2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Outlier Diagnostics in Logistic Regression: A Supervised Learning Technique

A. A. M. Nurunnabi[1+] and Mohammed Nasser[2]

[1] School of Business, Uttara University, Dhaka-1230, Bangladesh

[2] Department of Statistics, Rajshahi University, Rajshahi-6205, Bangladesh

Abstract. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. Logistic regression is one of the most popular supervised learning technique that is used in classification. Fields like computer vision, image analysis and engineering sciences frequently encounter data with outliers (noise). Presence of outliers in the training sample may be the cause of large training time, misclassification, and to design a faulty classifier. This article provides a new method for identifying outliers in logistic regression. The significance of the measure is shown by well-referred data sets.

Keywords: influential observation, logistic regression, outlier, regression diagnostics, and supervised learning.

1. Introduction

One of the goals of machine learning algorithms is to uncover relations among the predictors (data), to reveal new pattern and identify the underlying causes. Learning algorithm can be roughly categorized as both supervised and unsupervised. In supervised learning the goal is to predict the value of an outcome measure based on a number of input measures [8].

Logistic regression is one of the supervised machine learning technique that is used mostly for data analysis and inference. It is frequently used in epidemiology, medical imaging, computer science, electronics and electrical engineering. None of the areas having data sets to analyze without outliers. Noisy feature vectors (outliers) in the training data affect the hyperplane and a group of outliers can destroy the whole learning procedure. With the increasing use of this method different aspects of inference drawn from logistic regression are going through examinations. One of the most important issues is the estimation of parameters in the presence of unusual observations (outliers). In Logistic regression maximum likelihood (ML) method is used for estimating parameters and is extremely sensitive to 'bad' data [14]. In presence of outliers implicit assumption [4] breaks down and we have to find out the influence cases on the analyses. We discuss the idea of outliers, influential observations and diagnostics in logistic regression in section II. In section III, we present a new influence measure with numerical examples.

2. Outlier, Influential Observation and Logistic Regression Diagnostics

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. An observation is influential if it is individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations [1]. Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points after which these influential points can be removed or corrected. In binomial logistic regression outliers may occur as altercation (misclassification) between the

[+] Corresponding author. Tel.: 88-02-8932325; fax: 88-02-8918047;
E-mail address: pyal1471@yahoo.com

binary (1, 0) responses. Misclassification may refer to points which are on the wrong side of the hyperplane/classifier [17]. It may occur by meaningful deviation in predictor (explanatory) variables, which deviates the response (labels).

Let us consider the standard logistic regression model:

$$E(Y \mid X) = \pi(X)\ ; \quad \text{where} \quad \pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}},\ \ 0 \le \pi(X) \le 1. \tag{1}$$

This form gives an S-curve configuration. The well-known 'logit' transformation in terms of $\pi(X)$ is

$$g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = X\beta, \tag{2}$$

where X is an n × k matrix (k = p + 1), Y is an n × 1 vector of binary responses, β is the parameter vector.

For the model $Y = \pi(X) + \varepsilon$, the error term $\varepsilon = \begin{cases} 1-\pi(X) & w.p.\ \pi(X); if \quad y = 1 \\ -\pi(X) & w.p. 1-\pi(X); if \quad y = 0 \end{cases}$ has a distribution with mean zero and variance $\pi(X)[1-\pi(X)]$. The ε violates the least squares (LS) assumptions and the ML method based on iterative reweighed least squares (IRLS) is used to estimate the parameters. ML method is very sensitive to outlying responses and extreme points in design (predictor) space X.

Among the large body of literature (see [9, 14, 15]) Cook's distance (CD) [5] and DFFITS [1] have become very popular. For the logistic regression, ith Cook's distance is defined [15] as

$$\mathrm{CD}_i = \frac{\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)^T (X^T V X)\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)}{k\hat{\sigma}^2}, \qquad i = 1,2,\ldots,n \tag{3}$$

where $\hat{\beta}^{(-i)}$ is the estimated parameter of β without the ith observation. Using Pregibon's [14] linear regression like approximations (3) can be reexpressed as

$$\mathrm{CD}_i \approx \frac{1}{k} r_{si}^{\,2}\left(\frac{h_{ii}}{1-h_{ii}}\right); \text{where } r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1-h_{ii})}}, \quad i = 1,2,\ldots,n \tag{4}$$

r_{si} is the ith standardized Pearson residual , h_{ii} is the ith leverage value and v_i is the ith diagonal element of V,

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2} \text{ and } V(y_i \mid x_i) = v_i = \hat{\pi}_i(1-\hat{\pi}_i).$$

Observation with CD value greater than 1 is treated as an influential. DFFITS is expressed in terms of standardized Pearson residuals and leverage values as

$$\mathrm{DFFITS}_i = r_{si}\sqrt{\frac{h_{ii}}{(1-h_{ii})}\frac{v_i}{v_i^{(-i)}}}, \qquad i = 1,2,\ldots,n. \tag{5}$$

An influential observation has DFFITS value larger than c $\sqrt{k/n}$, where c is between 2 and 3.

3. New Diagnostic Measure

We develop a single-deletion influence measure [11] and name this measure as squared difference in beta (SDFBETA). We introduce the newly proposed measure as

$$SDFBETA_i = \frac{(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)^T (X^T V X)(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)}{v_i^{(-i)}(1-h_{ii})}, \tag{6}$$

where $v_i^{(-i)}(1-h_{ii})$ is the variance of the ith residual after deleting ith observation. We use confidence bound type cut-off value (see [12]), and consider the ith observation to be influential if

$$\left|\mathrm{SDFBETA}_i\right| \ge \frac{9k}{n-3p}. \tag{7}$$

Example: ***Brown Data***

To show the performance of the proposed SDFBETA with CD and DFFITS, we consider a part of Brown et al. [3] data. The original objective was to see whether an elevated level of acid phosphates (A.P.) in the blood serum would be of value as an additional regressor (predictor) for predicting whether or not prostate cancer patients also had lymph node involvement (L.N.I). The label (dependent) variable is nodal involvement, with 1 denoting the presence and 0 indicating the absence of involvement. Figure 1 identifies an unusual observation (case 24) among the patients without L.N.I. We apply new measure for the identification of the influential case. Table 1 shows that SDFBETA with Cook's distance and DFFITS successfully identifies the case 24 as an influential.

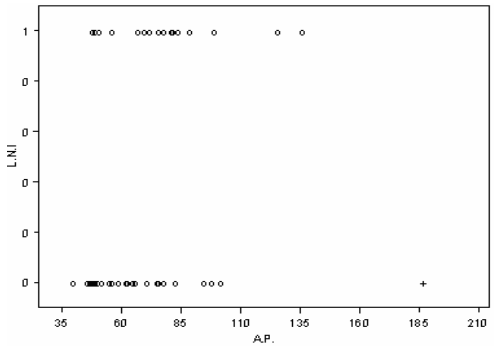

Fig.1 Scatter plot of acid phosphates versus nodal involvement

Table 1 Diagnostic measures for Brown data with one outlier

Ind	CD (1.00)	DFFITS (0.582)	SDFBTA (0.353)	Ind	CD (1.00)	DFFITS (0.582)	SDFBTA (0.353)	Ind	CD (1.00)	DFFITS (0.582)	SDFBTA (0.353)
2	0.005	-0.104	-0.107	20	0.025	-0.226	-0.236	38	0.032	-0.254	-0.267
3	0.006	-0.106	-0.109	21	0.006	-0.105	-0.108	39	0.008	-0.123	-0.125
4	0.006	-0.105	-0.108	22	0.007	-0.120	-0.122	40	0.021	-0.206	-0.214
5	0.006	-0.106	-0.109	23	0.025	0.222	0.232	41	0.006	-0.106	-0.108
6	0.006	-0.106	-0.110	24	**2.075**	**-2.149**	**-3.619**	42	0.018	0.188	0.193
7	0.006	-0.108	-0.112	25	0.044	0.294	0.343	43	0.017	0.184	0.188
8	0.005	-0.104	-0.106	26	0.017	0.185	0.190	44	0.016	0.180	0.184
9	0.025	0.228	0.234	27	0.006	-0.110	-0.115	45	0.016	0.182	0.186
10	0.005	-0.104	-0.107	28	0.006	0.106	0.109	46	0.016	0.181	0.185
11	0.005	-0.104	-0.106	29	0.006	-0.106	-0.109	47	0.016	0.182	0.186
12	0.006	-0.112	-0.114	30	0.006	-0.110	-0.115	48	0.017	0.187	0.191
13	0.006	-0.105	-0.108	31	0.005	-0.104	-0.107	49	0.017	0.185	0.190
14	0.017	0.187	0.191	32	0.005	-0.104	-0.106	50	0.017	0.187	0.191
15	0.006	-0.107	-0.111	33	0.038	0.280	0.289	51	0.016	0.181	0.184
16	0.006	-0.106	-0.110	34	0.033	0.259	0.266	52	0.020	0.198	0.204
17	0.006	-0.106	-0.109	35	0.036	0.272	0.281	53	0.040	0.281	0.316
18	0.008	-0.128	-0.131	36	0.006	-0.107	-0.110				

3.1. Generalized squared difference in beta

A group of (multiple) influential observations may distort the fitting of a model in such a way that influence measures may suffer masking and/or swamping. We develop a generalized group-deleted version of the residuals and weights that will be effective diagnostics for the identification of multiple influential observations and free from masking and swamping problems. We name the group-deletion measure as generalized squared difference in beta (GSDFBETA). The stepwise method is as follows.

Step 1: At the first step we try to find out all suspect influential cases. Some times graphical displays like index plot and character plot of explanatory and response variable could give us an idea about the influential observations, but these plots are not always helpful for higher dimension of regressors. We suspect that influential observations are potential outliers or high leverage points or both. Hence we can compute the generalized standardized Pearson residuals [10] and/or some leverage measures [9] to identify the suspect influential cases. In general, we prefer using any suitable robust techniques LMS and LTS [16], the BACON [2] or local influence measures [6] to find all suspect influential cases and to form the deletion set D.

Step 2: We assume that d observations among a set of n observations are deleted as the suspected cases. Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence R

contains (n-d) cases after d cases are deleted. Without loss of generality, X, Y and V (variance-covariance matrix) are

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

Let $\hat{\beta}_{(R)}$ be the corresponding vector based on R. The fitted values for the entire logistic regression model based on R set are defined as

$$\hat{\pi}_{i(R)} = \frac{\exp\left(x_i^T \hat{\beta}_{(R)}\right)}{1+\exp\left(x_i^T \hat{\beta}_{(R)}\right)}, \qquad i = 1,2,...,n. \tag{8}$$

Here we define the ith deletion residual and the corresponding variance as

$$\hat{\varepsilon}_{i(R)} = y_i - \hat{\pi}_{i(R)} \text{ and } v_{i(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}). \tag{9}$$

Hence the ith diagonal element of the leverage matrix can be expressed as

$$h_{ii(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}) x_i^T \left(X_R^T V_R X_R\right)^{-1} x_i \qquad i = 1,2,...,n. \tag{10}$$

We define the generalized squared difference in beta (GSDFBETA) in presence of multiple influential cases,

$$GSDFBETA_i = \begin{cases} \dfrac{\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)^T (X_R^T V_R X_R)\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)}{v_{i(R-i)}\left(1-h_{ii(R)}\right)}; & i \in R \\ \dfrac{\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)^T (X_D^T V_D X_D)\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)}{v_{i(R)}\left(1-h_{ii(R+i)}\right)}; & i \notin R. \end{cases} \tag{11}$$

To identify the multiple outliers Imon and Hadi [10] introduce the GSPR as

$$r_{si(R)} = \begin{cases} \dfrac{y_i - \hat{\pi}_{i(R)}}{\sqrt{v_{i(R)}\left(1-h_{ii(R)}\right)}} & \text{for} \quad i \in R \\ \dfrac{y_i - \hat{\pi}_{i(R)}}{\sqrt{v_{i(R)}\left(1+h_{ii(R)}\right)}} & \text{for} \quad i \in D. \end{cases} \tag{12}$$

Based on the R, let us define the generalized weights (GW) denoted by h_{ii}^* [11],

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} & \text{for} \quad i \in R \\ \dfrac{h_{ii(R)}}{1+h_{ii(R)}} & \text{for} \quad i \in D. \end{cases} \tag{13}$$

Now the GSDFBETA can be re-expressed in terms of GSPR and deleted leverages h_{ii}^* as

$$GSDFBETA_i = \begin{cases} \dfrac{h_{ii}^* r_{si(R)}^2}{1-h_{ii(R)}} \ for & i \in R \\ \dfrac{h_{ii}^* r_{si(R)}^2}{1+h_{ii(R)}} \ for & i \notin R. \end{cases} \tag{14}$$

We make a relationship between GSDFBETA with GDFFITS [13] in [11]. We consider the ith observation to be influential as like as GSDFBETA [12] in linear regression,

$$\left|GSDFBETA\right| \geq \frac{(3\sqrt{k/(n-d)})^2}{1-[3p/(n-d)]} = \frac{9k}{n-d-3p}. \tag{15}$$

4. Example: Modified Brown Data

We modify the Brown et al. [3] data by putting two more unusual observations, cases 54 (200, 0) and 55 (220, 0). Now we apply our newly proposed measure GSDFBETA. The method identifies all of the three suspect cases (Figure 2) as influential properly. Table 2 shows, besides the 3 cases it identifies one more (case 38) as influential, which was masked before the identification of the 3. We show GDFFITS [11] gives the same result.

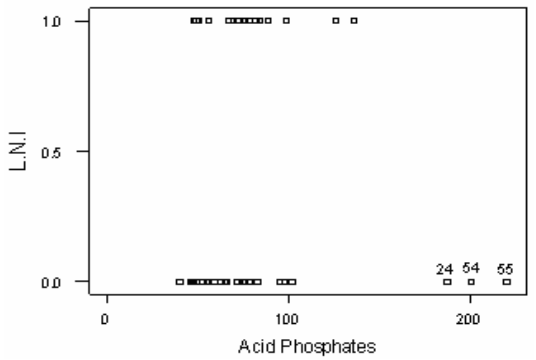

Fig. 2 Scatter plot of L.N.I. against acid phosphates for Imon and Hadi's modified Brown data

Table 2. Proposed diagnostic measures for modified Brown data; three outlier

Index	GSDFBETA(0.367)	Index	GSDFBETA(0.367)	Index	GSDFBETA(0.367)	Index	GSDFBETA(0.367)
1	0.0097	15	0.0096	29	0.0100	43	0.0377
2	0.0107	16	0.0099	30	0.0083	44	0.0358
3	0.0100	17	0.0100	31	0.0106	45	0.0372
4	0.0102	18	0.0370	32	0.0112	46	0.0364
5	0.0100	19	0.0629	33	0.1613	47	0.0372
6	0.0099	20	0.2694	34	0.1264	48	0.0406
7	0.0094	21	0.0102	35	0.1488	49	0.0381
8	0.0120	22	0.0274	36	0.0097	50	0.0406
9	0.0836	23	0.0399	37	0.0123	51	0.0360
10	0.0106	**24**	**4.3782**	**38**	**0.3710**	52	0.0410
11	0.0120	25	0.0082	39	0.0302	53	0.0150
12	0.0192	26	0.0381	40	0.2076	**54**	**5.5408**
13	0.0133	27	0.0083	41	0.0139	**55**	**7.5386**
14	0.0406	28	0.0100	42	0.0391		

5. References

[1] D. A. Belsley, E. Kuh, and R. E. Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity, 1980, *Wiley*, New York.

[2] N. Billor, A. S. Hadi, and F. Vellman. BACON: Blocked adaptive computationally efficient outlier nominator, Computat. Statist. *Data Anal.*, 2000, 34, pp. 279-298.

[3] B. W. Brown Jr.. Prediction analysis for binary data, in Biostatistics Casebook, R.G. Miller, Jr. B. Efron, B. W. Brown, Jr., L.E. Moses, Eds., 1980, *Wiley*, New York.

[4] S. Chatterjee, and A. S. Hadi. Sensitivity Analysis in Linear Regression, 1988, *Wiley*, New York.

[5] R. D. Cook. Detection of influential observations in linear regression, *Technometrics*, 1977, 19 pp. 15-18.

[6] R. D. Cook. *Assessment of local influence,* J. Roy. Stat. Soc., Ser – B, 1986, 48, pp. 131 – 169.

[7] A. S. Hadi, and J. S. Simonoff, Procedure for the identification of outliers in linear models, J. Amer. *Statist. Assoc.*, 1993, 88, pp. 1264 – 1272.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, 2001, Springer,* New York.

[9] D. W. Hosmer, and S. Lemeshow. *Applied Logistic Regression*, 2000, Wiley, New York.

[10] A. H. M. R. Imon, and A. S. Hadi. Identification of multiple outliers in logistic regression, Comm. Statist., *Theory and Meth.*, 2008, 37, pp. 1667 – 1709.

[11] A. A. M. Nurunnabi. Robust Diagnostic Deletion Techniques in Linear and Logistic Regression, *M. Phil Thesis*

(Unpublished), 2008, University of Rajshahi, Bangladesh.

[12] A. A. M. Nurunnabi, A. H. M. Rahmatullah Imon, and M. Nasser. A New Influence Statistic in Linear Regression, Proc. Of Intl. *Conference on Statistical Sciences,* (ICSS 08), December 2008, Dhaka, Bangladesh. pp. 165-173.

[13] A. A. M. Nurunnabi, A. H. M. Rahmatullah Imon, and M. Nasser. Identification of multiple influential observations in logistic regression, *Journal of Applied Statistics (Under review).*

[14] D. Pregibon, Logistic regression diagnostics, *Ann. Statist.*, 1981, 9, pp. 977-986.

[15] T. P. Ryan. Modern Regression Methods, 1997, *Wiley*, New York.

[16] P. J. Rousseeuw, and A. Leroy. Robust Regression and Outlier Detection, 2001, Wiley, New York.

[17] B. Scholkopf, and A. J. Smola. Learning with Kernels, 2002, MIT press, *Cambridge, Massachusetts.*