Research Article

Diagnostics through Residual Plots in Binomial Regression Addressing Chemical Species Data

Zawar Hussain and Atif Akbar

Department of Statistics, Bahauddin Zakariya University, Multan 60800, Pakistan

Correspondence should be addressed to Atif Akbar; atifakbar@bzu.edu.pk

Received 25 November 2021; Accepted 28 January 2022; Published 23 February 2022

Academic Editor: Firdous Khan

Copyright © 2022 Zawar Hussain and Atif Akbar. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Binomial regression is used as a generalized linear model (GLM) in natural sciences to identify the covariate structure that is responsible for outcomes. It is very important to assess the adequacy and effectiveness of any model before its implementation. In GLM context, this study explores the structure and usefulness of partial residual (PRES), augmented partial residual (APRES), and conditional expectation and residuals (CERES) plots for visualizing influence diagnostics as a function of selected predictors. Binomial regression is considered here with predictor transformation, and PRES, APRES, and CERES plots are constructed for diagnostics of outliers and multicollinearity. The efficacy of these plots for obtaining a good visual impression may be varied due to behaviour of response variable and allied link function with different covariates. The certain techniques are applied on the data of hindered internal rotational (HIR) treatment of chemical species to recognize patterns for efficient modelling. The power of the tests for different plots shows that APRES and CERES (L) endure maximum power for detection of outliers and multicollinearity. The results revealed that residuals plots are more effective as compared to the conventional methods and help the scientists to easily and effectively model the data for their diagnostics policies.

1. Introduction

In real patterns, thermodynamic properties are usually used to address different chemical and biological processes. A rigorous hindered internal rotation (HIR) treatment is generally used to acquire accurate thermodynamic data of chemical species for a variety of temperatures. This needs sufficient information regarding internal rotation. To discuss the patterns and relationships in the data regarding HIR, statistical methods have been established that lead to the reliable conclusions [1–3].

Applications of statistical methods in pure and applied sciences are worth mentioning. Among other tools, regression analysis is furthermost famous to quantify the relationships among different entities. Generalized linear models (GLM), being special case of regression models, use exponential family-based responses to handle relatively complex interactions that are usually found in chemical sciences, social sciences, medical and life sciences,

economics, finance, etc. Like other theories, regression analysis is based on some assumptions that are needed to be addressed prior to their implementations [4–6].

Statistical graphs are the most powerful techniques used in presenting, describing, supporting, and interpreting the numerical data. Statistical graphs usually show the relationship between two or more variables. The power of statistical graphics arises from the fact that they can suggest large quantities of information both quickly and efficiently because the human can quickly derive the understanding which would not come from the list of values [7, 8].

Different statistical techniques have been used to identify the violations of assumptions in regression models, and a plot of residuals depicts the relationship for the set of residuals against each of the explanatory variable in the model. This relationship would suggest whether a variable should be included in the model or not [9].

The use and interpretation of regression models depends on the estimates of the individual regression coefficient.

Influential outliers can bias parameter estimates and make the resulting analysis less useful. It is important to detect outliers since the outliers can provide misleading results. This estimator is generally best when the explanatory variables are independent. However, when multicollinearity exists, that is, when the explanatory variables are related, the estimator suffers a major setback in which the regression coefficients produce wrong signs with large standard error [10]. Different plots have been used to detect certain problems from the fitted model, but the most useful are partial residual plots [11–14].

These plots are usually used in regression diagnostics; the transformation is needed in the explanatory variables if they are noncollinear and not directly connected in the regression model, and also, these plots help us to improve the fit of logistic regression models in the generalized linear models [15, 16].

The added variable plots are used for the detection of influential observations, whereas the partial residual plots are more helpful for diagnosing nonlinearity [17, 18]. The partial residual plots help the researcher in detecting outliers, assessing the presence or absence of variance heterogeneity and determining if a transformation of the explanatory variable is needed or if another term needs to be added in the model [19].

To demonstrate the importance of statistics in chemistry is the realization of the imperative of widespread errors. There is good interest in basic statistics in chemistry, especially physical, environmental, and analytical chemistry which depends on quantitative measurements and computations. So, the chemist used statistics, as the causes can be broad [20]. Statistical thermodynamics also called statistical mechanics which include macroscopic properties of thermodynamics corresponding to individual particle behaviour and their interaction and compute the difference in energy among the staggered ground state and eclipsed transition state in angle of rotation [21].

Statistical analysis plays an important role in the field of chemical sciences, biology, bioinformatics, biochemistry, and other fields of natural sciences and developed the nonlinear regression model in the analytical chemistry to describe the relationship among the response and one or more stimulus variables. The example also shows some analysis of data, model compatibility, diagnostic verification, forecasting, and presentation of results [22, 23].

The application of logistic regression analysis in livestock data has been examined, and the logistic regression model used to determine the relationship between the binary response variable and the independent variables may be discrete, continuous, or the combination of them. Logistic regression is the most important mathematical modelling approach used to describe the relationship among variables, and they apply the maximum likelihood estimation procedure for the purpose of interpretations of coefficients [24–26]. He produced a detailed guide in the procedure for logistic regression to obtain the odd ratios. We take the examples to explain it as simple as possible and also discuss the main issues and interpret the result [27].

The logistic regression model, in the ipsilateral breast tumour relapse (IBTR) cancer patient's data, presents a common analysis framework for logistic regression modelling with a binary score classified and time associated with IBTR and also discusses the simulation study [28].

The logistic regression model is applied on cucumber data with a longer shelf life after the postharvest while taking marketing possibility as the response variable, and the storage days, varieties, fruit weight loss, and months of evaluation were used as explanatory variables. The logistic regression model describes that weight loss of an option, on which the fruit will be rejected on the market because the storage time was the main cause of cucumber weight loss strongly, influenced the probability of marketability [29].

Landwehr et al. [30] addressed the partial residual plots for GLM under canonical relations, for visualizing unknown function and identifying the transformation need for predictors in regression with a binary response. The applications of partial residual plots are for visualizing impression of curvature in binary logistic regression as a function of predictors [31]; also, we used the application of partial residual plots in inverse Gaussian regression and discussed the link function and stochastic behaviour of the predictors in the class of GLM [32].

The current study addresses the construction of PRES, APRES, CERS (K), and CERES (L) using response residual in the logistic regression model. The construction will be made using the transformation of regressors. These plots are then used for diagnostics in logistic GLM to generate a suitable model.

The paper is organized as follows. In Section 2, the construction of residual plots is presented. Section 3 contains real data example with residual plots, and Section 4 presents the empirical evaluation of different residual plots. In Section 5, the simulation study with diagnostic power is presented, while Section 6 concludes.

2. Partial Residual Plots' Construction in Logistic Regression

The process of construction of partial residuals in GLMs is available in [30] and in [31, 32]. The GLM is defined as

$$\begin{aligned} g(\mu_i) &= \mathbf{x}_i' \boldsymbol{\beta} \\ &= \eta_i. \end{aligned} \quad (1)$$

The development in the context of regression problems is with a univariate response y and a $p \times 1$ vector of random predictors $\mathbf{x}$, excluding the constant predictor. For a sample of independent and identically distributed observations $(y_i, \mathbf{x}_i')$, $i = 1, 2, \dots, N$, on the random vector $(y, \mathbf{x}')$, we take the probability mass function of the conditional distribution of $y|\mathbf{x}$ to be

$$f_{y|\mathbf{x}}(y|\theta, \varphi) = \exp \left\{ \frac{(y\theta - b(\theta))}{a(\varphi)} + c(y, \varphi) \right\}. \quad (2)$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known smooth functions, θ is the unknown scalar valued parameter that depends on $\mathbf{x}$, and

φ is an unknown dispersion parameter. In regression function, $E(y|\mathbf{x}) = \partial\mu/\partial\theta = \mu(\mathbf{x})$ and variance function is $\{\partial^2\mu/\partial\theta^2\}v(\phi)$.

Following [31], we partition $\mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2)$, where $\mathbf{x}_j$ is $p_j \times 1$, $j = 1, 2$. We assume that the regression function can be characterized adequately by the structure:

$$\begin{aligned}\eta(\mathbf{x}) &= h(\mu(\mathbf{x})) \\ &= \alpha_0 + \alpha'_1 \mathbf{x}_1 + g(\mathbf{x}_2),\end{aligned}\quad (3)$$

If the term $g(\mathbf{x}_2)$ is assessed by parametrically, then

$$\begin{aligned}\eta(\mathbf{x}) &= h(\mu(\mathbf{x})) \\ &= \alpha_0 + \alpha'_1 \mathbf{x}_1 + \alpha'_2 \mathbf{x}_2,\end{aligned}\quad (4)$$

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

Binomial distribution is given by

$$f(y_i|n, \mu) = \binom{n}{y} \mu^y (1 - \mu)^{n-y}, \quad y_i \in \{0\} \cup \mathbb{Z}^+. \quad (5)$$

Using (2), the term in (5) is explained as

$$\begin{aligned}&= \exp \left[\log \binom{n}{y} + y \log \mu + (n - y) \log (1 - \mu) \right], \\ &= \exp \left[y \log \frac{\mu}{1 - \mu} - (-\log (1 - \mu)) + \log \binom{n}{y} \right],\end{aligned}\quad (6)$$

where

$$\theta = \log \mu / (1 - \mu), \quad b(\theta) = -n \log (1 - \mu), \quad \text{and } c(y_i) = \log \binom{n}{y}.$$

Using (1) and (6), the link function can be written as

$$\begin{aligned}\eta &= \theta \\ &= h(\mu) \\ &= \log \frac{\mu}{(1 - \mu)},\end{aligned}\quad (7)$$

where

$$\mu(\eta) = \log (1 + \exp(\eta)). \quad (8)$$

The log likelihood function of binomial regression [33] is

$$L(\beta; y_i, \mathbf{x}) = \sum_{i=1}^N -\ln (1 + \exp(\mathbf{x}_i \beta) + y_i \mathbf{x}_i \beta). \quad (9)$$

The maximum likelihood estimator for β can be obtained by solving the following system of equations because solution of the system is nonlinear, so the iterative weighted least square procedure is used to estimate the unknown parameters.

The fitted binomial regression model is then given by

$$\begin{aligned}\eta_f(\mathbf{x}|\hat{\mathbf{b}}) &= h(\hat{\mu}_f) \\ &= \frac{\exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)}{1 + \exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)},\end{aligned}\quad (10)$$

where $\hat{\mathbf{b}}' = (\hat{b}_0, \hat{\mathbf{b}}'_1, \hat{\mathbf{b}}'_2)$ and the subscript ' f ' mean on η_f and $\hat{\mu}_f$ denotes the fitted model. The coefficient estimate $\hat{\mathbf{b}}_j$, $j = 0, 1, 2$, is obtained by minimizing the convex objective function:

$$\begin{aligned}\hat{\mathbf{b}}' &= (\hat{b}_0, \hat{\mathbf{b}}'_1, \hat{\mathbf{b}}'_2) \\ &= \arg \min L_N(\hat{\mathbf{b}}),\end{aligned}\quad (11)$$

where

$$\begin{aligned}L_N(\hat{\mathbf{b}}) &= \frac{1}{N} \sum_{i=1}^N L(\eta_f(\mathbf{x}_i|\hat{\mathbf{b}}, y_i)), \\ &= \frac{1}{N} \sum_{i=1}^N L \left(\frac{\exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)}{1 + \exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)}, y_i \right),\end{aligned}\quad (12)$$

where $L(\cdot, \cdot)$ is a user selected objective function assumed to be convex with respect to its first argument. This class is not very restrictive because, at the very least, it includes the use of ordinary least squares and maximum likelihood under (4) and (10) with canonical link, where $\theta = \eta$ and a certain robust estimates. The objective function with respect to maximum likelihood for logistic regression with link given in (7) is

$$L(\eta_f(\mathbf{x}|\hat{\mathbf{b}}, y)) = n \{ \log(1 + \exp(\eta_f)) - y \eta_f \}. \quad (13)$$

The class of objective function corresponding to (11) is a generalization of the class of convex objective function, $L(\eta_f, y) = L(y - \eta_f)$, used by [34] for additive error models (8). A partial residual (PRES2) for $\mathbf{x}_2$ is obtained by using (8)–(10) and is given by

$$\text{PRES2} = (y - \hat{\mu}_f) h'(\hat{\mu}_f) + \hat{\mathbf{b}}'_2 \mathbf{x}_2, \quad (14)$$

where $h'(\cdot)$ is the first derivative of $h(\cdot)$ w.r.t ' μ ' and $\hat{\mathbf{b}}$ can be obtained by (11).

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 first derivative of the binomial link function given in (7) is

$$h'(\hat{\mu}_f) = \frac{1}{(\mu - \mu^2)}. \quad (15)$$

Hence, the fitted model by using link for binomial regression can be expressed as

$$\hat{\mu}_f = \frac{\exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)}{1 + \exp(\hat{b}_0 + \hat{\mathbf{b}}'_1 \mathbf{x}_1 + \hat{\mathbf{b}}'_2 \mathbf{x}_2)}, \quad (16)$$

where $\hat{b}_0$, $\hat{\mathbf{b}}'_1$, and $\hat{\mathbf{b}}'_2$ are the regression estimators and $\hat{\mu}_f$ denote the fitted model and $\mathbf{x}_1$ and $\mathbf{x}_2$ are the predictors. PRES for $\mathbf{x}_2$ is

$$\text{PRES2} = (y - \hat{\mu}_f) \frac{1}{(\mu - \mu^2)} + \hat{\mathbf{b}}_2' \mathbf{x}_2. \quad (17)$$

Similarly, for the model having p explanatory variables, the partial residual can be expressed as

$$\text{PRES}_i = (y - \hat{\mu}_f) \frac{1}{(\mu - \mu^2)} + \hat{\mathbf{b}}_i' \mathbf{x}_i, \quad i = 1, 2, \dots, p, \quad (18)$$

where the fitted model for p explanatory variable is

$$\hat{\mu}_f = \frac{\exp(\hat{b}_0 + \hat{\mathbf{b}}_1' \mathbf{x}_1 + \hat{\mathbf{b}}_2' \mathbf{x}_2 + \dots + \hat{\mathbf{b}}_p' \mathbf{x}_p)}{1 + \exp(\hat{b}_0 + \hat{\mathbf{b}}_1' \mathbf{x}_1 + \hat{\mathbf{b}}_2' \mathbf{x}_2 + \dots + \hat{\mathbf{b}}_p' \mathbf{x}_p)}. \quad (19)$$

The partial residual available in (18) can be notated as RPRES due to response residual. So, the response partial residual is now given as

$$\text{RPRES}_i = (y - \hat{\mu}_f) \frac{1}{(\mu - \mu^2)} + g(\mathbf{x}_i), \quad i = 1, 2, \dots, p, \quad (20)$$

where $g(\mathbf{x}_i) = \hat{\mathbf{b}}_i' \mathbf{x}_i$ for PRES. The RAPRES, RCERES (L), and RCERES (K) can be obtained by using (20).

The augmented PRES (APRES), the conditional expectation, and residuals (CERES) can be obtained by using (20). When we plug quadratic and/or interaction term instead of $g(\mathbf{x}_i)$ in (20), APRES are obtained. Similarly, the conditional expectation ($\mathbf{x}_i | \mathbf{x}_j$) is used instead of $g(\mathbf{x}_i)$ in (20) to obtain CERES. $g(\mathbf{x}_i)$ is sometimes estimated nonparametrically by using LOESS and kernel function [31, 32].

3. Example: Hindered Internal Rotational Data

To demonstrate the applications of partial residual plots in logistic regression, we have taken the numerical example of the hindered internal rotational (HIR) treatment data of chemical species [1]. Having variables is a Boolean feature to check whether two sides are rotating in the same direction (SameDir = Y) as a response, where explanatory variables are out-of-axis angle (AxisAngle = X_1), magnitude of the out-of-axis vibrational vector (AxisMag = X_2), average value of the changes of the dihedral angles in the rotational group (DAngle = X_3), standard deviation of the changes of the dihedral angles in the rotational group (StDAngle = X_4), magnitude of the net vibrational vector of the rotational group (VecMag = X_5), average value of the changes of angle between every pair of atoms (BAngle = X_6), standard deviation of the changes of the angles between every pair of atoms (StBAngle = X_7), average value of the directional angles (DirAngle = X_8), standard deviation of the directional angles (StDirAngle = X_9), average value of the bending angles (StrAngle = X_{10}), standard deviation of the bending angles (StStrAngle = X_{11}), magnitude of the total torque of the current considered rotational group (STorque = X_{12}), magnitude of the total torque of the opposite side with respect to the considered rotational group (Torque = X_{13}), and magnitude of the difference between (X_{12}) and (X_{13}) (TTorque = X_{14}).

The logistic fitted regression model is given by

$$\hat{\mu}_f = \frac{\exp(\hat{b}_0 + \hat{\mathbf{b}}_1' \mathbf{x}_1 + \hat{\mathbf{b}}_2' \mathbf{x}_2 + \dots + \hat{\mathbf{b}}_{14}' \mathbf{x}_{14})}{1 + \exp(\hat{b}_0 + \hat{\mathbf{b}}_1' \mathbf{x}_1 + \hat{\mathbf{b}}_2' \mathbf{x}_2 + \dots + \hat{\mathbf{b}}_{14}' \mathbf{x}_{14})}. \quad (21)$$

Necessary calculations are presented as under by using iterative weighted least square estimation procedure.

By observing AIC, BIC, in Table 1, tells us that logistic regression shows the strong result as compared to other popular distribution because AIC and BIC have the minimum value among the other distribution so that binomial regression is more suitable than Poisson regression and negative binomial regression for this hindered internal rotational (HIR) treatment data of chemical species [1].

Table 2 presents descriptive statistics of the data as well as the summary statistics of all the measures of logistic regression model. In the analysis, it is observed that three variables are significant and other eleven variables are non-significant as all P values are greater than 0.05, but the values of R^2 (0.879) and Adj- R^2 (0.877) are very high which show the importance of these factors. It is well known that the explanatory variables considered here are very important towards in applied science. However, possibly due to data anomalies and deviations, these eleven variables found nonsignificant. So, it is very important to initially recognize the reasons that are responsible for this contradiction. The use of partial residual plots is presented here to serve the stated problem which is computationally less intensive and easy to interpret for a common reader. Figures 1–Figure 4 are available for PRES, APRES, CERES (K), and CERES (L), respectively.

From Figure 1(a), the outliers are clearly observed. It can be seen that there is some value that lies far away to the other data, which reveals an outlier's, and it can also be seen that some points are very close and overlaid to each other, which indicate the presence of multicollinearity. It can also be observed that the problems observed in Figure 1(a) can also be monitored in rest of the plots Figures 1(b)–1(n). Same tendencies are observed in Figures 2–4. So, in the binomial regression model, outlier's and multicollinearity diagnostics can be made by using PRES plots. Similarly, the APRES, CERES (K), and CERES (L) detect outliers. To validate our results and to justify our findings, we are now presenting the outlier's and multicollinearity diagnostic through some formal testing procedures.

To justify our findings, we are now presenting the diagnostics of the abovementioned problems through some formal testing procedures.

In Table 3, the evidence of outlier's (Grubb's test), multicollinearity (F-test), and nonnormality (Anderson–Darling test) is observed. Many other formal tests are available for outlier's diagnostics, and we know that all these tests are based on some regularity conditions and also more intensive, computationally; also, each test is used for single diagnostics only. In the light of the above discussion, we can conclude that a residual plot can be used for outlier's diagnostics and more effective than conventional tests.

TABLE 1: Goodness of fit test.

Distribution	Binomial	Poisson	Negative binomial
AIC	139.28	1632.062	1634.062
BIC	213.699	1637.023	1643.985

AIC, Akaike information criteria; BIC, Bayesian information criteria.

TABLE 2: Binomial regression analysis.

Parameter	β	SE(β)	Z	Pr(> Z)	VIF
Constant	-0.8391	2.4540	0.342	0.73244	
X_1	0.0024	0.0073	0.323	0.74689	1.272
X_2	4.9220	2.7060	1.819	0.06890	1.876
X_3	-0.1554	0.1174	-1.323	0.18590	2.790
X_4	0.1338	0.2038	0.657	0.51141	3.185
X_5	4.3800	3.0240	1.448	0.14755	5.760
X_6	-0.1595	0.1145	-1.393	0.16362	2.470
X_7	-0.2798	0.1770	-1.581	0.11398	1.837
X_8	-0.0444	0.0359	-1.237	0.21595	3.516
X_9	-0.0001	0.0503	-0.001	0.99923	3.889
X_{10}	-0.0028	0.1202	-0.023	0.98190	4.338
X_{11}	-0.1394	0.1351	-1.032	0.30195	2.668
X_{12}	3.2280	1.2490	2.585	0.00974	6.090
X_{13}	3.3000	1.2770	2.584	0.00976	6.761
X_{14}	-3.3000	1.2770	-2.584	0.00976	10.799

Note. $S = 0.520044$, $R^2 = 87.9\%$, and $R^2(\text{adj}) = 87.7\%$; VIF: variance inflation factor.

4. Empirical Evaluation of the Statistics Based on Residuals

The modified pseudolikelihood test based on residuals, partial residuals, augmented partial residual (APRES), and conditional expectation and residual CERES to test linearity and computed empirical powers of the stated statistics [35, 36]. In this study, we are using Grubb's and F-test based on RES, PRES, APRES CERES (L), and CERES (K) for detection of outlier's and multicollinearity, respectively. Powers of the stated statistics are computed and are available in Tables 4 and 5 (outliers) and Tables 6 and 7 (multicollinearity). The empirical power is calculated for $\alpha = 0.01$ and 0.05. Different sample sizes for outliers' detection ($n = 20, 30, 40, 50$) and for multicollinearity ($n = 20, 40, 60, 100$), and four values of ' a ' (2, 10, 20, 95) are chosen.

The binomial regression model is given in (10).

The aim of this section is to assess the performance of residuals (PRES, APRES, and CERES) for outlier and multicollinearity detection. The hypothesis comes to a form as

- (i) H_0 : there are no outliers
- (ii) H_1 : there is at least one outlier

$G_{\text{crit}} > (N - 1/\sqrt{N}) \sqrt{(t_{(\alpha/(2N)), N-2}^2)/(N - 2 + t_{(\alpha/(2N)), N-2}^2)}$, where N is the number of observation with $t_{\alpha/(2N), N-2}$ denoting the upper central value of the t-distribution with $N-2$ degree of freedom and significance level of $\alpha/(2N)$. Grubb's test statistics is

$$G = \frac{\max_i |Y_i - \bar{Y}|}{S}, \quad i = 1, 2, 3, \dots, N. \quad (22)$$

Therefore, we would like to construct the formal test of outliers based on Grubb's test. Statistic large value of G is significant, i.e., $G > G_{\text{crit}}$.

We used F-test for the presence and severity of multicollinearity in a function with several explanatory variables. $F^* = \sum \hat{y}^2/\nu_1 / \sum e^2/\nu_2$ has an F distribution with $\nu_1 = k - 1$ and $\nu_2 = n - k$ degree of freedom. We reject the null hypothesis if $F^* > F_{\text{crit}}$, and there is significant relationship between response and explanatory variables, and however, all explanatory variables have serious multicollinearity.

5. The Performance of Residual Plots for the Assessment of Outlier (s) and Multicollinearity

A small study was carried out to evaluate the diagnostic power of PRES, APRES, CERES (K), and CERES (L) in the presence of outliers and multicollinearity. Data of response variable were generated for the binomial regression model, in which we generate response variable from binomial ($n, 0.5$), and $\mathbf{x}_1$ are generated from a uniform random variable on the interval (0, 30) and $g(\mathbf{x}_1) = a/(1 + e^{-\mathbf{x}_1})$, $\mathbf{x}_2 = \log(\mathbf{x}_1) + N(0, 0.25^2)$, and $\mathbf{x}_3 = \mathbf{x}_1^{-1} + N(0, 1)$. We have fixed the values for the regression parameters as $\beta_1 = 0.05$, $\beta_2 = 0.01$, and $\beta_3 = 0.003$.

We generate outliers in x_{ij} , by replacing the first observation in x 's by $x_{1j} = x_{1j} + 100$ with $Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}$ which are generated from standard normal distribution for fixed degree of multicollinearity (i.e., $\rho^2 = 0.9$). Then, $\mathbf{x}_{ij} = \sqrt{(1 - \rho^2)}Z_{ij} + \rho Z_{ij}(p + 1)$, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$.

FIGURE 1: PRES using response residual in binomial fits.

No error was included so that y is a deterministic function of the three covariates. This allows the conclusion to be illustrated more clearly than when an additive error was included but will not change the qualitative nature of the results. Sample sizes, $n = 20, 30, 40, 50$, are used to address outliers and $n = 40, 60, 80, 100$ are used to address multicollinearity. In each case, 2000 samples were generated and the proportion of times that the observed significances was below $\alpha = 0.05, 0.01$ and was counted. The bandwidth of

kernel function is 0.5 and the number of smoothing data in LOESS is fixed at $n/2$ for obtaining conditional expectation in CERES. The bandwidth of kernel function for obtaining test statistic is 0.25. We choose $a = 2, 10, 20, 95$ in curve $g(\cdot)$ as roughness measure.

5.1. Simulation Result and Discussion. We used the notations, RES, PRES, APRES, CERES (L), and CERES (K), for

FIGURE 2: APRES using response residual in binomial fits.

residual, partial residual, augmented partial residual, conditional expectations residual based on LOESS, and conditional expectations residual based on kernel, respectively. We present the empirical power obtained by the outliers and multicollinearity test statistic in Tables 4 and 5 and Tables 6 and 7, respectively. In Tables 4 and 5 and Tables 6 and 7, it is observed that the test statistic is highly affected by roughness measure ' α '. In Tables 4 and 5, empirical powers of RES and CERES(L) decreased as sample size increased, but the

powers of PRES, APRES, and CERES (K) are increased as sample size increases. It is concluded that test statistic based on PRES, APRES, and CERES (K) is more sensitive to outlier. Similarly, by observing Tables 6 and 7, empirical powers of RES, APRES, and PRES decreased as sample size increased, but the power of CERES (L), CERES (K), and APRES are increased as sample size increased. It is concluded that test statistic based on CERES (L) and CERES (K) is more sensitive to multicollinearity. Thus, we suggest and

FIGURE 3: CERES (K) using response residual in binomial fits.

recommend to use Grubb's and F-test statistic based on residual plots to check the outliers and multicollinearity of selected covariates in the binomial regression model. Figure 5 presents the PRES, APRES, CERES (L), and CERES (K) plots for simulated data. From the evaluation of Figures 5(a)–5(l), we can see that some points lie outside the main course of the data that show outlying cases. Moreover, it can also be observed that some values overlaid each other

and form the congestion and over crowdedness in the plots, which is the evidence of multicollinearity. The joint reflection of outlier and multicollinearity can be seen in PRES, APRES, CERES (L), and CERES (K) plots, which lead to the conclusion that we can detect two data problems in a single shot by the use of stated plots.

Tables 4 and 5 (outliers) and Tables 6 and 7 (multicollinearity) present power computations of five different

FIGURE 4: CERES (L) using response residual in binomial fits.

graphical methods on the basis of Grubb's and F-test for two different level of ' α .' It can be observed from the tables that APRES bear is standing good among others as it has maximum power values for outlier's and CERES (L) performed well has maximum power for multicollinearity. It is also important to mention that the values of ' α ' have notable

impact on the performance of these methods. The impact of sample size is observed in a sense that, with its increment, the power increases; also, some impact of alpha is observed. For the larger values of sample size, a higher value of power is observed for residual plots but APRES and CERES (L) show more consistent behaviour as compared to all other methods.

TABLE 3: Regression diagnostics tests.

Test statistic	Statistic	P value
Grubbs' test	1.2074	0.0005
F-test	5.4567	0.0001
Anderson darling test	198.204	0.0000

TABLE 4: Empirical power of different residuals based on Grubb's test statistic for outlier's detection in binomial regression.

$\alpha = 0.05$						
n	a	RES	PRES	APRES	CERES (L)	CERES (K)
10	2	0.308	0.610	0.852	0.442	0.650
	10	0.344	0.612	0.872	0.426	0.650
	20	0.314	0.674	0.854	0.428	0.584
	95	0.318	0.650	0.878	0.456	0.612
20	2	0.102	0.742	0.938	0.458	0.866
	10	0.082	0.702	0.926	0.462	0.856
	20	0.124	0.752	0.932	0.422	0.866
	95	0.076	0.734	0.932	0.468	0.870
30	2	0.034	0.762	0.954	0.492	0.944
	10	0.018	0.792	0.936	0.502	0.938
	20	0.026	0.750	0.936	0.522	0.924
	95	0.038	0.776	0.940	0.510	0.936
40	2	0.000	0.786	0.944	0.588	0.996
	10	0.002	0.818	0.948	0.636	0.984
	20	0.002	0.784	0.940	0.586	0.990
	95	0.000	0.796	0.946	0.596	0.990

TABLE 5: Empirical power of different residuals based on Grubb's test statistic for outlier's detection in binomial regression.

$\alpha = 0.01$						
n	a	RES	PRES	APRES	CERES (L)	CERES (K)
10	2	0.018	0.534	0.822	0.342	0.600
	10	0.024	0.540	0.806	0.292	0.590
	20	0.014	0.550	0.788	0.290	0.602
	95	0.016	0.568	0.794	0.310	0.582
20	2	0.000	0.684	0.876	0.352	0.830
	10	0.001	0.698	0.894	0.386	0.878
	20	0.000	0.686	0.892	0.360	0.838
	95	0.002	0.670	0.902	0.420	0.832
30	2	0.003	0.738	0.932	0.510	0.950
	10	0.004	0.746	0.920	0.502	0.940
	20	0.001	0.728	0.900	0.476	0.946
	95	0.000	0.716	0.904	0.502	0.940
40	2	0.000	0.704	0.902	0.584	0.944
	10	0.000	0.848	0.906	0.516	0.972
	20	0.005	0.862	0.904	0.500	0.976
	95	0.001	0.976	0.974	0.500	0.978

TABLE 6: Empirical power of different residuals based on F-test statistic for multicollinearity detection in binomial regression.

$\alpha = 0.05$						
n	a	RES	PRES	APRES	CERES (L)	CERES (K)
20	2	0.001	0.615	0.652	0.744	0.641
	10	0.002	0.615	0.652	0.754	0.655
	20	0.040	0.576	0.654	0.756	0.714
	95	0.018	0.576	0.678	0.756	0.716
40	2	0.122	0.746	0.770	0.858	0.766
	10	0.132	0.751	0.771	0.862	0.756
	20	0.144	0.753	0.732	0.872	0.766
	95	0.176	0.763	0.832	0.878	0.780

TABLE 6: Continued.

$\alpha = 0.05$						
n	a	RES	PRES	APRES	CERES (L)	CERES (K)
60	2	0.184	0.764	0.854	0.882	0.844
	10	0.212	0.794	0.856	0.882	0.858
	20	0.226	0.795	0.856	0.898	0.874
	95	0.238	0.796	0.970	0.910	0.886
100	2	0.256	0.796	0.974	0.918	0.980
	10	0.257	0.878	0.978	0.936	0.884
	20	0.257	0.894	0.988	0.995	0.995
	95	0.334	0.896	0.996	0.996	0.990

TABLE 7: Empirical power of different residuals based on F-test statistic for multicollinearity detection in binomial regression.

$\alpha = 0.01$						
n	a	RES	PRES	APRES	CERES (L)	CERES (K)
20	2	0.000	0.623	0.656	0.744	0.670
	10	0.001	0.625	0.657	0.754	0.673
	20	0.002	0.676	0.658	0.756	0.714
	95	0.008	0.676	0.638	0.856	0.732
40	2	0.023	0.741	0.739	0.858	0.766
	10	0.182	0.751	0.747	0.865	0.756
	20	0.154	0.753	0.892	0.874	0.766
	95	0.176	0.764	0.832	0.878	0.780
60	2	0.178	0.764	0.854	0.881	0.844
	10	0.216	0.791	0.866	0.886	0.858
	20	0.128	0.795	0.876	0.992	0.874
	95	0.132	0.796	0.878	0.912	0.886
100	2	0.146	0.796	0.941	0.918	0.990
	10	0.150	0.878	0.995	0.936	0.984
	20	0.151	0.894	0.948	0.996	0.995
	95	0.151	0.796	0.946	0.998	0.990

FIGURE 5: Continued.

FIGURE 5: Residual plots using simulation data in binomial fits.

6. Conclusion

Residual plots are the data mining approach that solves existing as well as future challenges of approach based on rule for the determination of the rotational frequency (i.e., the most significant so far complex parameter of the hindered internal rotation) interested of chemical species. They explored the partial residual plots in GLM and checked the violation of assumption of regression using inverse Gaussian regression [32]. They investigated the condition in which the partial residual plots provided the useful transformation of the predictor in the class of GLM by using response residual in the binomial regression model [31].

This study articulates the construction and evaluation of RES, PRES APRES, CERES (L), and CERES (K) for hindered internal rotational (HIR) treatment data of chemical species [1] plots for binomial regression by calculating the power of the tests. It is observed that APRES and CERES (L) plots are fantastic techniques for influence diagnostics. So, these plots are recommended for use for detection of outlier's and multicollinearity. The usefulness of these plots and tests for obtaining best results may be restricted to the specified GLM, the link function, and the stochastic behaviour of the predictors.

Results are also validated by the simulation study with diagnostic power; it is observed that residual plots are fantastic technique for outlier's and multicollinearity diagnostics, while identifying the various problems in a single shot. So, it is concluded that residual plots are useful graphical technique for the outlier and multicollinearity diagnostics in a single graph in chemical processes.

To deal with the applied networks science, scientists usually expect the relationship of certain elements and

variables on the basis of natural phenomena. However, when statistical methodologies are applied, some unanticipated results arise. This is due to the reasons that (i) the wrong applications of statistical methodology and (ii) contradictory and misleading assumptions for data under the study. Therefore, in later case, initially diagnostics for assumptions are made and then particular statistical methodology is applied after suitable transformations [16]. In our case, two different techniques produce two opposite results, i.e., P values show unimportance of underlying factors while large values of R^2 (0.879) and $\text{Adj} - R^2$ (0.877) show great importance of these factors. So, it very necessary that diagnostic checking should be made before proceeding further. Moreover, the diagnostic checking is usually made by different statistical tests and one test is dedicated to one diagnostic checking. So, it is computationally intensive and also requires strong background of statistical knowledge. In spite of this, the study suggests a use of residual plots for diagnostic checking in a very simple way. These plots also have capability to diagnose multiple problems on a single graph. This will help users to assess their data anomalies initially by using these plots and then proceed further in their planning and making policies.

Data Availability

The numerical example of the hindered internal rotational (HIR) treatment data of chemical species was taken (Le et al., 2018). The data used to support the findings of the study can be obtained from the references cited within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] T. H. M. Le, T. T. Tran, and L. K. Huynh, "Identification of hindered internal rotational mode for complex chemical species: a data mining approach with multivariate logistic regression model," *Chemometrics and Intelligent Laboratory Systems*, vol. 172, pp. 10–16, 2018.
- [2] T. H. M. Le, S. T. Do, and L. K. Huynh, "Algorithm for auto-generation of hindered internal rotation parameters for complex chemical systems," *Computational and Theoretical Chemistry*, vol. 1100, pp. 61–69, 2017.
- [3] A. Y. Ben-Naim, *Statistical Thermodynamics for Chemists and Biochemists*, Springer Science & Business Media, Berlin, Germany, 2013.
- [4] A. F. Lukman, K. Ayinde, S. Binuomote, and O. A. Clement, "Modified ridge-type estimator to combat multicollinearity: application to chemical data," *Journal of Chemometrics*, vol. 33, no. 5, Article ID e3125, 2019.
- [5] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, London, United Kingdom, 1989.
- [6] M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, John Wiley & Sons, NJ, USA, 2016.
- [7] D. J. Marchette, "Random graphs for statistical pattern recognition," *Wiley Series in Probability and Statistics*, John Wiley & Sons, vol. 565, NJ, USA, 2005.
- [8] W. G. Jacoby, "Statistical graphics for univariate and bivariate data," *Quantitative Applications in the Social Sciences*, Sage, vol. 117, CA, USA, 1997.
- [9] P. Stefansson, U. G. Indahl, K. H. Liland, and I. Burud, "Orders of magnitude speed increase in partial least squares feature selection with new simple indexing technique for very tall data sets," *Journal of Chemometrics*, vol. 33, no. 11, Article ID e3141, 2019.
- [10] G. C. Fernandez, "Detection of model specification, outlier, and multicollinearity in multiple linear regression models using partial regression/residual plots," in *SAS Conference Proceedings*, vol. 22, pp. 16–19, 1997.
- [11] W. A. Larsen and S. J. McCleary, "The use of partial residual plots in regression analysis," *Technometrics*, vol. 14, no. 3, pp. 781–790, 1972.
- [12] R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman & Hall, London, UK, 1982.
- [13] M. Ezekiel, "A method of handling curvilinear correlation for any number of variables," *Journal of the American Statistical Association*, vol. 19, no. 148, pp. 431–453, 1924.
- [14] C. L. Huynh, "Using partial residual plots in assessing and improving the construct validity of multiple regression models," *Multiple Linear Regression Viewpoints*, vol. 26, 2000.
- [15] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics*, John Wiley, NJ, USA, 1980.
- [16] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, NJ, USA, 2018.
- [17] P. F. Velleman and R. E. Welsch, "Efficient computing of regression diagnostics," *The American Statistician*, vol. 35, no. 4, pp. 234–242, 1981.
- [18] S. Chatterjee and A. S. Hadi, "Influential observations: higher leverage points and outliers in linear regression," *Statistical Science*, vol. 1, no. 3, pp. 379–393, 1986.
- [19] N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley, NJ, USA, 3rd edition, 1998.
- [20] F. Vogt, *Applied Chemometrics for Scientists*, R. G. Brereton, Ed., John Wiley & Sons, NJ, USA, 2007.
- [21] R. G. Brereton, "Chemometrics and Multivariate Analysis," *Digital Encyclopedia Of Applied Physics*, pp. 1105–1148, 2009.
- [22] P. Watkins and B. Venables, "Non-linear regression for optimising the separation of carboxylic acids," *The Newsletter of the R Project*, vol. 6, p. 3, 2006.
- [23] Z. Y. Algamal, "Diagnostic in Poisson regression models," *Electronic Journal of Applied Statistical Analysis*, vol. 5, no. 2, pp. 178–186, 2012.
- [24] M. Korkmaz, S. Guney, and S. Yigiter, "The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields," *Har-ran Tarım ve Gıda Bilimleri Dergisi*, vol. 16, no. 2, pp. 25–36, 2012.
- [25] Z. Y. Algamal and M. H. Lee, "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.
- [26] Z. Y. Algamal and M. H. Lee, "High dimensional logistic regression model using adjusted elastic net penalty," *Pakistan Journal of Statistics and Operation Research*, vol. 11, no. 4, pp. 667–676, 2015.
- [27] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014.
- [28] S. Luo, X. Su, S. M. DeSantis, X. Huang, M. Yi, and K. K. Hunt, "Joint model for a diagnostic test without a gold standard in the presence of a dependent terminal event," *Statistics in Medicine*, vol. 33, no. 15, pp. 2554–2566, 2014.
- [29] M. Díaz-Pérez, Á. Carreño-Ortega, J.-A. Salinas-Andújar, and Á.-J. Callejón-Ferre, "Application of logistic regression models for the marketability of cucumber cultivars," *Agronomy*, vol. 9, no. 1, p. 17, 2019.
- [30] J. M. Landwehr, D. Pregibon, and A. C. Shoemaker, "Graphical methods for assessing logistic regression models," *Journal of the American Statistical Association*, vol. 79, no. 385, pp. 61–71, 1984.
- [31] R. D. Cook and R. Croos-dabrera, "Partial residual plots in generalized linear models," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 730–739, 1998.
- [32] M. Imran and A. Akbar, "Diagnostics via partial residual plots in inverse Gaussian regression," *Journal of Chemometrics*, vol. 34, no. 1, 2020.
- [33] J. Hilbe, *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK, 2011.
- [34] R. D. Cook, "Exploring partial residual plots," *Technometrics*, vol. 35, no. 4, pp. 351–362, 1993.
- [35] A. Azzalini and A. Bowman, "On the use of nonparametric regression for checking linear relationships," *Journal of the Royal Statistical Society: Series B*, vol. 55, no. 2, pp. 549–557, 1993.
- [36] K. S. Oh, "Regression diagnostics using residual plots," *Communications for Statistical Applications and Methods*, vol. 8, no. 2, pp. 311–317, 2001.