Research Article

Identification of Multiple Outliers in a Generalized Linear Model with Continuous Variables

Loo Yee Peng,¹ Habshah Midi,^{1,2} Sohel Rana,^{1,2} and Anwar Fitrianto^{1,2}

¹Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

²Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Correspondence should be addressed to Habshah Midi; habshahmidi@gmail.com

Received 18 April 2016; Accepted 8 August 2016

Academic Editor: M.I. Herreros

Copyright © 2016 Loo Yee Peng et al. 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.

In the statistical analysis of data, a model might be awfully fitted with the presence of outliers. Besides, it has been well established to use residuals for identification of outliers. The asymptotic properties of residuals can be utilized to contribute diagnostic tools. However, it is now evident that most of the existing diagnostic methods have failed in identifying multiple outliers. Therefore, this paper proposed a diagnostic method for the identification of multiple outliers in GLM, where traditionally used outlier detection methods are effortless as they undergo masking or swamping dilemma. Hence, an investigation was carried out to determine the capability of the proposed GSCPR method. The findings obtained from the numerical examples indicated that the performance of the proposed method was satisfactory for the identification of multiple outliers. Meanwhile, in the simulation study, two scenarios were considered to assess the validity of the proposed method. The proposed method consistently displayed higher percentage of correct detection, as well as lower rates of swamping and masking, regardless of the sample size and the contamination levels.

1. Introduction

Generalized linear model (GLM) is a continuation of the familiar linear regression model for modeling a nonnormal response variable [1]. In the statistical analysis of data, the model might be awfully fitted with the presence of outliers. In fact, any individual observation that appears to depart in some way from the remainder of that set of data is called an outlier [2]. Hence, the identification of outliers is a necessary step to obtain appropriate results in GLM [3]. Moreover, it has been well established to make use of residuals for the identification of outliers [4]. However, there is evidence that the maximum likelihood estimates for GLM most probably get distorted when the sample size, n , of the data set is small [5]. The distribution of residuals, sometimes, differs from the distribution of the true residuals to the order n^{-1} , where n is the sample size. Furthermore, the asymptotic properties of residuals for the selected regression model are available in the literature. For instance, Cordeiro and McCullagh [5] derived the formulae for first-order biases of maximum likelihood

estimates of linear parameters, linear predictors, dispersion parameter, and fitted values in GLM. Cordeiro [6] defined the adjusted Pearson residuals where the mean and the variance are approximately zero and one. Later, Cordeiro and Simas [7] obtained an explicit formula for the density of the Pearson residuals to order n^{-1} , which hold for all continuous GLM and defined corrected residuals for these models. Recently, Anholeto et al. [8] acquired the matrix formula of order n^{-1} for the first two moments of Pearson residuals and adjusted Pearson residuals in beta regression models. These asymptotic properties of residuals can be utile to contribute as diagnostic tools [7].

However, it is now evident that a majority of the existing diagnostic methods are inadequate in identifying multiple outliers. A single case diagnostic method cannot correctly identify outliers if multiple outliers exist in a data set. The reasons are that a group of outliers is able to distort the fitting of a model as the outliers can have artificially tiny residuals that appear as inliers [9]. This type of troublesomeness is cognized as the masking effect. The opposite effect of masking

is swamping, where the inliers may resemble the outliers. Thus, the outlier detection methods in GLM sustain from masking or swamping effects.

In addition, several published papers have considered the detection of multiple outliers. For example, Lee and Fung [10] proposed an adding-back procedure with an initial high breakdown point for the detection of multiple outliers in GLM and nonlinear regressions. On the other hand, Imon and Hadi [9] proposed a generalized version of standardized Pearson residuals based on group deletion method (GSPR) to overcome the difficulty of multiple outliers detection in logistic regression.

These have motivated the researchers to modify the corrected Pearson residuals [7] to adapt to the problems related to multiple outliers. Hence, this paper presents the development of a method for the identification of multiple outliers by employing corrected Pearson residuals based on group deletions. This method is called Generalized Standardized Corrected Pearson Residuals (GSCPR). Furthermore, a few frequently used diagnostics related to residuals have been briefly discussed for the identification of outliers in Section 2. In Section 3, the proposed GSCPR method is described for the identification of multiple outliers. Next, the usefulness of this proposed method is examined in Section 4 via a real data set, and finally, Section 5 reports the Monte Carlo simulations.

2. Residuals in GLM

As mentioned earlier, it is often very important to be aware of the existence of outliers in a data set. The outliers have been found to tremendously influence the covariate pattern, and hence, their existence may mislead the interpretation of the statistical analysis. The deviance residuals and the Pearson residuals are two common types of residuals in GLM. In this paper, the focus was only on the Pearson residuals. Hilbe and Robinson [11] pointed out that, by normalizing Pearson or deviance residuals to a standard deviation of 1.0, the entire deviance residual tends to normalize the residual better than the Pearson residual. Although deviance residual is the favorable statistics in GLM at the time the study was conducted, more tests developed based on the Pearson residuals are essential. The Pearson residuals are defined as $r_i = (y_i - \hat{\pi}_i)/\sqrt{v_i}$, $i = 1, 2, \dots, n$, where $\hat{\pi}_i$ is the i th fitted values and v_i is variance function. Meanwhile, the standardized Pearson residuals are defined as $r_{s_i} = (y_i - \hat{\pi}_i)/\sqrt{v_i(1 - h_i)}$, $i = 1, 2, \dots, n$, where $\hat{\pi}_i$ is the i th fitted values, v_i is the variance function, and hat-values h_i for GLM is the i th diagonal element of the $n \times n$ matrix $H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2}$, where V is a diagonal matrix with diagonal elements v_i . $X_i^T = [1, x_{1i}, x_{2i}, \dots, x_{pi}]$ is the $1 \times k$ vector of observations corresponding to the i th case. Furthermore, an observation is stated as a residual outlier when its corresponding r_i or r_{s_i} exceeds a quantity c in absolute term. A popular and well-reasoned choice for c could be 3 as it matches the three-sigma distance rule practice in the normal theory [12]. Another suitable constant between 3

and 5 is considered if the cutoff value 3 identifies too many observations as outliers [13].

In general, the distribution of the Pearson residuals deviates from its true distribution by terms of order n^{-1} . In addition, the mean and the variance of its true distribution are zero and one, respectively. Cordeiro [6] claimed that the adjusted Pearson residuals have closer distribution to standard normal distribution compared to the corresponding moments of the unmodified residuals. The adjusted Pearson residual corrects the residuals to equal mean and variance, but its distribution is not equivalent to the distribution of the true Pearson residuals to order n^{-1} . Hence, the corrected Pearson residuals were proposed to remedy this problem. According to Cordeiro and Simas [7], it is essential to discern the ϕ - and the n -asymptotic in GLM. The n -asymptotic was applied in this method; that is, the dispersion parameter, ϕ , is fixed, whereas the size of the number of observations, n , becomes large. This asymptotic theory is widely concerned about estimation and hypothesis testing with regard to the unknown parameter β . Nonetheless, these methods are restricted to data set with continuous distributions, such as normal, gamma, and inverse Gaussian distribution. The corrected Pearson residuals for continuous GLMs are defined as $r'_i = r_i + \rho_i(r_i)$, where r_i is the Pearson residuals and $\rho(\cdot)$ is a function of order $O(n^{-1})$ created to produce residual r'_i with the same distribution of ε_i to order n^{-1} . The correction function is equivalent to

$$\begin{aligned} \rho_i(x) &= e_i(x) \left\{ -\frac{1}{2\phi} V_i^{-1} V_i^{(1)} \mu'_i z_{ii} - B(\hat{\eta}_i) - w_i^{1/2} z_{ii} x \right\} \\ &\quad - \frac{z_{ii}}{2\phi} h_i(x) \\ &\quad + \frac{z_{ii}}{2\phi} e_i(x)^2 \left\{ \phi \sqrt{V_i} q(\mu_i) + \frac{d}{dx} c \left(\sqrt{V_i} x + \mu_i, \phi \right) \right\}. \end{aligned} \quad (1)$$

Furthermore, the term $\phi^{-1} z_{ii}$ in the equation is equivalent to $\text{Var}(\hat{\eta}_i)$. This expression can be easily implemented to any continuous model as only $e_i(x)$, $h_i(x)$, and $(d/dx)c(\sqrt{V_i}x + \mu_i, \phi)$ need to be calculated. The equations for the correction $\rho_i(\cdot)$ for some important GLM are shown in the work conducted by Cordeiro and Simas [7]. Besides that, the values such as μ' , μ'' for several useful link functions, $q(\mu)$, V , w , and $(d/dx)c(\sqrt{V_i}x + \mu_i, \phi)$ for various distributions, are shown by Cordeiro and Simas [7]. Meanwhile, for normal models, $V_i = 1$, $w_i = \mu_i'^2$, $c(x, \phi) = -\{x^2\phi + \log(2\pi/\phi)\}/2$, $e_i(x) = -\mu_i'$ and $h_i(x) = -\mu_i''$. There is also $\rho_i(x) = B(\hat{\eta}_i)\mu_i' + \mu_i'' z_{ii}/2\phi + (\mu_i'^2 z_{ii}/2)x$. As for gamma models, $V_i = \mu_i^2$, $w_i = \mu_i^{-2} \mu_i'^2$, $c(x, \phi) = (\phi - 1) \log(x) + \phi \log(\phi) - \log \Gamma(\phi)$ and $(d/dx)c(\mu x + \mu, \phi) = (\phi - 1)/(1 + x)$. There are also $e_i(x) = -\mu_i^{-1} \mu_i'$ - $\mu_i^{-1} \mu_i' x$ and $h_i(x) = -\mu_i^{-1} \mu_i'' + 2\mu_i^{-2} \mu_i'^2 - \mu_i^{-1} \mu_i'' x + 2\mu_i^{-2} \mu_i'^2 x$.

Then,

$$\rho_i(x) = (1+x) \left(\mu_i^{-1} \mu_i' B(\hat{\eta}_i) + \frac{\mu_i^{-1} \mu_i''}{2\phi} z_{ii} - \frac{\mu_i^{-2} \mu_i'^2}{2\phi} z_{ii} + \frac{\mu_i^{-2} \mu_i'^2 z_{ii}}{2} x \right). \quad (2)$$

Furthermore, for inverse Gaussian models, $V_i = \mu_i^3$, $w_i = \mu_i^{-3} \mu_i'^2$, $c(x, \phi) = (1/2) \log\{\phi/(2\pi x^3)\} - \phi/(2x)$ and $(d/dx)c(\mu^{3/2}x + \mu, \phi) = -3\mu^{3/2}/2(\mu^{3/2}x + \mu) + \phi\mu^{3/2}/2(\mu^{3/2}x + \mu)^2$.

Next, $e_i(x) = -\mu_i^{-3/2} \mu_i' - (3/2)\mu_i^{-1} \mu_i' x$ and $h_i(x) = -\mu_i^{-3/2} \mu_i'' + 3\mu_i^{-5/2} \mu_i'^2 + (15/4)\mu_i^{-2} \mu_i'^2 x - (3/2)\mu_i^{-1} \mu_i'' x$.

Then,

$$\begin{aligned} \rho_i(x) = & \left(\mu_i^{-3/2} \mu_i' + \frac{3\mu_i'}{2\mu_i} x \right) B(\hat{\eta}_i) + \frac{\mu_i^{3/2} \mu_i'' z_{ii}}{2\phi} \\ & + \left(\frac{3\mu_i'^2 z_{ii}}{8\phi\mu_i^2} + \frac{3\mu_i'' z_{ii}}{4\phi\mu_i} \right) x + \frac{\mu_i'^2 z_{ii}}{\mu_i^3} x + \frac{3\mu_i'^2 z_{ii}}{2\mu_i^{5/2}} x^2 \\ & + \frac{\mu_i^{-3} z_{ii}}{4\phi} \left(\mu_i'^2 + 3\mu_i^{1/2} \mu_i'^2 x + \frac{9\mu_i \mu_i'^2}{4} x^2 \right) \\ & \cdot \left\{ -\frac{\phi}{\mu_i^{1/2}} - \frac{3\mu_i^{3/2}}{(\mu_i^{3/2}x + \mu_i)} + \frac{\phi\mu_i^{3/2}}{(\mu_i^{3/2}x + \mu_i)^2} \right\}. \end{aligned} \quad (3)$$

Moreover, the QQ plot was employed for the quantiles of the corrected Pearson residuals against the quantiles of the estimated shifted gamma distribution to detect outliers. The corrected Pearson residuals were found to be superior than the uncorrected Pearson residuals in discerning the discrepancies among the fitted model and the data.

3. Identification of Multiple Outliers by Using GSCPR

The diagnostics of multiple outliers are crucial because it is difficult to ensure that the data set possesses only single outlier in real life problem. The explanation from above is comprehensive and emphasized that the diagnostic methods for the identification of multiple outliers are influenced by masking or swamping effect. For identification of multiple outliers, a group-deleted version of the residuals is used to develop effective diagnostics. By assuming that d observations in a set of n observations are excluded, hence, only $(n-d)$ cases of observations are considered for model fittings. Denoting a set of “remaining” observations in the study by R and a set of “deleted” observations by D , it is assumed that the “deleted” observations are the last of d rows of X , Y , and V so that $X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}$, $Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}$, and $V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}$. Let $\hat{\beta}^{(-D)}$ be the corresponding vector of the estimated coefficients when a group of observations indexed by D is excluded. The corrected Pearson residuals for GLMs are defined as $r_i' = r_i + \rho_i(r_i)$; thus, the i th deletion of the corrected Pearson

residuals is $\hat{r}_i^{(-D)} = \hat{r}_i^{(-D)} + \rho_i(\hat{r}_i^{(-D)})$, $i = 1, 2, \dots, n$, where $\hat{r}_i^{(-D)} = (y_i - \hat{\pi}_i^{(-D)})/\sqrt{v_i^{(-D)}}$, $\hat{\pi}_i^{(-D)}$ is the i th deletion for the fitted values and $\rho(\cdot)$ is a function of order $O(n^{-1})$. The respective deletion variances, $v_i^{(-D)}$ (for gamma distribution), and deletion leverages, $h_i^{(-D)}$, for the overall data set are defined as

$$\begin{aligned} v_i^{(-D)} &= \hat{\pi}_i^{2(-D)}, \\ h_i^{(-D)} &= x_i^T (X_R^T V_R X_R)^{-1} x_i. \end{aligned} \quad (4)$$

On the other hand, concerning logistic regression, Imon and Hadi [9] defined the i th GSPR as

$$\begin{aligned} t_i^{(-D)} &= \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)} (1 - h_i^{(-D)})}} \quad \text{for } i \in R, \\ &= \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)} (1 + h_i^{(-D)})}} \quad \text{for } i \in D. \end{aligned} \quad (5)$$

The GSPR were obtained by using the principles of the scaled types residuals and linear regression like approximation modulus. According to Imon and Hadi [9], the implantation of the scaled types residuals principle enabled the residuals $t_i^{(-D)}$ for the R set and D set to be measured on a similar scale.

In order to obtain the standardized corrected Pearson residuals, slight modifications were made to the standardized Pearson residuals; that is, $r_{s_i}' = (r_i + \rho_i(r_i))/\sqrt{1 - h_i}$, where r_i is the Pearson residuals. Thus, the standardized corrected Pearson residuals are also defined as $r_{s_i}' = (y_i - \hat{\pi}_i)/\sqrt{v_i(1 - h_i)} + \rho_i((y_i - \hat{\pi}_i)/\sqrt{v_i})/\sqrt{1 - h_i}$ or equivalent to $r_{s_i}' = (y_i - \hat{\pi}_i + \sqrt{v_i}\rho_i((y_i - \hat{\pi}_i)/\sqrt{v_i}))/\sqrt{v_i(1 - h_i)}$.

By using the principles of the scaled types residuals, the i th GSCPR for GLM is defined as

$$\begin{aligned} r_{\text{GSCPR}_i}^{(-D)} &= \frac{y_i - \hat{\pi}_i^{(-D)} + \sqrt{v_i^{(-D)}} \rho_i\left(\frac{(y_i - \hat{\pi}_i^{(-D)})}{\sqrt{v_i^{(-D)}}}\right)}{\sqrt{v_i^{(-D)} (1 - h_i^{(-D)})}} \quad \text{for } i \in R, \\ &= \frac{y_i - \hat{\pi}_i^{(-D)} + \sqrt{v_i^{(-D)}} \rho_i\left(\frac{(y_i - \hat{\pi}_i^{(-D)})}{\sqrt{v_i^{(-D)}}}\right)}{\sqrt{v_i^{(-D)} (1 + h_i^{(-D)})}} \quad \text{for } i \in D. \end{aligned} \quad (6)$$

Therefore, the GSCPR method is summarized in the following.

Step 1. For each i point, calculate the corrected Pearson residuals, r_i' .

Step 2. An i th point with r_i' exceeding the cutoff point of value ± 3 is suspected as outlier. These points are taken into

consideration to be assigned to the deleted set D . The rest of the points are assigned to R set.

Step 3. Calculate the $r_{\text{GSCPR}_i}^{(-D)}$ values based on the decided D and R sets in Step 2.

Step 4. Any deleted points in accordance with $r_{\text{GSCPR}_i}^{(-D)}$ exceeding value 3 in absolute term are finalized and declared as the outlier.

The choice of deleted observations played a pivotal role in the method as the exclusion of this group decided the residuals resulting from the D set and the R set. Besides, all suspected outliers were included in the initial deletion set, since the entire GSCPR set would have been faulty if any outlier was left in the R set. Meanwhile, in logistic regression, Imon and Hadi [9] suggested a diagnostic-robust approach, which employed graphical procedures or robust procedures to identify suspect outliers at the early stage. The next stage employed the diagnostic tools to the resulting residuals so as to assign the inliers (if any), which were incorrectly identified as outliers at the early stage, back into the estimation subset. In this study, corrected Pearson residuals were applied to detect the suspected outliers in the early stage, and then, the suspected points were declared as outliers if every member subset in set D fulfilled the rule given in Step 4. Otherwise, the observations were placed back to R set. Moreover, the deletion of set D was continuously inspected by recalculating the GSCPR until every member in the final deletion set individually fulfilled the rule outlined in Step 4. The member subset in this final set D was, eventually, declared as outliers.

4. Example Using Real Data Set

A real data set was used to illustrate the capability of the newly proposed tool for the identification of multiple outliers. The drill experiment comprised a 2^4 unreplicated factorial in order to investigate the response variable advance rate. This example was reanalyzed by using a generalized linear model developed by Lewis and Montgomery [14]. As for the GLM analysis, the gamma distribution was selected with a log link function. Table 1 displays the design matrix and the response data. In order to observe the effect of outliers, the original response data had been modified. For the single outlier case, the 16th observation was replaced with a value of 36.30 instead of 16.30. Meanwhile, for the case of two outliers, the 13th and 16th observations were replaced with values 27.77 and 36.30, respectively. The results of the Pearson residuals, corrected Pearson residuals, and GSCPR measures are shown in Table 2.

Table 2 and index plots of Pearson residuals, as given in Figures 1(a) and 1(b), show that the Pearson residuals method failed in identifying the outliers. For data set with an outlier, as observed in Figures 1(c) and 1(e), the values for both corrected Pearson residuals and GSCPR for observation 13 had been remarkably high and, therefore, could be revealed as outlier. Meanwhile, for the data set with two outliers, as depicted in Figure 1(d), the corrected Pearson residuals

TABLE 1: Drill experiment data for design matrix and response data.

Run	X_1	X_2	X_3	X_4	Y
1	-1	-1	-1	-1	1.68
2	1	-1	-1	-1	1.98
3	-1	1	-1	-1	3.28
4	1	1	-1	-1	3.44
5	-1	-1	1	-1	4.98
6	1	-1	1	-1	5.70
7	-1	1	1	-1	9.97
8	1	1	1	-1	9.07
9	-1	-1	-1	1	2.07
10	1	-1	-1	1	2.44
11	-1	1	-1	1	4.09
12	1	1	-1	1	4.53
13	-1	-1	1	1	7.77 (27.77)
14	1	-1	1	1	9.43
15	-1	1	1	1	11.75
16	1	1	1	1	16.30 (36.30)

method correctly identified observations 13 and 16 as outliers but swapped a good observation (observation 15) as an outlier. Nevertheless, Figure 1(f) shows that the proposed GSCPR correctly identified observations 13 and 16 as outliers in the data set.

5. Simulation Results

In this section, simulation studies based on 2^k factorial design data and explanatory variable data were conducted to verify the conclusion of the numerical data set that the proposed method had been indeed capable of correctly identifying multiple outliers. The measures of Pearson residuals, corrected Pearson residuals, and GSCPR measures were reported. Besides, only the model that adhered to Gamma distribution and log link function had been considered. Furthermore, two different scenarios were considered in this study. In the first scenario (Scenario 1), a 2^k factorial design where $k = 2$ with replication of 8, 16, 32, and 64 runs had been applied. For each experimental design condition, a model was set up with known parameters and a known true mean, as suggested by Lewis et al. [15]. On top of that, the true linear predictor, η_i , was created as $\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$, where the true linear predictor was taken as $\eta_i = 0.5 + x_1 - x_2$, $\phi = 4$ for 2^2 factorial design [7]. The true mean was defined as $\mu_i = g^{-1}(\eta_i)$, where g is the link function and $i = 1, 2, \dots, n$. The actual observation was acquired by attaching an error drawn at arbitrary from a specified distribution to the linear predictor; that is, $y_i = g^{-1}(\eta_i) + \varepsilon_i$, where $i = 1, 2, \dots, n$, while residuals ε_i were induced from a Gamma distribution, with a shape equivalent to 0.1 and a scale that equaled one. Furthermore, in order to generate 8 run designs' matrix, two replicates of the 2^2 factorial design were used, four replicates gave 16 run designs matrix, and so on. The outlying observations were generated from the model $y_i = g^{-1}(\mu_i) + y_{\text{shift}} + \varepsilon_i$, where $i = 1, 2, \dots, n$

TABLE 2: Outlier measures based on GLM for the drill experiment data with single outlier and two outliers.

Run	One outlier			Two outliers		
	Pearson residuals	Corrected Pearson residuals	GSCPR	Pearson residuals	Corrected Pearson residuals	GSCPR
1	0.2215	0.2046	2.6815	0.0691	-0.0111	2.8615
2	0.1115	0.0233	2.2698	0.2246	0.1881	2.5292
3	0.1642	0.0648	2.1541	0.2664	0.2441	2.2357
4	-0.0571	-0.3568	1.6509	0.2908	0.2925	1.9941
5	0.0034	-0.3083	1.8457	-0.2844	-1.1473	1.5493
6	-0.1132	-0.6985	1.7087	-0.2040	-1.0058	1.6476
7	-0.0192	-0.5886	2.9959	-0.1309	-1.2699	2.5177
8	-0.3111	-2.2933	1.3435	-0.2316	-1.7617	1.4277
9	-0.0378	-0.2063	1.4597	-0.3205	-0.6520	1.4446
10	-0.1242	-0.3826	1.3911	-0.2216	-0.5477	1.4168
11	-0.0718	-0.4398	1.6409	-0.1855	-0.6065	1.4863
12	-0.2062	-0.8296	1.4652	-0.1233	-0.6065	1.4693
13	0.0008	-0.4118	2.5986	1.0577	16.7608	18.790
14	-0.0620	-0.8023	2.7891	-0.3208	-2.4704	2.3355
15	-0.2611	-2.6258	1.7722	-0.4717	-4.9425	2.0991
16	0.7625	19.1545	26.1150	0.5859	15.3120	21.2990

and y_{shift} represents the distance of the outliers that were placed away from good observations. In this study, y_{shift} was taken as 10. The residual outliers were created in the first n' observation, where n' is the total number of the outlying observations. The 5,000 replications of the simulation study are summarized in Table 3. This table presents the percentage of correct detection of outliers, masking rates, and swamping rates for 2^2 factorial design. In the second scenario (Scenario 2), a simulation study was designed for a set of explanatory variables data, that is, two explanatory variables. The true linear predictor was taken as $\eta_i = 0.5 + x_1 - x_2$, $\phi = 4$, by adhering to that suggested by Cordeiro and Simas [7]. The true mean and the actual observation were obtained as described in Scenario 1. The explanatory variables were generated as Uniform (0,1). In addition, the sample sizes n were considered as equivalent to 20, 40, 60, 100, and 200 with different contamination levels $\alpha = 0.05, 0.10, 0.15$, and 0.20 of outliers with varying weights. The initial $100\alpha\%$ observations were further constructed as outliers in the data set. Thus, in order to induce the outlying values of the varying weights, the first outlier point was unchanged and remained at value 10, whereas those successive values increased as much as value two. The results of the 5,000 replications are summarized in Table 4.

Additionally, for Scenario 1, Table 3 clearly shows that the Pearson residuals method was excellent in detecting single outlier only when the size of the experimental run had been large, but its performance became extremely poor for multiple outliers detection. However, the proposed GSCPR method consistently displayed higher rate of detection of outliers with almost negligible masking rates and swamping rates regardless of the size of the experimental run and the existing number of outliers. The performance of the corrected Pearson residuals method was also encouraging, but it did not outperform the proposed method. Moreover, the rate of

correct detection for the corrected Pearson residuals method had been found to become lower as the amount of outliers increased.

Other than that, Table 4 portrays the eminence of the proposed method for Scenario 2. The Pearson residuals method had been discovered to be excellent only when the contamination level was low, but it exhibited an extremely poor performance when the contamination level was more than 5% in the cases. In general, it generated less percentage in correct detection, high masking, and low swamping effects. However, the GSCPR method that outperformed the other methods revealed higher rate of detection of outliers and almost negligible masking rates regardless of the sample size, n , and contamination levels, α . This method also had a tendency of swamping the inliers as outliers when the contamination level had been low, such as 5% and 10% cases. Moreover, the performance of the corrected Pearson residuals method was better than that of the Pearson residuals method since the rate of detection of outlier was higher when the contamination level was low. Although the swamping effects for the corrected Pearson residuals method were low, its masking effects increased as the contamination level increased. Hence, the outcomes of the study indicated that the GSCPR method had the best performance, followed by the corrected Pearson residuals method and the Pearson residuals method.

6. Conclusion

This paper proposed a diagnostic method for the identification of multiple outliers in GLM, where traditionally used outlier detection methods are effortless as they undergo masking or swamping dilemma. Thus, an investigation had been conducted to determine the capability of the proposed

TABLE 3: Percentages of correct detection of outlier, masking rates, and swamping rates based on 5,000 simulations for Scenario 1.

n'	Run	% Correct detection			% Masking			% Swamping		
		Pearson residuals	Corrected Pearson residuals	GSCPR	Pearson residuals	Corrected Pearson residuals	GSCPR	Pearson residuals	Corrected Pearson residuals	GSCPR
1	8	0.00	99.24	98.94	100.00	0.76	1.06	0.00	80.02	3.42
	16	0.18	99.98	99.84	99.82	0.02	0.16	0.00	1.54	3.22
	32	85.64	100.00	100.00	14.36	0.00	0.00	0.32	11.00	8.72
	64	100.00	100.00	100.00	0.00	0.00	0.00	25.52	26.78	5.92
2	8	0.00	97.86	91.28	100.00	2.14	8.72	0.00	12.22	4.52
	16	0.00	26.64	99.90	100.00	73.36	0.10	0.00	1.84	2.28
	32	0.10	18.74	100.00	99.90	81.26	0.00	0.00	8.94	6.24
	64	26.02	27.50	100.00	73.98	72.50	0.00	0.92	1.26	5.26
3	8	0.00	0.20	91.02	100.00	99.8	8.98	0.00	0.02	6.74
	16	0.00	0.36	97.88	100.00	99.64	2.12	0.00	0.00	4.96
	32	0.00	86.00	99.62	100.00	14.00	0.38	0.00	0.00	2.70
	64	0.00	5.70	100.00	100.00	94.28	0.00	0.00	0.42	5.02

TABLE 4: Percentages of correct detection of outlier, masking rates, and swamping rates based on 5,000 simulations for Scenario 2.

$\alpha\%$	n	n'	% Correct detection			% Masking			% Swamping		
			Pearson residuals	Corrected Pearson residuals	GSCPR	Pearson residuals	Corrected Pearson residuals	GSCPR	Pearson residuals	Corrected Pearson residuals	GSCPR
5	20	1	100.00	100.00	100.00	0.00	0.00	0.00	0.00	0.12	39.16
	40	2	88.56	100.00	100.00	11.44	0.00	0.00	0.00	0.08	40.66
	60	3	94.68	100.00	100.00	5.32	0.00	0.00	0.04	0.30	26.44
	100	5	100.00	100.00	100.00	0.00	0.00	0.00	0.16	0.46	32.69
	200	10	100.00	100.00	100.00	0.00	0.00	0.00	0.12	0.18	39.20
10	20	2	0.02	100.00	100.00	99.98	0.00	0.00	0.00	0.08	32.66
	40	4	0.00	99.80	100.00	100.00	0.20	0.00	0.00	0.02	15.69
	60	6	0.00	0.10	100.00	100.00	99.90	0.00	0.00	0.00	32.36
	100	10	0.00	95.98	100.00	100.00	4.02	0.00	0.00	0.00	38.74
	200	20	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	27.62
15	20	3	0.00	1.92	40.98	100.00	98.08	59.02	0.00	0.04	5.08
	40	6	0.00	20.98	86.62	100.00	79.02	13.38	0.00	0.02	24.04
	60	9	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	0.18
	100	15	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	4.28
	200	30	0.00	0.00	100.00	100.00	100.00	1.32	0.00	0.00	14.40
20	20	4	0.00	2.26	98.68	100.00	97.74	0.00	0.00	0.00	10.26
	40	8	0.00	4.66	100.00	100.00	95.34	0.00	0.00	0.00	4.74
	60	12	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	5.46
	100	20	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	1.60
	200	40	0.00	0.00	100.00	100.00	100.00	0.00	0.00	0.00	1.20

FIGURE 1: Index plot of (a) Pearson residuals for one outlier, (b) Pearson residuals for two outliers, (c) corrected Pearson residuals for one outlier, (d) corrected Pearson residuals for two outliers, (e) GSCPR for one outlier, and (f) GSCPR for two outliers.

GSCPR method. The findings obtained from the numerical examples indicated that the performance of the proposed method was satisfactory in identifying multiple outliers. Other than that, in the simulation study, two scenarios were considered to assess the validity of the GSCPR method. The results retrieved from the simulation study exhibited the superiority of the proposed method under wide assortment of conditions. The proposed method consistently displayed higher percentage of correct detection, as well as lower rates of swamping and masking, regardless of the sample size, n , and the contamination level, α .

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Research University Grant of the University Putra Malaysia, Malaysia.

References

- [1] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society, Series A: General*, vol. 135, no. 3, pp. 370–384, 1972.
- [2] J. K. Lindsey, *Applying Generalized Linear Models*, Springer Science & Business Media, 1997.
- [3] S. Kuhnt and J. Pawlitschko, *Outlier Identification Rules for Generalized Linear Models*, Springer, Berlin, Germany, 2005.
- [4] M. Habshah, M. R. Norazan, and A. H. Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression," *Journal of Applied Statistics*, vol. 36, no. 5-6, pp. 507–520, 2009.
- [5] G. M. Cordeiro and P. McCullagh, "Bias correction in generalized linear models," *Journal of the Royal Statistical Society—Series B: Methodological*, vol. 53, no. 3, pp. 629–643, 1991.
- [6] G. M. Cordeiro, "On Pearson's residuals in generalized linear models," *Statistics & Probability Letters*, vol. 66, no. 3, pp. 213–219, 2004.
- [7] G. M. Cordeiro and A. B. Simas, "The distribution of Pearson residuals in generalized linear models," *Computational Statistics & Data Analysis*, vol. 53, no. 9, pp. 3397–3411, 2009.
- [8] T. Anholetto, M. C. Sandoval, and D. A. Botter, "Adjusted Pearson residuals in beta regression models," *Journal of Statistical Computation and Simulation*, vol. 84, no. 5, pp. 999–1014, 2014.

- [9] A. H. M. R. Imon and A. S. Hadi, "Identification of multiple outliers in logistic regression," *Communications in Statistics—Theory and Methods*, vol. 37, no. 11-12, pp. 1697–1709, 2008.
- [10] A. H. Lee and W. K. Fung, "Confirmation of multiple outliers in generalized linear and nonlinear regressions," *Computational Statistics and Data Analysis*, vol. 25, no. 1, pp. 55–65, 1997.
- [11] J. M. Hilbe and A. P. Robinson, *Methods of Statistical Model Estimation*, CRC Press, 2013.
- [12] T. P. Ryan, *Modern Regression Methods*, vol. 655, John Wiley & Sons, 2008.
- [13] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, 1993.
- [14] S. L. Lewis, D. C. Montgomery, and R. H. Myers, "Examples of designed experiments with nonnormal responses," *Journal of Quality Technology*, vol. 33, no. 3, pp. 265–278, 2001.
- [15] S. L. Lewis, D. C. Montgomery, and R. H. Myers, "The analysis of designed experiments with non-normal responses," *Quality Engineering*, vol. 12, no. 2, pp. 225–243, 1999.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>