Global and Stochastic Analysis
Vol. 4 No. 1, January (2017), 71-76

MUK PUBLICATIONS
Open Access Publisher

IDENTIFICATION OF OUTLIERS IN LOGISTIC REGRESSION BASED ON ROBUST DEVIANCE COMPONENT

K. SENTHAMARAI KANNAN AND G. SAMPATH*

ABSTRACT. Logistic regression (LR) is widely used in the medical field and behavioral sciences to develop the model for the dependence of the binomial response variable on a set of predictor variables. Diagnostic is one of most indispensable thing that needs to be taken into account for analyzing data. Leverage outlier can easily biased the parameter estimates and imprecise the other observation in LR model. This study examined the daily variation in hospital for cardiovascular disease. Robust Deviance Component (robDEVC) method is based on deviance component. The aim of this paper is to analyze, whether the high leverage points can either be good or bad and the outliers are identified with numerical illustrations.

1. Introduction

Multivariate methods are most common statistical analyses appearing in the medical literature. It is relationship between two or more predictor (independent, exposure) variables and one outcome (dependent, response) variable. Formally, the model states for the predicted value of the outcome variable as a sum of products, each product formed by multiplying the value of the variable and its coefficient. This coefficients are calculated from the data taken. A regression model assists two purposes: First one is to predict the outcome variable for new values of the predictor variables and second one is to help for answer the queries about the area in study. The coefficient of each predictor variable clearly describes the relative role of that variable to the outcome variable, it automatically controls the influences of the other predictor variables.

LR is a multivariate method that was developed for dichotomous outcomes (Vollmer RT, 1996; Hosmer DW *et. al.,* 1989). The particularly appropriate for the models which are involving syndrome state (diseased/healthy) and decision making (yes/no) and therefore it is broadly used in the health science studies. LR technique is used to find the logarithm of the odds of a positive outcome (here "positive" means the encoding of the outcome variable, (i.e., Y = 1)); a straightforward algebraic manipulation transforms this into the outcome's probability.

In multivariate methods, the calculation of the coefficients from the original data is more complex than can be conveniently performed by hand. The coefficients calculation for logistic models involves equations that are not able to solve

2000 *Mathematics Subject Classification.* Primary 62G35; Secondary 62G08.
Key words and phrases. logistic regression, leverage, Robust Deviance Component, outlier, Cardiovascular disease .

explicitly but can solved by an iterative procedure, easily expressed in computational form. The quality of the regression analysis depends heavily on researchers understanding. The following principles are developed to ensure their sound application (Feinstein AR, 1996; Concato J, 1994). Generally, Diagnostics methods are used in all branches of regression analysis. In modern years diagnostic has become an obligatory part of LR (Hosmer and Lemeshow, 2000). A diagnostic method on the identification of residual outliers in LR based on deviance components (Syaiba B.A, 2010; Sanizah A, 2011; Norazan M.R, 2012).

2. Materials and Methods

In this study, the secondary data was used. The data was taken from the Worcester Heart Attack Study data from Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School.

2.1. Logistic Regression. Logistic model describe the expected value of y that is E(y) in terms of the following "logistic" formula:

$$E(y) = \frac{1}{1 + exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_j\right)\right]}$$

For (0,1) random variable say 'y' it follows from basic statistical principles about expected values. So that E(y) is equal to the probability pr(y=1); so, the formula for the logistic model can be written in the form which describes the probability of occurrence of one of the two possible outcomes of y, as follows:

$$pr\,(y = 1) = \frac{1}{1 + exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_j\right)\right]}$$

The logistic model is useful in many important practical situation where the response variable can yield one of two possible values. For example, a study of the development of a specific disease in some human population could employ a logistic model to describe in probabilistic terms, whether a specified individual in the study group will (y = 1) or will not (y =0) develop the disease in question during a follow up period of concern.

The first step in LR analysis is to propose (based on knowledge about and experience, with the process under study) a mathematical model describing the average of y as a function of the x_j and β_j values. Using maximum likelihood, the model is then fitted to the data, are eventually appropriate statistical inferences are made (after the model's adequacy of fit is verified, including consideration of relevant regression diagnostic indices).

2.2. Residuals. In LR, we can get a feel for exactly in what way the model agrees with the data by comparing the observed and predicted logits or probabilities for all possible covariate patterns. In multiple LR, the residuals provided useful information about possible problem with the model. We can also use the residuals in LR to examine the fit of the logistic model. Two common forms of residuals used in LR are Standardized Pearson Residuals (SPR), Deviance Residual and

Pearson Residuals (PR). These residuals are useful for identifying outlying and influential points (Pregibon 1981). The PR is defined as

$$r_i = \frac{y_i - n_i\widehat{\pi}_i}{\sqrt{n_i\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)}}$$

where n_i is the number of observations with the i^{th} covariate pattern, y_i is the number of observations with the outcome of interest among n_i observations, $\widehat{\pi}_i$ is the predicted probability of the outcome of interest for the i^{th} covariate pattern. The form of the PR is familiar – dividing the difference in the observed count. Note that we can also express the numerator of r_i as $y_i - \hat{y}_i$, where $\hat{y}_i$ is equal to $n_i\widehat{\pi}_i$.

Some recommend, a slightly different form of the PR. For example, according to collet (2003), a better procedure is to divide the raw residual, $y_i - \hat{y}_i$, by its standard error (se), $se(y_i - \hat{y}_i)$. This standard error is complicated to derive, but it is used in many of the LR programs. Residuals based on the $se(y_i - \hat{y}_i)$ are known as the (SPD). The DR is defined as

$$d_i = sgn\left(y_i - n_j\widehat{\pi}_j\right)\left[2y_i\ \ln\left(\frac{y_i}{n_i\widehat{\pi}_i}\right) + 2\left(n_i - y_i\right) ln\left(\frac{n_i - y_i}{n_i\left(1-\widehat{\pi}_i\right)}\right)\right]^{1/2}$$

where *sgn* is plus if the quantity in the parenthesis is positive and negative if the quantity is negative. It is known that the MLE is not resistant to outlying observation. Since the estimated probability relies on the initial $\hat{y}_i$ a good initial parameter is needed by using the robust method. Thus, diagnostic method based on the robDEVC was used. Even though this is a robust estimate, it down weights high leverage points (both good and bad leverage points) and not just bad leverage points. This causes the problem of swamping by identifying to many suspected outlying observation. We accommodate this problem by using a robust initial estimate, $\hat{y}_i^{rob}$. Calculate the estimated probability

$$\hat{\pi}_i^{rob} = \frac{\exp(x_i^T\hat{y}_i^{rob})}{1+\exp(x_i^T\hat{y}_i^{rob})}$$

Calculate the robDEVC

$$rdc_i = \begin{cases} 2\log\left(\frac{1}{1-\hat{\pi}_i^{rob}}\right) & if\ y_i = 0 \\ 2\log\left(\frac{1}{\hat{\pi}_i^{rob}}\right) & if\ y_i = 1 \end{cases}$$

Bad leverage points occurs when the following condition are satisfied:

(1) If $y_i = 1,\ \hat{\pi}_i^{rob} \to 0$
(2) If $y_i = 1,\ \hat{\pi}_i^{rob} \to 1$

3. Results and Discussion

In this study, the outliers are identified using SPR, DR, PR and robDEVC methods from the Table 1. It can be noted that SPR method shows two high leverage points, DR method shows one high leverage point, PR method shows two high leverage points and robDEVC method shows that two high leverage points.

TABLE 1. Bad Leverage Diagnostics for Cardiovascular Disease data

Obs.	SPR	DR	PR	robDEVC	Obs.	SPR	DR	PR	robDEVC
1	-0.7006	-0.7082	-0.7049	0.4014	51	-1.1503	-1.1720	-1.1239	-1.1257
2	-0.7006	-0.7082	-0.7049	0.3384	52	-0.7181	-0.7325	-0.7202	0.2913
3	1.3667	1.3939	1.3885	0.5793	53	0.9344	0.9521	0.8898	0.4014
4	-1.1503	-1.1720	-1.1239	-0.4874	54	-1.1503	-1.1720	-1.1239	-1.293
5	-0.7006	-0.7082	-0.7049	0.4267	55	1.1047	1.2225	1.0362	0.6829
6	0.1458	0.1643	-0.0002	-0.297	56	0.9344	0.9521	0.8898	0.9281
7	-0.7006	-0.7082	-0.7049	-0.7343	57	1.3842	1.3992	1.4186	1.1269
8	-0.7006	-0.7082	-0.7049	-1.0704	58	-0.7006	-0.7082	-0.7049	1.942
9	**5.9344**	0.9521	**3.8898**	0.8792	59	1.3842	1.3992	1.4186	0.8619
10	-0.7006	-0.7082	-0.7049	0.7038	60	-0.7006	-0.7082	-0.7049	1.2093
11	-0.7181	-0.7325	-0.7202	1.1254	61	-0.5479	-0.6149	-0.6184	-0.6566
12	-1.1503	-1.1720	-1.1239	1.1825	62	0.9344	0.9521	0.8898	0.7775
13	-0.7006	-0.7082	-0.7049	2.5194	63	1.3842	1.3992	1.4186	0.7222
14	0.9344	0.9521	0.8898	0.2827	64	0.9344	0.9521	0.8898	-0.6735
15	-0.7181	-0.7325	-0.7202	-0.636	65	-1.1503	-1.1720	-1.1239	-1.0023
16	-0.7181	-0.7325	-0.7202	0.5328	66	1.1047	1.2225	1.0362	0.748
17	0.9344	0.9521	0.8898	-1.3412	67	-0.7006	-0.7082	-0.7049	1.5824
18	0.9169	0.9548	0.8709	2.0716	68	-0.7181	-0.7325	-0.7202	1.528
19	-0.7181	-0.7325	-0.7202	2.7972	69	-0.7006	-0.7082	-0.7049	-1.0126
20	-0.3039	-0.3465	-0.0003	0.3529	70	-0.5479	-0.6149	-0.6184	1.1536
21	1.3842	1.3992	1.4186	**5.0744**	71	0.9344	0.9521	0.8898	-0.9976
22	-1.1503	-1.1720	-1.1239	-0.6202	72	-0.7006	-0.7082	-0.7049	1.1588
23	1.3842	1.3992	1.4186	**9.9723**	73	-0.7181	-0.7325	-0.7202	1.8606
24	-0.7181	-0.7325	-0.7202	0.132	74	1.3667	1.3939	1.3885	-0.1184
25	-0.7006	-0.7082	-0.7049	1.0686	75	-0.7181	-0.7325	-0.7202	1.5531
26	-0.7006	-0.7082	-0.7049	-0.2378	76	-1.1503	-1.1720	-1.1239	1.1674
27	-1.1679	-1.2161	-1.1482	0.4258	77	-0.7006	-0.7082	-0.7049	2.65
28	1.3842	1.3992	1.4186	-0.3712	78	1.3667	1.3939	1.3885	-1.9104
29	-0.7181	-0.7325	-0.7202	0.2494	79	1.3667	1.3939	1.3885	1.8891
30	1.3667	1.3939	**5.3885**	0.7768	80	1.3842	1.3992	1.4186	1.3931
31	-0.7006	-0.7082	-0.7049	0.4998	81	1.3842	1.3992	1.4186	-0.7975
32	0.9169	0.9548	0.8709	1.2142	82	**4.3667**	1.3939	1.3885	0.4787
33	0.9344	0.9521	0.8898	-1.9098	83	-0.7006	-0.7082	-0.7049	0.4721
34	1.3842	1.3992	1.4186	-0.7158	84	-0.7006	-0.7082	-0.7049	0.7953
35	0.9344	0.9521	0.8898	0.3246	85	1.3842	1.3992	1.4186	0.6985
36	-0.7006	-0.7082	-0.7049	0.7968	86	1.3842	1.3992	1.4186	0.6231
37	1.3667	1.3939	1.3885	2.7072	87	-0.7006	-0.7082	-0.7049	0.8357
38	0.9344	0.9521	0.8898	1.6087	88	-0.9801	-1.0846	-0.9651	0.7893
39	-1.1503	-1.1720	-1.1239	1.846	89	-0.7006	-0.7082	-0.7049	0.6782
40	-0.7006	-0.7082	-0.7049	1.136	90	1.3842	1.3992	1.4186	0.8876
41	-1.1503	-1.1720	-1.1239	-1.3963	91	-0.7006	-0.7082	-0.7049	1.304
42	1.3667	**4.3939**	1.3885	0.8734	92	-0.7006	-0.7082	-0.7049	-0.6566
43	0.9344	0.9521	0.8898	-1.5329	93	-0.7181	-0.7325	-0.7202	0.7775
44	-0.7006	-0.7082	-0.7049	1.086	94	0.1458	0.1643	-0.0002	0.7222
45	-0.7006	-0.7082	-0.7049	1.2364	95	-0.7181	-0.7325	-0.7202	1.1674
46	-0.7006	-0.7082	-0.7049	1.1289	96	-0.7006	-0.7082	-0.7049	2.65
47	-1.1503	-1.1720	-1.1239	2.5194	97	-0.7181	-0.7325	-0.7202	1.9104
48	0.1458	0.1643	-0.0002	1.3914	98	-0.1336	-0.1620	-0.0003	-0.4787
49	1.3842	1.3992	1.4186	1.8523	99	-1.1679	-1.2161	-1.1482	-0.4721
50	-0.7006	-0.7082	-0.7049	1.2696	100	1.3842	1.3992	1.4186	1.2358

The results of the iteration analysis using cardiovascular disease are shown in Table 2. Effect modification was found for cardiovascular disease as shown by the significant Wald test for iteration items. Odds ratio value is 0.0522, p-value

is 0.9110 and residual deviance value is 132.74 in cardiovascular disease data. Cardiogenic shock odds ratio is 0.0251, p-value is 0.9552 and residual deviance value is 143.04. Congestive heart complications odds ratio is 0.8536, p -value is 0.0607 and residual deviance value is 129.24.

TABLE 2. Residual deviance for cardiovascular disease using logistic regression

Variable	Estimate (β)	Pr($>|z|$) (P value)	Residual Deviance
Cvd	0.0522	0.9110	132.74
Sho	0.0251	0.9552	143.04
Chf	0.8536	0.0607	129.24

Abbreviations: cvd - Cardiovascular Disease, sho - Cardiogenic Shock, chf - Congestive Heart Complications

The scatter plot of the cardiovascular disease shown Figure 1. All the identified bad leverage points are represented in bold and the good and bad leverage points are underlined in Table 1. The results in the table suggest that the diagnostic measure using the SPR, DR, PR and robDEVC fail to identify the case 23 as bad leverage. Only robDEVC correctly identify the one bad leverage points. The Box plots in Figure 2 confirm these results.

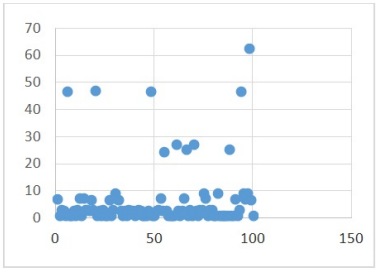

Fig 1: Scatter plot of cardiovascular disease data.

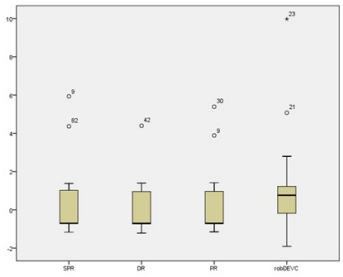

Fig 2: Box plot of SPR, DR, PR, and robDEVC cardiovascular disease data.

4. Conclusions

High leverage points contain both the good leverage points and bad leverage points. As already mentioned, the bad points are harmful to the fit while the good points improve it. There are many diagnostics available for identifying high leverage points but these methods cannot distinguish between the good and bad ones. There is still a lack of diagnostic methods on the identification of bad leverage points which can pose the real danger to the parameter estimation of the LR model.

From the results, we can observed that the outliers are identified by the SPR, DR, PR and robDEVC. The performance of the four diagnostic measures are compared. Among the four methods robDEVC method identified the bad leverage point. The numerical examples shows that the robDEVC method performed better and successfully identifies all the bad leverage point when both the good leverage and bad leverage points are present in the data.

Acknowledgment. The first author express his gratitude to the UGC for providing financial support to carry out this work under scheme UGC SAP (DRS-I). The second author acknowledges the UGC for awarding the scheme of Basic Scientific Research (BSR) Fellowship for providing financial support to carry out this work.

References

1. Belsley, D.A., Kuh, E. and Welsch., R.E.: *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*, Wiley, New York,1980.
2. Collett, D.: *Modelling Survival Data in Medical Research (second edition) ISBN 1584883251; 408 pages*,CRC Press, 2003.
3. Concato J, Feinstein AR, Holford TR.: The risk of determining risk with multivariable models, *Ann Intern Med* 118:20110, 1993.
4. Dary Pregibon.: Logistic Regression Diagnostics the Annals of Statistics, Vol. 9, No.4, 705-724, 1981.
5. Feinstein AR.: *Multivariable Analysis: An Introduction.*,Yale University Press, New Haven CT 1996.
6. Glantz SA, Slinker BK.: *Primer of Applied Regression and Analysis of Variance* McGraw-Hill, Inc., New York 1990.
7. Hosmer, D. W., Lemeshow, S. *Applied Logistic Regression.* 2nd ed. Wiley, New York, 2000.
8. Norazan, M.R., Sanizah, A., and Habshah, M.: Identification Bad Leverage Points in Logistic Regression Model Based on Robust Deviance Components: *Mathematical Models and Methods in Modern Science* 2011.
9. Sanizah Agmad, Habshah Midi and Norazan Maogamed Ramli.: (2011). Diagnostics for Residual Outliers Using Deviance Component in Binary Logistic Regression. *World Applied Science Journal* 14 8, 1125-1130.
10. Syaiba, B.A., and Habshah, M.: Robust Logistic Diagnostic for the Identification of High Leverage Points in Logistic Regression Model. *Journal of Applied Science* 10 23, 3042-3050.

K. Senthamarai Kannan: Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli-12
E-mail address: senkannan2002@gmail.com

G. Sampath*: Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli-12
E-mail address: sampathg49@gmail.com