Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (6): 828-836, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.828.836

Modified Standardized Pearson Residual for the Identification of Outliers in Logistic Regression Model

Habshah Midi and Syaiba Balqish Ariffin
Department of Mathematics, Faculty of Science, Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia

Abstract: Detection of outlier based on standardized Pearson residuals has gained widespread use in logistic regression model in the presence of a single outlier. An innovation attempts in the same direction but dealing for a group of outliers have been made using generalized standardized Pearson residual which requires a graphical or a robust estimator to find suspected outliers to form a group deletion. In this study, an alternative measure namely modified standardized Pearson residual is derived from the robust logistic diagnostic. The weakness of standardized Pearson residuals and the usefulness of generalized standardized Pearson residual and modified standardized Pearson residual are examined through several real examples and Monte Carlo simulation study. The results of this study signify that the generalized standardized Pearson residual and the modified standardized Pearson residual perform equally good in identifying a group of outliers.

Key words: Logistic regression, outliers, masking, swamping, standardized Pearson residuals, group deletion

INTRODUCTION

The popularity of the logistic regression model stem from its wider acceptance in research studies that rely on the binary response. Traditionally, the Maximum Likelihood Estimator (MLE) is used to fit the logistic regression model and it has good optimally properties in ideals settings. Unfortunately, it is enormously sensitive to outliers and high leverage points (Pregibon, 1981; Hao, 1992; Victoria-Feser, 2002; Croux *et al*., 2002; Croux and Haesbroeck, 2003; Rousseeuw and Christmann, 2003; Imon, 2006; Cizek, 2007; Imon and Hadi, 2008; Habshah and Syaiba, 2012).

The term of outliers have very close ties with high leverage points. The outliers which are severely outlying in the response as well as in covariate space produce large residuals with respect to the fitted values can be classified as one type of high leverage points called bad leverages (Croux and Haesbroeck, 2003). In addition, Hao (1992) illuminated that outlier need not to be outlying (in space of Y or X) in the sense of having large influence to model fit by departing away from the fitted pattern set by rest of data. Therefore, the outliers have similar adverse effect with the high leverage points on parameter estimates and the error variances of the logistic regression line to some degrees, consequently their presence can mislead the model fit and interpretation.

Logistic regression model is a special case of generalized linear model, where usual approach to outlier detection is based on deviances, Pearson residuals and standardized Pearson residuals. There are a number of diagnostic tools based on these measures can be found in the literatures (Pregibon, 1981, 1982; Jennings, 1986; Pierce and Schafer, 1986; William and Aelst, 2005; McCullagh and Nelder, 1989; Bedrick and Hill, 1990; Ryan, 1997; Hosmer and Lemeshow, 2000; Sarkar *et al*., 2011). These methods, however, are successful only if the data contain a single outlier. Hampel *et al*. (1986) claimed that real data normally contains about 1-10% outliers and even the highest quality of real data is not free from outliers. Therefore, similar approaches for single case detection to be applied to a group case detection are faulty due to masking and swamping effects. The effect of masking can be seen when the diagnostic methods fail to detect all outliers correctly. In contrast to masking effect, swamping effect occurs when good points are wrongly detected as outliers (Hadi and Simonoff, 1993; Imon, 2006; Imon and Apu, 2007; Imon and Hadi, 2008; Nurunnabi *et al*., 2010; Habshah *et al*., 2009; Syaiba and Habshah, 2010).

Recently, the Generalized Standardized Pearson Residual (GSPR) appears to be the most efficient method in identifying the outliers in logistic regression model. The GSPR proposed by Imon and Hadi (2008) gives very promising results if an initial deletion set include all the

Corresponding Author: Syaiba Balqish Ariffin, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia Tel: +60139748959 Fax: +60389437958

suspected outliers. In order to perform the initial deletion set, one might suggest to apply robust estimator or to use graphical method. Croux and Haesbroeck (2003) noted that robust estimators for logistic regression are computationally expensive. Meanwhile, the outliers are more difficult to detect especially when the number of covariates is high in the model (Nurunnabi *et al*., 2010; Syaiba and Habshah, 2010).

In this study, the Robust Logistic Diagnostic (RLGD) method proposed by Syaiba and Habshah (2010) is employed to detect the high leverage points in logistic regression model. The RLGD is applied at the beginning of analysis to detect all the suspected outliers which hold same behaviour with the high leverage points and to assign new weights for the outliers in constructing alternative version of the GSPR method, namely the Modified Standardized Pearson Residual (MSPR).

MATERIALS AND METHODS

Logistic regression model: In this model, we would like to predict the probabilities that the response Y takes either value 0 or 1. Let π denote the probability of occurrence that Y = 1 when X = x. The probability is modelled as:

$$\begin{aligned}\pi &= \Pr(Y=1 \mid X_1=x_1, X_2=x_2,\ldots,X_p=x_p)\\ &= \frac{\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\right)}{1+\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\right)}\end{aligned} \quad (1)$$

The Eq. 1 is called the logistic regression function where the range of values of π is $0\leq\pi\leq1$. The function is nonlinear in the parameters β_0, β_1, β_2, …, β_p. However, it can be linearized by the logit transformation. Since:

$$\begin{aligned}1-\pi &= \Pr(Y=0 \mid X_1=x_1, X_2=x_2,\ldots,X_p=x_p)\\ &= \frac{1}{1+\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\right)}\end{aligned} \quad (2)$$

Then, the odds ratio or the probability of non-occurrence is:

$$\frac{\pi}{1-\pi} = \exp\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\right) \quad (3)$$

Taking the natural logarithm of both sides of Eq. 3, we obtain:

$$\begin{aligned}g(x_1, x_2,\ldots, x_p) &= \log\left(\frac{\pi}{1-\pi}\right)\\ &= \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\end{aligned} \quad (4)$$

The logarithm of the odds ratio is called the logit where the range of values of is logit (π) is $-\infty\leq$logit (π)$\leq\infty$. Suppose we have a sample of n observations with k = p+1 regression coefficients. Let the model in Eq. 1 is written in terms of Y as:

$$Y = \pi(x)+\varepsilon \quad (5)$$

Here, Y is an n×1 vector of responses, X is an n×k matrix of covariates, $\beta^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients and ε is an n×1 vector of unobserved random errors. The fitted values in logistic regression model are calculated for each covariate which depend on the estimated probability for that covariates, denoted as $\hat{y}_i = \hat{\pi}_i$. Thus, the ith residuals are defined as:

$$\hat{\varepsilon}_i = y_i-\hat{\pi}_i \quad (6)$$

In logistic regression, the fitting is carried out by working with the logit. The widely used method estimation is the maximum likelihood method. The maximum likelihood estimates are obtained numerically using iterative procedure. The iterative estimates of $\hat{\beta}$ are obtained as:

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + (X^TVX)^{-1}X^T\varepsilon \quad (7)$$

where, V is a diagonal matrix with elements of $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$.

Identification of single outlier: In this section, we will briefly review the identification methods for a single outlier. We begin with hat matrix for logistic regression derived by Pregibon (1981) from linear approximation to the fitted values which is:

$$H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2} \quad (8)$$

where, V is an n×n diagonal matrix with elements of $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$. Leverage values are diagonal elements of the hat matrix, H are defined as:

$$h_i = \hat{\pi}_i(1-\hat{\pi}_i)x^T_i(X^TVX)^{-1}x_i \quad (9)$$

where, $x^T_i = (1, x_{1i}, x_{2i}, \ldots, x_{pi})$ is the 1×k vector of observations corresponding to ith case.

Commonly, the residuals as defined in Eq. 6 are purposely used to measure the extent of ill-fitted of covariate patterns by examining any observation that possesses large residual to be suspected as outlier. However, the residuals are not appropriate to be used since they are unscaled. Since the response take only the values 0 and 1, the residual may assume one of the two

possible values, if $y_i = 1$ then $\hat{\varepsilon}_i = 1-\hat{\pi}_i$ with probability $\hat{\pi}_i$ and if $y_i = 0$ then $\hat{\varepsilon}_i = -\hat{\pi}_i$ with probability $1-\hat{\pi}_i$. Moreover, the distribution of residuals is not well determined. A scaled version of the residuals is Pearson residuals which are elements of the Pearson chi-square goodness-of-fit statistic. In logistic regression, there are binomial errors which mean, the error variance $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$ is a function of the conditional mean. The Pearson residuals for ith covariate pattern are given by:

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i}} \tag{10}$$

Any observation which correspond to Pearson residual exceeds $|r_i|>c$ is called an outlier. Perhaps c = 3 could be a reasonable choice that matches with 3σ distance rule used in the normal theory as mentioned by Ryan (1997). Imon and Hadi (2008) pointed out that the cut-off point c = 3 identify too many observations as outliers as they suggested to follow Chen and Liu (1993) by considering c as a constant from 3 to 5. The Pearson residuals do not have variance equal to 1. A better approach is to use a standardized version of the residuals, called Standardized Pearson Residuals (SPR) which are defined as:

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1-h_i)}} \tag{11}$$

with $|r_{si}|>3$ as a cut-off point to declare that the ith observation as an outlier.

Generalized standardized pearson residual: Imon and Hadi (2008) proved that the GSPR method is successful method in handling outliers compared to the current methods and this method is also free from the effect of masking and swamping. The GSPR is a group-deleted version of the residuals. Assume that d observations among a set of n observations are omitted before the fitting of the model. Denote a set of cases 'remaining' by R and a set of cases 'deleted' by D. Hence, R contains (n-d) cases with good observations after d cases with suspected outliers in D are deleted. Set D are allocated at the last of d rows of X, Y and V, yielding:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \text{ and } V = \begin{pmatrix} V_R & 0 \\ 0 & V_D \end{pmatrix}$$

We estimate the coefficients, $\hat{\beta}^{(-D)}$ after all the suspected outliers in set D is excluded. Thus, the fitted values can be written as:

$$\hat{\pi}_i^{(-D)} = \frac{\exp\left(x_i^T\hat{\beta}^{(-D)}\right)}{1+x_i^T\hat{\beta}^{(-D)}} \tag{12}$$

Then, the respective ith deletion of residuals $\hat{\varepsilon}_i^{(-D)} = y_i-\hat{\pi}_i^{(-D)}$, variances $v_i^{(-D)} = \hat{\pi}_i^{(-D)}(1-\hat{\pi}_i^{(-D)})$ and leverages $h_i^{(-D)} = \hat{\pi}_i^{(-D)}(1-\hat{\pi}_i^{(-D)})x^T_i (X^T_R V_R X_R)^{-1} x_i$ for the entire data are computed. The residuals that are used for fitting the model should be computed differently. This idea was pointed out for linear regression model by Hadi and Simonoff (1993). They recommended to use scaled residuals as follows:

$$\begin{aligned} t_i^* &= \frac{y_i - x_i^T\hat{\beta}^{(-D)}}{\hat{\sigma}_R\sqrt{1-w_{ii}^{(-D)}}}; \text{ for } i\in R \\ &= \frac{y_i - x_i^T\hat{\beta}^{(-D)}}{\hat{\sigma}_R\sqrt{1+w_{ii}^{(-D)}}}; \text{ for } i\in D \end{aligned} \tag{13}$$

This type of scaled residuals is extensively used in regression diagnostics (Atkinson, 1994; Munier, 1999; Imon, 2005). Using the above scaled residuals and also by using linear regression model approximation, Imon and Hadi (2008) defined the GSPR as:

$$\begin{aligned} r_{si}^{(-D)} &= \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1-h_i^{(-D)}\right)}}; \text{ for } i\in R \\ &= \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1+h_i^{(-D)}\right)}}; \text{ for } i\in D \end{aligned} \tag{14}$$

The declaration of an observation as outlier is when its corresponding GSPR value exceed the cut-off point of $|r_{si}^{(-D)}|>3$.

In the initial detection process to perform on set D, it is very important to ensure that none of suspected outliers remain in set R. The GSPR may not work properly if this condition fails. Many approaches have been considered for initial detection. Cook and Hawkins (1990) noted that robust methods prone to swamp inliers as outliers. Imon and Hadi (2008) suggested to apply robust methods e.g. Least Median Squares (LMS), Least Trimmed Squares (LTS), the Block Adaptive Computationally Efficient Outlier Nominator (BACON) or best omitted from the ordinary least squares techniques (BOFOL). Imon and Hadi (2008) declared the observations as outliers when $|r_{si}^{(-D_0)}|>3$.

Robust logistic diagnostic: As already mentioned, Imon and Hadi (2008) developed a generalized version of standard Pearson residuals based on group deletion for identifying of outliers. They have shown through some real examples that the GSPR values correctly identify all the outliers and not affected by the masking and swamping problems. They also highlighted an important

issue that need to be considered before performing the GSPR method. As an alternative procedure to robust method or graphical technique used in the GSPR, we propose to use the robust logistic diagnostic (RLGD) by Syaiba and Habshah (2010) as initial detection process to form deleted set D.

Let:

$$\tilde{X} = V^{\frac{1}{2}}X$$

Thus, group deleted distance from mean based on group deleted cases D is:

$$b_i^{(-D)} = x_i^T \left(\tilde{X}_R^T \tilde{X}_R\right)^{-1} x_i \quad (15)$$

The value of Eq. 15 is computed using $\hat{\beta}^{(-D)}$ after completing the first stage of the RLGD method and $x_i^T = [1, x_{1i}, x_{2i}, \ldots, x_{pi}]$ is the 1×k vector of observations corresponding to the ith case. The relationship between potential values proposed by Hadi (1992) and Eq. 1 gives:

$$b_i^{(-D+i)} = x_i^T \left(\tilde{X}_R^T \tilde{X}_R + x_i x_i^T\right)^{-1} x_i = \frac{b_i^{(-D)}}{1+b_i^{(-D)}} \quad (16)$$

Then, the group deleted potential is defined by:

$$p_{ii}^{*(-D)} = \begin{cases} \dfrac{b_i^{(-D)}}{1+b_i^{(-D)}}; & i \in R \\ b_i^{(-D)}; & i \in D \end{cases} \quad (17)$$

Since the distribution of $p_{ii}^{*(-D)}$ is not determined, we apply cut-off point based on median and Median Absolute Deviation (MAD) for $p_{ii}^{*(-D)}$ as suggested by Hadi (1992). The MAD for $p_{ii}^{*(-D)}$ is computed by:

$$MAD(p_{ii}^{*(-D)}) = \frac{Median\{|p_{ii}^{*(-D)} - Median(p_{ii}^{*(-D)})|\}}{0.6745} \quad (18)$$

and the cut-off point with the constant c = 3 is given as:

$$p_{ii}^{*(-D)} > Median\left(p_{ii}^{*(-D)}\right) + cMAD\left(p_{ii}^{*(-D)}\right) \quad (19)$$

Hence, any observation corresponding to excessively large potential values with above cut-off point, shall be declared as high leverage points. The step of RLGD method is summarized as follows:

Step 1: For each ith point, compute Robust Mahalanobis Distance (RMD) either using Minimum Covariance Determinant (MCD) or Minimum Volume Ellipsoid (MVE) estimators

Step 2: An ith point with $RMD_i > Median(RMD_i) + cMAD(RMD_i)$, are suspected as high leverage points and will be included in the deleted set D

Step 3: Based on the remaining points in set R, compute the $p_{ii}^{*(-D)}$

Step 4: Any deleted points with $p_{ii}^{*(-D)} > Median(p_{ii}^{*(-D)}) + cMAD(p_{ii}^{*(-D)})$ are finalized and declared as high leverage points

Modified standardized pearson residual: Our proposed method is called the Modified Standardized Pearson Residual (MSPR). The MSPR serves as an alternative method to the GSPR. Firstly, the RLGD is utilized to identify the suspected high leverage points. Then, we form a deletion set and estimate $\hat{\beta}^{(-D)}$ of the MLE. Thus, the MSPR for group deleted residual is computed as follows:

$$r_{si}^{MLE(-D)} = \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1-h_i^{(-D)}\right)}} \quad (20)$$

We shall call the above new measure as MSPR1. The RLGD also used to determine weights for the Weighted Maximum Likelihood Estimator (WMLE). We define the weight as:

$$w_i^{RLGD} = \min\left\{1, \frac{p}{p_{ii}^{*(-D)}}\right\} \quad (21)$$

where, p is the number of parameter excluding the intercept terms. The Eq. 21 was first introduced by Hubert and Rousseeuw (1997). They computed positive weights $w_i = \min\{1, p/RMD(x_i)^2\}$ based on the RMD. Thus, iterative estimates of $\hat{\beta}^{WMLE}$ are then obtained as follow:

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + \left(X^T X^{1/2} W^{RLGD} W^{1/2} X\right)^{-1} X^T W^{RLGD} e \quad (22)$$

Now, the MSPR for WMLE is computed as follows:

$$r_{si}^{WMLE} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1-h_i)}} \quad (23)$$

We shall call the above new measure as MSPR2. Therefore, for the MSPR1 value and MSPR2 value, the cut-off point to declare the ith observation as outliers are $|r_{si}^{MLE(-D)}| > 3$ and $|r_{si}^{WMLE}| > 3$.

Detail steps of the MSPR1 method are described as follows:

Step 1: For each ith point, compute $p_{ii}^{*(-D)}$

Step 2: Suspected high leverage points are included in the set D. The remaining points are put into the set R

Step 3: Based on set R, estimate $\hat{\beta}^{(-D)}$. Then, compute $r_{si}^{MLE(-D)}$

Step 4: Any points with $r_{si}^{MLE(-D)}$ are finalized and declared as outliers

Meanwhile for the MSPR2 method:

Step 1: For each ith point, compute $p_{ii}^{*(-D)}$

Step 2: Based on $p_{ii}^{*(-D)}$, compute weight w_i^{RLGD}

Step 3: Estimate $\hat{\beta}^{WMLE}$. Then compute r_{si}^{WMLE}

Step 4: Any points with $|r_{si}^{WMLE}|>3$ are finalized and declared as outliers

RESULTS AND DISCUSSION

The usefulness of the proposed MSPR1 and MSPR2 methods is investigated on two well-known data sets and their performance is compared to the SPR and the GSPR methods. The evaluation is based on number of outliers that can be correctly detected by these identification methods.

The modified prostate cancer data: We first consider the modified prostate cancer data with a single regressor given by Imon and Hadi (2008). Here, we are interested to determine whether an elevated level of Acid Phosphate (AP) in the blood serum would contribute for predicting whether prostate cancer patients had Lymph Nodes Involvement (LNI). Ryan (1997) pointed out that the original data set on 53 cancer patients contain a single outlier at observation 24, $Z_{24(y,x)} = (0,186)$. Imon and Hadi (2008) modified this data by putting two more outliers in variables AP for observation 54, $Z_{54(y,x)} = (0,200)$ and observation 55, $Z_{55(y,x)} = (0,220)$.

Many authors claimed that modified prostate cancer data contains three outliers for cases (24, 54 and 55). The index plot of AP for modified prostate cancer data (Fig. 1) clearly reveals those observations 24, 54 and 55 are outlying in both Y and X spaces and may severely break the covariate pattern. Hence, we consider these three points as outliers. From the RLGD(MCD) method, we manage to detect observations 24, 25, 53, 54 and 55 as high leverage points. Our initial detection result includes observations 24, 25, 53, 54 and 55 as illustrated by Fig. 1. Then, we apply the MSPR1 and the MSPR2 to identify the outliers. We also compared the results with the GSPR and the SPR methods.

Table 1 present the outliers diagnostic for the modified prostate cancer data. Setting the cut-off point as 3 in absolute terms for detection methods, we observe that from this table, the single detection of classical SPR failed to identify even a single outlier. As to be expected,

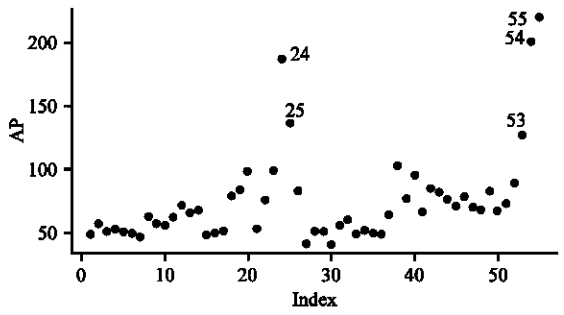

Fig. 1: Index plot of AP for modified prostate cancer data

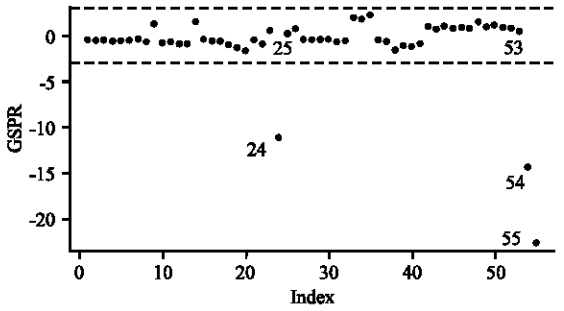

Fig. 2: Index plot of GSPR for modified prostate cancer data

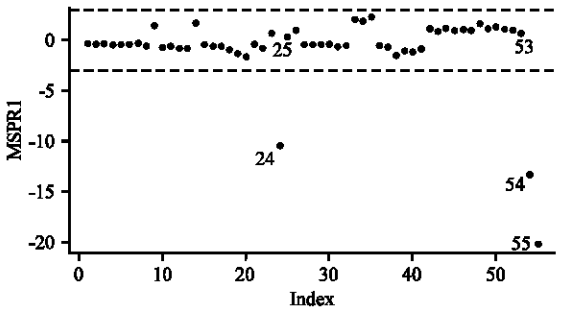

Fig. 3: Index plot of MSPR1 for modified prostate cancer data

our alternative method of the MSPR1 and the MSPR2 and the GSPR method give similar results and successfully detect three outliers (cases 24, 54 and 55).

Similar conclusions may be drawn from the index plot of the GSPR (Fig. 2) and the MSPR1 (Fig. 3). All three suspected cases are clearly separated from the rest of the data and correctly identified as outliers. From the initial detection by RLGD(MCD), we also identify case 25 and case 53 in order to perform initial deletion set, but these cases do not appear as outliers after applying the MSPR1, the MSPR2 and the GSPR methods. We suspect that cases 25 and 53 as high leverage points but having small influence to model fit.

Table 1: Outliers diagnostics for modified prostate cancer data

	Cut-off points			
Index	3 SPR	3 GSPR	3 MSPR1	3 MSPR2
1	-0.7323	-0.4872	-0.5205	-0.4763
2	-0.7405	-0.5904	-0.6152	-0.5056
3	-0.7343	-0.5112	-0.5426	-0.6670
4	-0.7363	-0.5363	-0.5657	-0.4419
5	-0.7343	-0.5112	-0.5426	-0.6670
6	-0.7333	-0.4990	-0.5314	-0.4926
7	-0.7303	-0.4643	-0.4992	-0.2534
8	-0.7471	-0.6812	-0.6985	-0.6191
9	1.3818	1.6484	1.6722	1.7748
10	-0.7394	-0.5764	-0.6024	-0.4888
11	-0.7471	-0.6812	-0.6985	-0.6191
12	-0.7578	-0.8415	-0.8491	-0.5917
13	-0.7506	-0.7313	-0.7449	-0.4818
14	1.3539	1.2755	1.3153	0.5459
15	-0.7313	-0.4756	-0.5097	-0.2622
16	-0.7333	-0.4990	-0.5314	-0.4926
17	-0.7343	-0.5112	-0.5426	-0.6670
18	-0.7668	-0.9880	-0.9934	-1.0853
19	-0.7736	-1.1060	-1.1142	-0.9021
20	-0.7959	-1.5464	-1.5844	-1.5459
21	-0.7363	-0.5363	-0.5657	-0.4419
22	-0.7629	-0.9227	-0.9283	-0.6798
23	1.2873	0.5837	0.6894	0.6986
24	-1.0151	-12.8727	-10.5325	-29.8940
25	1.2369	0.2432	0.3172	0.1953
26	1.3200	0.8883	0.9637	0.2388
27	-0.7247	-0.4021	-0.4405	-0.2946
28	-0.7343	-0.5112	-0.5426	-0.6670
29	-0.7343	-0.5112	-0.5426	-0.6670
30	-0.7247	-0.4021	-0.4405	-0.2946
31	-0.7394	-0.5764	-0.6024	-0.4888
32	-0.7438	-0.6343	-0.6553	-0.3932
33	1.4038	1.9821	1.9982	1.8717
34	1.3954	1.8498	1.8688	1.3770
35	1.4010	1.9370	1.9541	1.7875
36	-0.7323	-0.4872	-0.5205	-0.4763
37	-0.7483	-0.6976	-0.7136	-0.4502
38	-0.8023	-1.6921	-1.7404	-1.7796
39	-0.7642	-0.9440	-0.9494	-1.0089
40	-0.7912	-1.4459	-1.4763	-1.3890
41	-0.7518	-0.7488	-0.7612	-0.4985
42	1.3159	0.8457	0.9259	1.1261
43	1.3221	0.9104	0.9832	1.2396
44	1.3330	1.0284	1.0887	0.5400
45	1.3467	1.1879	1.2339	0.8591
46	1.3286	0.9797	1.0449	0.4386
47	1.3467	1.1879	1.2339	0.8591
48	1.3539	1.2755	1.3153	0.5459
49	1.3200	0.8883	0.9637	0.2388
50	1.3539	1.2755	1.3153	0.5459
51	1.3421	1.1325	1.1830	1.6628
52	1.3058	0.7474	0.8386	0.9609
53	1.2477	0.3063	0.3954	0.2778
54	-1.0664	-17.5581	-13.8270	-46.3479
55	-1.1618	-28.2343	-21.1062	-91.1242

SPR: Standardized Pearson residuals, GSPR: Generalized standardized Pearson residual, MSPR: Modified standardized Pearson residual

The modified vaso-constriction skin digits data: We now consider the modified vaso-constriction skin digits data given by Imon and Hadi (2008). The original data set were obtained to study the effect of the rate (RATE) and volume of air inspired (VOL) on the occurrence or non occurrence of vaso-constriction skin digits after air inspiration. This data is difficult to handle due to the presence of outliers which are not outlying in the Y space. At the same time, existence of estimation is highly dependent on these points. Pregibon (1981) pointed out that the original data might contain two outliers, namely observation 4, $z_{4(y,x_1,x_2)} = (1, 0.750, 1.500)$ and observation 18, $z_{18(y,x_1,x_2)} = (1, 0.850, 1.415)$. Croux and Haesbroeck (2003) detected the same observations as outliers in their study. They noted that observation 4 and 18 can be classified as influential observations for the MLE, therefore, deleting these points may leads to a data set whose overlap relies only on these observations. Imon and Hadi (2008) modified this data by putting two more outliers for observation 10, $z_{10(y,x_1,x_2)} = (1, 0.900, 0.450)$ and observation 11, $z_{11(y,x_1,x_2)} = (1, 0.800, 0.570)$, where the non-occurrences are replaced by occurrences.

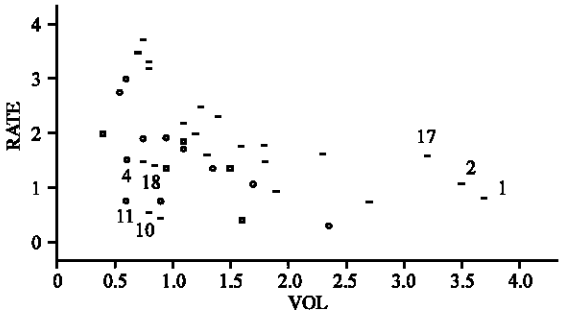

Fig. 4: Scatter plot of RATE vs. VOL for modified vaso-constriction skin digits data

Figure 4 presents the character plot of the modified vaso-constriction skin digits data where RATE is plotted against VOL and the character corresponding to occurrence Y = 1 and non-occurrences Y = 0 are denoted by symbols of triangle and circle, respectively. Looking at the pattern of occurrence and non-occurrence in relation to RATE and VOL, we observe that observations 4, 10, 11 and 18 are suspected as outliers and observations 1, 2 and 17 are suspected high leverage points by the RLGD(MCD) at the initial detection stage. Then we perform the deletion set, that consist of cases 1, 2, 4, 10, 11, 17 and 18.

Table 2 presents the outliers diagnostic for the modified vaso-constriction skin digits data. We observe from this table, that the classical SPR failed to identify even a single outlier. As to be expected, our alternative method of the MSPR1, the MSPR2 and the GSPR proposed by Imon and Hadi (2008) perform similar results and successfully detect four outliers (cases 4, 10, 11 and 18). We observed that the MSPR1 performs better compared to MSPR2 for the modified vaso-constriction skin digits data by looking at the MSPR2 values which are

Table 2: Outliers diagnostics for modified vaso-constriction skin digits data

	Cut-off points			
Index	3 SPR	3 GSPR	3 MSPR1	3 MSPR2
1	0.1604	0.0000	0.0000	0.1193
2	0.1655	0.0000	0.0000	0.1270
3	0.5908	0.0918	0.0966	0.6372
4	1.6682	97.1021	97.2680	3.9529
5	0.6179	0.0864	0.0912	0.7140
6	0.5796	0.0489	0.0502	0.6839
7	-0.3684	-0.0002	-0.0002	-0.3067
8	-0.9795	-0.1780	-0.1922	-0.8881
9	-0.4864	-0.0013	-0.0013	-0.4252
10	2.6864	2476.8020	2476.8210	3.0982
11	2.7459	2812.0060	2812.0230	3.2111
12	-1.1082	-0.3832	-0.4913	-0.9372
13	-1.3480	-1.2977	-2.0230	-1.1548
14	0.5655	0.0735	0.0762	0.5981
15	0.4806	0.0137	0.0137	0.5643
16	0.3699	0.0051	0.0051	0.3472
17	0.1624	0.0000	0.0000	0.1302
18	1.5926	74.4758	74.6675	3.8363
19	-1.2046	-0.5031	-0.5919	-1.2116
20	0.5304	0.0542	0.0555	0.5349
21	-0.6254	-0.0090	-0.0091	-0.5128
22	-0.7097	-0.0196	-0.0197	-0.6282
23	-1.0144	-0.1999	-0.2178	-0.9568
24	-1.1738	-0.5020	-0.5700	-1.1380
25	0.6413	0.1872	0.2081	0.6630
26	-0.5619	-0.0042	-0.0042	-0.4724
27	0.6296	0.1706	0.1883	0.6372
28	-0.9579	-0.1596	-0.1715	-1.2290
29	0.7970	0.7707	0.9818	0.8038
30	-0.7742	-0.0220	-0.0222	-0.7618
31	0.4388	0.0155	0.0156	0.3952
32	-1.5263	-1.0957	-1.8793	-1.7623
33	-1.0542	-0.2927	-0.3267	-0.9567
34	0.7992	0.6930	0.8366	0.8789
35	0.8142	0.8507	0.9867	0.8838
36	0.5761	0.0524	0.0539	0.6669
37	-0.9579	-0.1596	-0.1715	-1.2290
38	-0.8023	-0.0493	-0.0502	-0.6916
39	0.9188	2.0751	2.3320	0.9863

SPR: Standardized Pearson residuals, GSPR: Generalized standardized Pearson residual, MSPR: Modified standardized Pearson residual

closer to the cut-off point 3. It is evident that down weighting scheme for the high leverage points are less efficient for influential observations (Nurunnabi *et al*., 2010). These four observations (cases 4, 10, 11 and 18) are not outlying in covariate space but are considered as outliers as mentioned by Croux and Haesbroeck (2003).

Similar conclusion may be drawn from the index plot of the GSPR (Fig. 5) and the MSPR1 (Fig. 6). All four suspected cases are clearly separated from the rest of the data and correctly identified as outliers. At the initial detection by RLGD(MCD), we omit cases (1, 2, 4, 10, 11, 17 and 18) in order to perform initial deletion set. Then, after applying the GSPR and the MSPR1 methods, we observe that case 10 and case 11 are highly influential to fitted values by looking at the GSPR and the MSPR1 values which are far from the cut-off points 3 while case 4 and

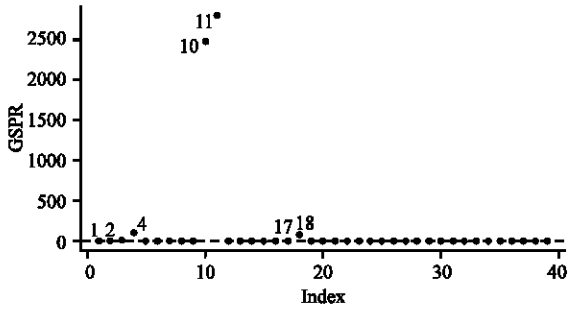

Fig. 5: Index plot of GSPR for modified vaso-constriction skin digits data

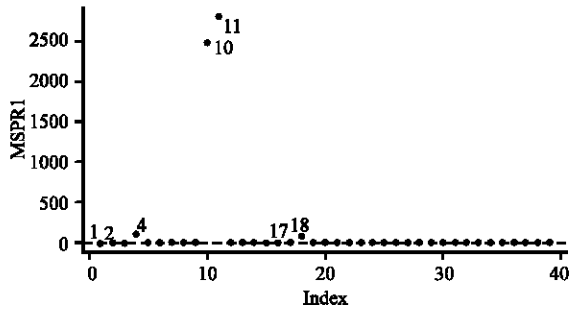

Fig. 6: Index plot of MSPR1 for modified vaso-constriction skin digits data

case 18 are less influential to fitted values. It seems that cases 1, 2 and 17 are the suspected high leverage points but may not be the outliers.

Monte Carlo simulation study: A simulation study is carried out in order to further assess the performance of GSPR, MSPR1 and MSPR2 with the cut-off points 3 in absolute terms. The performance of these methods is evaluated based on probability of the Detection Capability (DC) and probability of the False Alarm Rate (FAR) (Kudus *et al*., 2008). We conducted a similar simulation study to the RLGD method as discussed in Syaiba and Habshah (2010). The RLGD method is applied with the 50% breakdown point of MCD estimator and median and MAD cut-off point, setting c as 3. The choice for sample size starting with n = 100 is to ensure the existence and stability of the MLE as recommended by Victoria-Feser (2002). We considered different percentage of contaminations denoted as s, such that s = (50, 10, 15, 20%). In Type 1 data (uncontaminated), x are generated according to a standard normal distribution, $x_1 \sim N(0, 1)$ and $x_2 \sim N(0, 1)$ with the error terms is generated according to logistic distribution, $\varepsilon_i \sim \Lambda(0, 1)$. The response y is generated in the following manner:

Table 3: The measures of performance on the diagnostic methods on moderate contamination

Detection methods	Measures of performance	Intermediate contamination (%) 0	5	10	15	20
GSPR	FAR	0	0.0088	0.0087	0.0086	0.0085
	DC	-	0.9974	0.9972	0.9974	0.9967
MSPR1	FAR	0	0.0101	0.0092	0.0079	0
	DC	-	0.9948	0.9948	0.9895	1
MSPR2	FAR	0	0.0035	0.0049	0.0015	0
	DC	-	0.9876	0.9998	0.9968	1

GSPR: Generalized standardized Pearson residual, MSPR: Modified standardized Pearson residual, FAR: Probability of false alarm rate, DC: Probability of detection capability

$$y_i = \begin{cases} 0 \text{ if } \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i < 0 \\ 1 \text{ if } \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \geq 0 \end{cases} \quad (24)$$

Meanwhile, for Type 2 (5% intermediate contamination), z are generated according to a same standard normal distribution, $z_1 \sim N(0, 1)$ and $z_2 \sim N(0, 1)$ with magnitude of outlying shift distance in X-space is taken as $\delta = 5$, respectively. Now, new contaminated data, x^* are written as $x^*_1 = z_1 + \delta$ and $x^*_2 = z_2 - \delta$ with the error $\varepsilon_i \sim \Lambda(0, 4)$. The response is defined as the following model equations:

$$y^*_i = \begin{cases} 0 \text{ if } \beta_0 + \beta_1 x^*_1 + \beta_2 x^*_2 + \varepsilon_i \geq 0 \\ 1 \text{ if } \beta_0 + \beta_1 x^*_1 + \beta_2 x^*_2 + \varepsilon_i \leq 0 \end{cases} \quad (25)$$

The last Type 3 (5% extreme contamination) followed Type II procedure with magnitude of outlying shift distance in X-space is given as $\delta = 10$. We set the true parameters as $\beta = (\beta_0, \beta_1, \beta_2)^T = (0.5, 1, -1)^T$. The simulation for each detection method is computed over M = 1000 replications. The FAR gives probabilities of swamping occur, meanwhile the DC gives probabilities of masking occur for the GSPR, MSPR and MSPR2 methods. For every repetition, these detection methods assigned the outliers with weights $w_i = 1$ and $w_i = 0$ for good observations. Then, $\bar{w}_1$ is denoted as average for weights of the uncontaminated $\bar{w}_1 = \sum_{i=1}^{n} w_i / n$ and $\bar{w}_2$ is average for weights of the contaminated, $\bar{w}_2 = \sum_{j=1}^{s} w_j / s$. Hence, the probabilities of FAR and DC can be computed as $FAR = \sum_{i=1}^{m} \bar{w}_1 / M$ and $DC = \sum_{i=1}^{M} \bar{w}_2 / M$. A criterion for 'good' diagnostic method is based on which detection method has probability of the FAR closest to 0 and the DC closest to 1. Result on the simulation study in the identification of the outliers for the GSPR, MSPR1 and MSPR2 are shown as follows:

A 'good' method of identifying the outliers is the method which performs the probability of DC closer or exactly 1 and the probability of FAR closer or exactly 0. In this simulation study, we consider two covariates in the model. Refer to Table 3 and Table 4, all detection methods

Table 4: The measures of performance on the diagnostic methods on extreme contamination

Detection methods	Measures of performance	Extreme contamination (%) 0	5	10	15	20
GSPR	FAR	0	0.0088	0.0088	0.0088	0.0088
	DC	-	1	1	1	1
MSPR1	FAR	0	0.0102	0.0104	0.0103	0.0103
	DC	-	1	1	1	1
MSPR2	FAR	0	0.007	0.0041	0.0017	0.0006
	DC	-	1	1	1	1

GSPR: Generalized standardized Pearson residual, MSPR: Modified standardized Pearson residual, FAR: Probability of false alarm rate, DC: Probability of detection capability

give good results. In the 20% of intermediate contamination, the MSPR1 and the MSPR2 performs slightly better compared to the GSPR with DC value exactly 1 and FAR value exactly 0. In the extreme outliers, the MSPR2 is the best detection method followed by the GSPR and the MSPR1. It is important to note here, our simulation results are different compared to real data results regarding on the MSPR1 and MSPR2. From real examples, the MSPR1 shows better performance compared to MSPR2. Meanwhile, in simulation examples, we intended to investigate the usefulness of down weighting scheme in the MSPR2 by generating the outliers which are outlying in X space and highly influential to fitted values. Therefore, in simulation examples, down weighting scheme in the MSPR2 method works better compared to the MSPR1 method.

CONCLUSION

The main focus of this study was to develop an alternative method for identification of outliers in logistic regression model. The commonly used SPR only good for identification of single outlier but failed to detect multiple outliers. The GSPR successfully identify multiple outliers in a data, but the accomplishment of the GSPR depends on the initial deleted set. In this study, two alternative methods were proposed namely the MSPR1 and MSPR2 based on the RLGD. From the numerical examples on modified prostate cancer data and modified vaso-constriction skin digits data, the GSPR, MSPR1 and MSPR2 identify all the outliers correctly. The simulation results indicate that the MSPR2 have slightly better detection probability and have false alarm rate up to 20% of contamination data set for two covariates in the model. Generally, the GSPR, MSPR1 and MSPR2 perform equally good in identifying the outliers in logistic regression model.

REFERENCES

Atkinson, A., 1994. Fast very robust methods for the detection of multiple outliers. J. Am. Stat. Assoc., 89: 1329-1339.

Bedrick, E.J. and J.R. Hill, 1990. Outlier tests for logistic regression: A conditional approach. Biometrika, 77: 815-827.

Chen, C. and L.M. Liu, 1993. Joint estimation of model parameter and outliers effect in time series. J. Am. Stat. Assoc., 88: 284-297.

Cizek, P., 2007. Robust and efficient adaptive estimation of binary-choice regression models. CentER Discussion Paper, Tilburg University.

Cook, R.D. and D.M. Hawkins, 1990. Unmasking multivariate outliers and leverage points: Comment. J. Am. Stat. Assoc., 85: 640-644.

Croux, C. and G. Haesbroeck, 2003. Implementing the bianco and yohai estimator for logistic regression. Comput. Statist. Data Anal., 44: 273-295.

Croux, C., C. Flandre and G. Haesbroeck, 2002. The breakdown behaviour of the maximum likelihood estimator in the logistic regression model. Stat. Probab. Lett., 60: 377-386.

Habshah, M. and B.A. Syaiba, 2012. The performance of classical and robust logistic regression estimators in the presence of outliers. Pertanika J. Sci. Technol., 20: 313-325.

Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Statist., 36: 507-520.

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Comput. Statist. Data Anal., 14: 1-27.

Hadi, A.S. and J.S. Simonoff, 1993. Procedure for the identification of multiple outliers in linear models. J. Am. Statistical Assoc., 88: 1264-1272.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Sathel, 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Hao, Y., 1992. Maximum median likelihood and maximum trimmed likelihood estimations. Doctoral Thesis, University of Toronto, Canada.

Hosmer, D.W. and S. Lemeshow, 2000. Applied logistic Regression. 2nd Edn., Wiley, New York, pp: 143-199.

Hubert, M. and P.J. Rousseeuw, 1997. Robust regression with both continuous and binary regressors. J. statist. Plann. inference, 57: 153-163.

Imon, A.H.M.R., 2005. Identifying multiple influential observations in linear regression. J. Applied Statist., 32: 929-946.

Imon, A.H.M.R., 2006. Identification of high leverage points in logistic regression. Pak. J. Statist., 22: 147-156.

Imon, A.H.M.R. and M.R. Apu, 2007. Identification of multiple high leverage points using robust mahalanobis distance. J. Statist. Stud., 32: 929-946.

Imon, A.H.M.R. and A.S. Hadi, 2008. Identification of multiple outliers in logistic regression. Commun. Statist. Theory Meth., 37: 1697-1709.

Jennings, D.E., 1986. Outliers and residual distributions in logistic regression. J. Am. Statist. Assoc., 81: 987-990.

Kudus, A., I.N. Akma and D. Isa, 2008. Simulation on group deleted generalized potentials for diagnostics of censored survival regression. Proceedings of the 16th National Mathematical Science Symposium, June 2-5, 2008, Kota Bharu, Kelantan, Malaysia, pp: 280-287.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapman and Hall, New York. pp: 98-148.

Munier, S., 1999. Multiple outlier detection in logistic regression. J. Statist. Student, 3: 117-126.

Nurunnabi, A.A.M., A.H.M.R. Imon and M. Nasser, 2010. Identification of multiple influential observations in logistic regression. J. Applied Stat., 37: 1605-1624.

Pierce, D.A. and D.W. Schafer, 1986. Residuals in generalized linear models. J. Am. Stat. Assoc., 81: 977-986.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist., 9: 705-724.

Pregibon, D., 1982. Resistants fits for some commonly used logistic regression models with medical applications. Biometrics, 38: 485-498.

Rousseeuw, P.J. and A. Christmann, 2003. Robustness against separation and outliers in logistic regression. J. Comput. Stat. Data Anal., 43: 315-332.

Ryan, T.P., 1997. Modern Regression Methods. John Wiley and Sons, Inc., New York pp: 255-313.

Sarkar, S.K., H. Midi and S. Rana, 2011. Detection of outliers and influential observations in binary logistic regression: An empirical study. J. Applied Sci., 11: 26-35.

Syaiba, B.A. and M. Habshah, 2010. Robust logistic diagnostic for the identification of high leverage points in logistic regression model. J. Applied Sci., 10: 3042-3050.

Victoria-Feser, M.P., 2002. Robust inference with binary data. Psychometrika, 67: 21-32.

William, G. and S.V. Aelst, 2005. Fast and robust boostrap for LTS. Comput. Stat. Data Anal., 48: 703-715.