also present in the remaining plots. Consequently, we may view the outlier's diagnostics using residual plots in the geometric regression model.

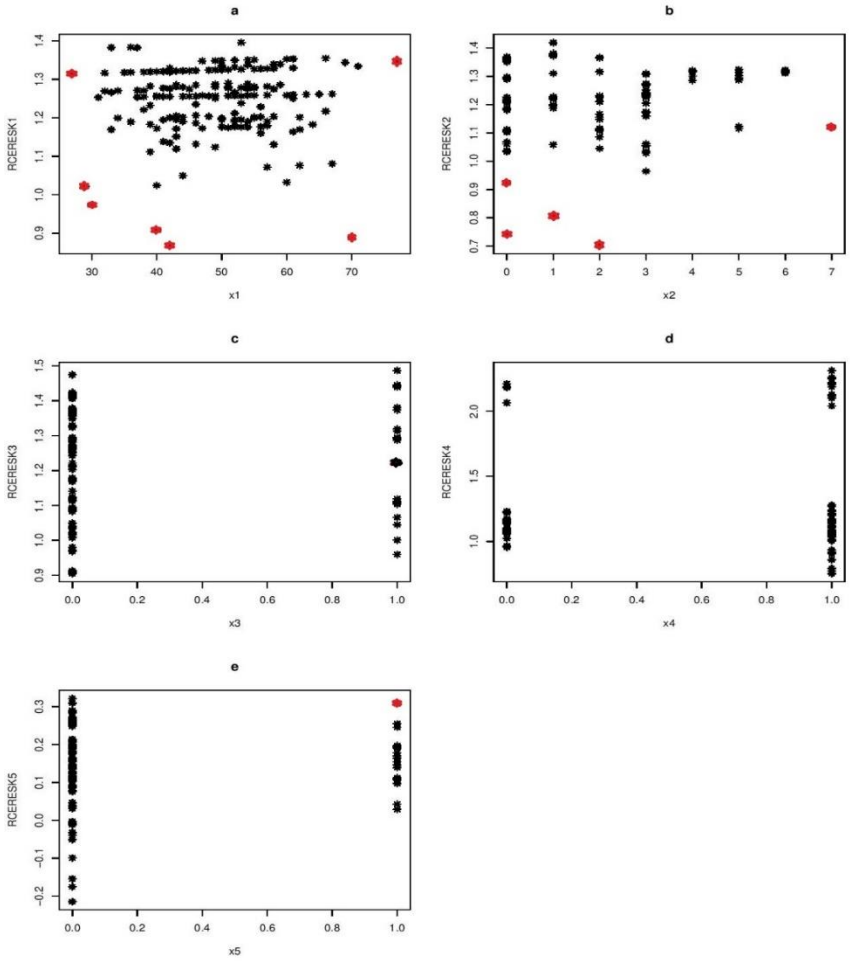

Figure 4. CERES (K) Plots using response residuals with geometric fits.

4. Analyzing statistics based on residuals using empirical methods

Azzalini and Bowman [31] and Oh [29] used modified pseudo-likelihood tests based on RES, PRES, APERS, conditional expectation, and residual CERES to assess the non-linearity and estimated empirical powers of the presented statistics. In this article, outliers on the RES, PRES, APRES, CERES (L), and CERES (K) were detected using the Grubb's test [26]. The computed and accessible powers of the aforementioned statistics are shown in Tables 4 and 5. The empirical power is determined for α = 0.01 and 0.05. Four sample sizes (n=20, 30, 40, 50) and four values of 'a' (2, 10, 20, 95) are chosen.

Equation (2.9) contains the geometric regression model.

This section evaluates how well the residuals (PRES, APRES, CERES, etc.) work for identifying outliers. The hypothesis takes the following shape:

H_0: There isn't any outlier;

H_1: At least one outlier exists;

where,

$$G_{crit} > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}},$$

where, N is the quantity of the observation, with $t_{\alpha/(2N),N-2}$ indicating the higher central value of the t- distribution with N-2 degrees of freedom and a significance level of $\alpha/(2N)$. The Grubb's test statistics are as follows:

$$G = \frac{max_i|Y - \bar{Y}|}{S}, i = 1,2,3, \dots, N.$$

Therefore, we construct the formal test based on the Grubb's test of outliers and statistically significant large value of G (i.e., $G > G_{crit}$).

5. Performance of partial residual plots for outlier assessment using simulation

To compare the effectiveness of the outlier's test, a modest power study was conducted. By using the geometric regression model, where we create the response variable from a geometric regression, data of the response variable were generated, and $\mathbf{x}_1$ was produced using a homogeneous random variable on the range (0, 30) and $g(\mathbf{x}_1) = \frac{a}{1+e^{-\mathbf{x}_1}}, \mathbf{x}_2 = \log(\mathbf{x}_1) + N(0, 0.25^2), \mathbf{x}_3 = g{\mathbf{x}_1}^{-1} + N(0,1)$.

We've set the regression parameters' fixed values as follows:

$$\boldsymbol{\beta}_1 = 0.05, \boldsymbol{\beta}_2 = 0.01, \boldsymbol{\beta}_3 = 0.003,$$

where Y is a deterministic function of the three covariates because no errors were present. This preserves the qualitative integrity of the results while allowing the conclusion to be conveyed more succinctly than if an additional mistake was added. Sample sizes of n = 20, 30, 40, 50 were considered. Two thousand samples were generated for each case, and the frequency with which the observed significances were below 0.05 or 0.01 was counted. For deriving the conditional expectation in CERES, the LOESS smoothing data count was fixed at n|2 and the bandwidth of the kernel function was 0.5. The bandwidth of the kernel function to obtain a test statistic was 0.25. We chose a = 2, 10, 20, 95 in curve $g(.)$ using a roughness measure [26].

5.1. Simulation result and discussion

We used the notations RES, PRES, APRES, CERES (L), and CERES (K). We demonstrate the empirical validity of the outliers' test statistic in Tables 4 and 5. In these tables, the roughness metric had a significant impact on the test statistic 'a' Additionally, the empirical powers of RES and APRES decreased as the sample size increased; alternatively, the power of PRES, CERES (L), and CERES (K) increased as the sample size increased. The test statistic based on PRES, CERES (L), and CERES (K) was found to be more outlier-sensitive. Therefore, in order to check the outliers of chosen covariates in a geometric regression model for diagnostics, we propose and advise to employ an outliers' test statistic based on PRES plots.

Table 4. Empirical power of various residuals using Grubb's test statistic for geometric regression for detecting outliers for $\alpha = 0.05$.

$\alpha = 0.05$						
n	a	RES	PRES	APRES	CERES(L)	CERES(K)
10	2	0.160	0.865	0.980	0.980	0.965
	10	0.180	0.800	0.925	0.990	0.930
	20	0.165	0.870	0.945	0.975	0.960
	95	0.115	0.845	0.905	0.990	0.950
20	2	0.360	0.860	0.930	0.980	0.950
	10	0.265	0.850	0.950	0.945	0.925
	20	0.360	0.835	0.930	0.950	0.950
	95	0.405	0.880	0.965	0.990	0.950
30	2	0.475	0.825	0.925	0.965	0.920
	10	0.505	0.860	0.925	0.995	0.940
	20	0.540	0.885	0.970	0.970	0.920
	95	0.530	0.875	0.990	0.980	0.900
40	2	0.605	0.895	0.980	0.920	0.935
	10	0.600	0.820	0.950	0.965	0.910
	20	0.590	0.835	0.995	0.935	0.910
	95	0.545	0.820	0.945	0.900	0.895

Table 5. Empirical power of various residuals using Grubb's test statistic for geometric regression for detecting outliers for $\alpha = 0.01$.

$\alpha = 0.01$						
n	a	RES	PRES	APRES	CERES(L)	CERES(K)
10	2	0.030	0.775	0.950	0.985	0.980
	10	0.025	0.750	0.995	0.995	0.920
	20	0.045	0.780	0.905	0.910	0.990
	95	0.055	0.890	0.915	0.930	0.985
20	2	0.200	0.870	0.970	0.975	0.935
	10	0.215	0.800	0.995	0.990	0.925
	20	0.125	0.840	0.960	0.960	0.930
	95	0.230	0.860	0.965	0.945	0.980
30	2	0.290	0.855	0.980	0.990	0.980
	10	0.240	0.840	0.925	0.965	0.945
	20	0.370	0.805	0.920	0.980	0.990
	95	0.305	0.865	0.910	0.965	0.945
40	2	0.385	0.815	0.915	0.985	0.885
	10	0.335	0.800	0.945	0.995	0.850
	20	0.380	0.830	0.940	0.930	0.855
	95	0.385	0.840	0.980	0.900	0.850

Tables 4 and 5 display the empirical power computations of five altered graphical methods on the basis of the Grubb's test for two altered level of 'α'. It is evident from the above tables that APRES, CERES (L), and CERES (K) bear good standings among others since they have the highest power levels. Moreover, it is crucial to note that the values of 'a' significantly affect how well these techniques work. The effect of the sample size is since the power and some alpha influences are also seen to have an effect. In comparison to RES and PRES techniques, APRES, CERES (L, and CERES (K) exhibit more consistent behaviors for larger values of the sample size and have higher values of power.

As shown in Table 6, by observing he null and residual deviances, the geometric regression shows a strong result as compared to other popular distributions because the null and residual deviances have minimal values among the other distributions; therefore, a geometric regression is more suitable in simulation data.

Table 6. Goodness of fit test using simulation data.

Distribution	Null Deviance	d.f	Residual Deviance	d.f
Geometric	188.08	197	185.10	194
Negative Binomial	194.02	197	193.17	194
Poisson	367.36	197	357.57	194

Tables 6 and 7 show collective results for simulation data and confirm the real data results. In Figure 5, similar results are observed in the PRES, APRES, CERES (K), and CERES (L) plots using the simulation data. Thus, residual plots in the geometric regression model allow us to see the diagnostics of the outlier.

Table 7. Geometric regression analysis of variance using simulation data.

Variable	Coefficients	SE(β)	t-value	Pr(>|t|)	VIF
Intercept	0.06095	0.43023	0.142	0.887	
X_1	−0.03631	0.02676	−1.357	0.176	3.1028
X_2	0. 22528	0.29039	0.776	0.439	6.1230
X_3	0.06954	0.21218	0.328	0.743	4.1524
AIC = 561.91; BIC = 566.4865; R^2 (%) = 98.90%; Adj-R^2(%) = 97.96%					

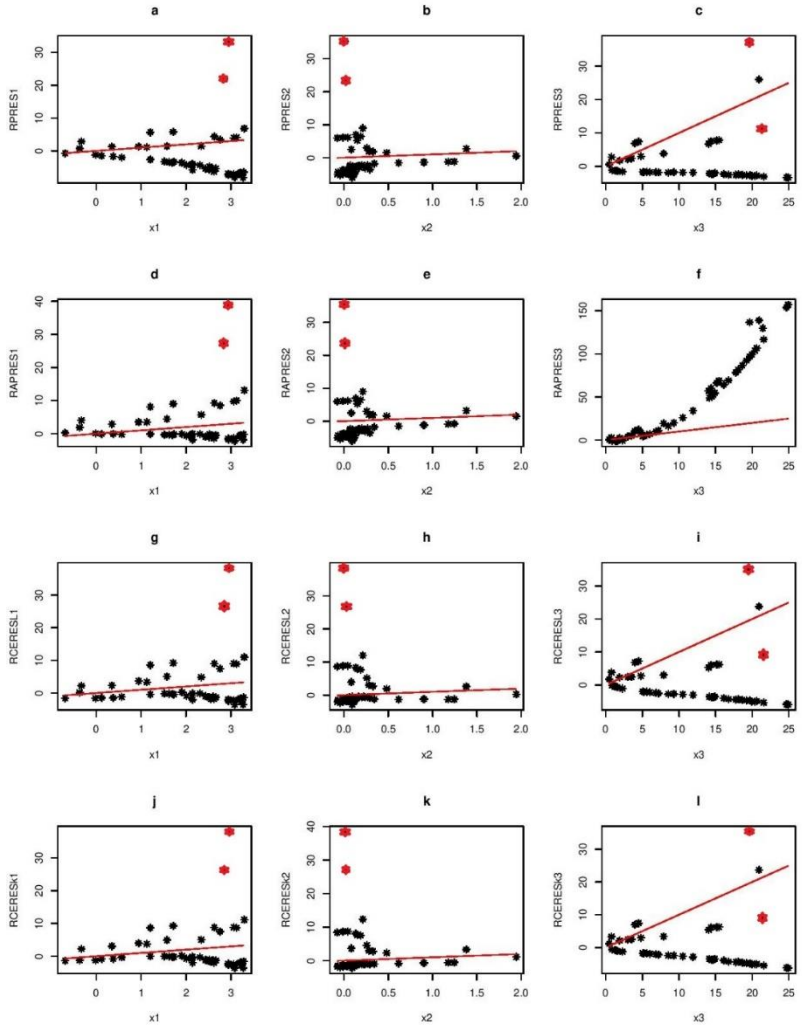

Figure 5. Response residual plots using simulation data with geometric fits.

6. Conclusions

Cook and Croos-Dabrera [24] studied the modification of the explanatory variable in several regression settings using the PRES plots in GLM. They investigated the circumstances under which the PRES plots can be used to alter the predictor in the GLM class [25]. By determining the statistical significance power of the tests, Hussain and Akbar [26] created and observed the pattern of RES, PRES, APRES, CERES (L), and CERES (K) for hindered internal rotational (HIR) treatment data of the chemical species plots of a binomial regression.

The development and assessment of this study's is the construction and evaluation of PRES, APRES, CERES (L), and CERES (K) plots for geometric regression in a dataset for cervical cancer patients. It was observed that PRES, APRES, CERES (L), and CERES (K) plots were a fantastic technique for the outlier's diagnostics. Therefore, residual plots are an effective graphical approach for diagnosing outliers in a geometric regression. Moreover, APRES, CERES (L), and CERES (K) perform better than PRES plots and in a diagnostic power simulation scheme.

While dealing with the cervical cancer data and in simulation study, it was observed that outliers are present in the data. Both conventional and graphical methods confirm this diagnostic. It is very

important for the applied scientist to address these diagnostics prior to the modelling of the data. It is a well-known fact that the violation of the aforementioned assumptions may result in an insignificance of results, which arise in our case of cervical cancer. Therefore, according to the results, none of the factors are responsible for executions in cervical cancer. This is very misleading and erroneous. To deal with such situations, statistics suggest corrective actions and data handling. In order to identify issues that need to be taken into account prior to running the final model that can be utilized for policy purposes, the current study addresses these crucial scenarios and largely offers easy graphical methods.

The specified GLM, the link function, and the stochastic behavior of the predictors may all have limitations on the usefulness of residual plots for generating a clear visual picture of the curvature. Due to their broad applicability in the outlier's diagnostics employing visual impression underlying predictor transformation, residual plots were shown to be more colorful than conventional methods. Due to the performance of PRES, CERES (K), and CERES (L), these are recommended for outlier diagnostics in the presence of conventional tests.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP.2/44/44) and this study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444)

Conflict of interest

The authors declared that they have no conflicts of interest regarding the publication of this work.

References

1. P. McCullagh, J. A. Nelder, *Generalized linear models*, Chapman and Hall, 1989. Available from: https://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/glmbook.pdf.
2. M. Otto, Chemometrics: statistics and computer application in analytical chemistry, John Wiley & Sons, 2016. Available from: https://www.wiley.com/en-us/exportProduct/pdf/9783527699384.
3. A. F. Lukman, K. Ayinde, S. Binuomote, O. A. Clement, Modified ridge-type estimator to combat multicollinearity: Application to chemical data, *J. Chemometr.*, **33** (2019), e3125. https://doi.org/10.1002/cem.3125
4. A. Zeileis, C. Kleiber, S. Jackman, Regression models for count data in R, *J. Stat. Softw.*, 27 (2008), 1–25.
5. W. S. Cleveland, Graphs in scientific publications, *Am. Stat.*, **38** (1984), 261–269. https://doi.org/10.1080/00031305.1984.10483223
6. W. G. Jacoby, Statistical graphics for univariate and bivariate data, Sage, 1997.

7. J. Textor, J. Hardt, S. Knuppel, Dagitty: A graphical tool for analyzing causal diagram, *Epidemiology*, **22** (2011), 745. https://doi.org/10.1097/EDE.0b013e318225c2be
8. W. A. Larsen, S. J. McCleary, The use of partial residual plots in regression analysis, *Technometrics*, 14 (1972), 781–790. https://doi.org/10.1080/00401706.1972.10488966
9. E. R. Mansfield, M.D. Conerly, Diagnostic value of residual and partial residual plots, *Am. Stat.*, **41** (1987), 107–116. https://doi.org/10.1080/00031305.1987.10475457
10. A. C. Atkinson, Regression diagnostics, transformations and constructed variables, *J. R. Stat. Soc. Ser. B (Meth.)*, **44** (1982), 1–22. https://doi.org/10.1111/j.2517-6161.1982.tb01181.x
11. A. C. Davison, C. L. Tsai, Regression model diagnostics, *Int. Stat. Rev.*, **60** (1992), 337–353. https://doi.org/10.2307/1403682
12. R. J. O'Hara Hines, E. M. Carter, Improved added variable and partial residual plots for the detection of influential observations in generalized linear models, *J. R. Stat. Soc. Ser C. (Appl. Stat.)*, **42** (1993), 3–20. https://doi.org/10.2307/2347405
13. P. C. Wang, Residual plots for detecting nonlinearity in generalized linear models, *Technometrics*, **29** (1987), 435–438. https://doi.org/10.1080/00401706.1987.10488271
14. R. D. Cook, S. Weisberg, *Residuals and influence in regression*, New York: Chapman and Hall, 1982.
15. M. M. A. Almazah, T. Erbayram, Y. Akdoğan, M. M. Al Sobhi, A. Z. Afify, A new extended geometric distribution: Properties, regression model, and actuarial applications, *Mathematics*, **9** (2021), 1336. https://doi.org/10.3390/math9121336
16. J. Makcutek, A generalization of the geometric distribution and its application in quantitative linguistics, *Rom. Rep. Phys.*, **60** (2008), 501–509.
17. F. Jahan, B. Siddika, M. A. Islam, An application of the generalized linear model for the geometric distribution, *J. Stat.: Adv. Theory. Appl.*, **16** (2016), 45–65. http://doi.org/10.18642/jsata_7100121695
18. B. Pradhan, D. Kundu, A choice between Poisson and geometric distributions, *J. Indian Soc. Prob. Stat.*, **17** (2016), 111–123. https://doi.org/10.1007/s41096-016-0008-2
19. Z. M. D. Al-Balushi, M. M. Islam, Geometric regression for modelling count data on the time-to-first antenatal care visit, *J. Stat.: Adv. Theory. Appl.*, **23** (2020), 35–57. http://doi.org/10.18642/jsata_7100122148
20. P. J. Saulnier, M. Darshi, K. M. Wheelock, H. C. Looker, G. D. Fufaa, W. C. Knowler, et al., Urine metabolites are associated with glomerular lesions in type 2 diabetes, *Metabolomics*, **14** (2018), 84. https://doi.org/ 10.1007/s11306-018-1380-6
21. G. Xie, J. T. Lundholm, J. S. MacIvor, Phylogenetic diversity and plant trait composition predict multiple ecosystem functions in green roofs, *Sci. Total Environ.*, **628-629** (2018), 1017–1026. https://doi.org/10.1016/j.scitotenv.2018.02.093
22. J. M. Wouters, J. B. Gusmao, G. Mattos, P. Lana, Polychaete functional diversity in shallow habitats: Shelter from the storm, *J. Sea. Res.*, **135** (2018), 18–30. https://doi.org/10.1016/j.seares.2018.02.005
23. J. M. Landwehr, D. Pregibon, A. C. Shoemaker, Graphical methods for assessing logistic regression models, *J. Am. Stat. Assoc.*, **79** (1984), 61–71. https://doi.org/10.1080/01621459.1984.10477062
24. R. D. Cook, R. Croos-Dabrera, Partial residual plots in generalized linear models, *J. Am. Stat. Assoc.*, **93** (1998), 730–739. https://doi.org/10.1080/01621459.1998.10473725

25. M. Imran, A. Akbar, Diagnostics via partial residual plots in inverse Gaussian regression, *J. Chemometr.*, **34** (2020), e3203. https://doi.org/10.1002/cem.3203
26. Z. Hussain, A. Akbar, Diagnostics through residual plots in binomial regression addressing chemical species data, *Math. Probl. Eng.*, **2022** (2022), 437594. https://doi.org/10.1155/2022/4375945
27. J. L. Hintz, User guide–III: Regression and curve fitting, kaysville: NCSS, 2007. Available from: https://www.ncss.com/wp-content/uploads/2012/09/NCSSUG3.pdf.
28. R. D. Cook, Exploring partial residual plots, *Technometrics*, **35** (1993), 351–362. https://doi.org/10.1080/00401706.1993.10485350
29. K. Oh, Regression diagnostics using residual plots, *Korean. Commun. Stat.*, **8** (2001), 311–317. Available from: https://koreascience.kr/article/JAKO200111920779561.pdf.
30. A. R. Irawan, Pemodelan perulangan pengobatan pasien kanker serviks di rsud dr. soetomo dengan bayesian geometric regression dan bayesian mixture geometric regression, Ph D thesis, Institut teknologi sepuluh nopember, surabaya, 2017. Available from: https://core.ac.uk/download/pdf/291465419.pdf.
31. A. Azzalini, A. W. Bowman, On the use of nonparametric regression for checking linear relationship, *J. R. Stat. Soc. Ser. B (Meth.)*, **55** (1993), 549–557. https://doi.org/10.1111/j.2517-6161.1993.tb01923.x

AIMS Press