University Of AL-Qadisiyah

Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

Robust logistic regression in the presence of high leverage points

Mohammed A. Mohammed

Department of Accounting Techniques Al-Dewanyia Technical Institute - Al-Furat Al-awsat Technical University.

Email : *dw.moh2@atu.edu.iq*

ARTICLE INFO

Article history:
Received: 01 /04/2019
Rrevised form: 17 /04/2019
Accepted : 23 /04/2019
Available online: 30 /05/2019

Keywords:

Logistic regression model, MVE, RMD, SGD and HLPs

ABSTRACT

In this article we conceder the logistic regression model with high leverage points. For the logistic regression model with a binary response, we suggested a new robust approach called robust logistic regression (RLR) based on the robust mahalanobis distance (RMD) which depends on the minimum volume ellipsoid (MVE) estimators. The RMD is computed by using the algorithm of stochastic gradient descent (SGD). In order to assist the new suggested approach we compare it with some existing method such as maximum likelihood estimator and robust M-estimator in logistic regression model. The simulation study points that the RLR has supreme performances throw some measurement comparison.

1. Introduction

Many types of outlying data can occur in the logistic regression model (LRM). Outliers can appear in the Y- direction, X- direction or in both directions. This study focuses in the case of outliers in the X-direction which called high leverage points (HLPs). It is important to distinguish between two types of leverage points, good leverage points (GLPs) and bad leverage points (BLPs) (see; [1], [10]). GLPs in logistic regression appear when outcomes equal to one $(y = 1)$ with large value of x's or when the outcome equal to zero $(y = 0)$ with small value x's and vice versa for BLPs [14]. GLPs have no effect in the estimators and it may improve the solution, whereas, BLPs have high affect in the estimators and it may lead to misclassification ([14], [17]). The maximum likelihood estimator (ML- estimator) is commonly used to estimate the LRM by using Newton Raphson numerical method.

Corresponding author Mohammed A. Mohammed

Email addresses: *dw.moh2@atu.edu.iq*

Communicated by Qusuay Hatim Egaar

Unfortunately, the ML estimator is sensitive to outliers and HLPs and it can easily influenced by outlying observations (see; [12], [13]).

In literature, there is a massive number of works to robustness of the ML estimators in the LRM. Most of researchers attempt to accomplish robustness by down weighting unusual values (see; [1], [2], [3], [7] and [12] and [16]). Pregibon (1982) proposed new robust fitting methods which taper the standard likelihood to decrease an effect of outliers [13]. Kordzakhia et al. (2001) suggested a resistant method by minimize the mean-squared deviance for the worst extreme observations [11]. Hobza et al. (2008) suggested a new approach by taken a robust median estimator in LRM [9]. Bianco and Yohai suggested a new class of robust M-estimates for the LRM. They illustrate that these estimates are consistent and asymptotically normal [4].

Another method for estimate the ML estimator is the stochastic gradient descent (SGD). The SGD is an iterative and efficient numerical optimization approach ([5], [15]). It tries to find such a minimum *x* by using information from the first derivative of function. SGD is almost never as fast as Newton Raphson method but it is much more robust ([5], [14]). Moreover, it does not require that the second derivative as Newton Raphson [5]. In many cases, SGD is so robust even that it is not a hard requirement. In addition, gradient descent typically has a much larger region of convergence than Newton Raphson.

This article is setting as follows: the logistic regression model is briefly explained in Section 2. In Section 3, the robust logistic regression methods and the proposed method are given. The simulation study is introduced in Section 4. The results of simulation experiments are discussed in Section 5. Finally, in Section 6, the conclusions are explained.

2. Logistic Regression Model

The Logistic Regression model (LRM) is firstly developed by the statistician David Cox in 1958 to solve the problems which were not directly suited for linear regression model [7]. LRM is considered a special case of linear regression model when the response is a binary variable (False/ true, Yes/ No, 0/1, Male/Female… etc.), whereas, a set of explanatory variables can be a discrete (categorical, nominal or ordinal), or continuous [12]. For the following multiple linear regression model [4]:

$$y_j = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_j \qquad \ldots 1$$

where $\varepsilon_j \sim IND(0, \sigma^2)$, $j = 1, 2, \ldots, n$

Let y_j be a binary response variable (0/1), where a 1 represent a success and 0 represent a failure [18]. Hence, y_j is distributed as Bernoulli distribution with $E(y_j) = \pi_j$ and $var\left(y_j\right) = \pi_j(1 - \pi_j)$. Unlike of linear regression model, it clearly to see the variance of LRM could be potentially different for each value of (y_j) [1]. Moreover, in LRM, we are only dealing with the probability of outcome of response variable, success (p) and failure $(1 - p)$. The p should be satisfying the following conditions:

1- $p > 0$, it always must be positive, and

2- $p \leq 1$, it always must be less than or equal to1

The LRM can be also expressed as:

$$\begin{aligned} \text{logit}\left(\pi_j\right) &= \log\left\{\frac{\pi_j}{1 - \pi_j}\right\} \\ &= x_j^T \beta \qquad \ldots 2 \end{aligned}$$

where

$$\begin{aligned} \pi_j &= pr\left(Y_j = 1 \middle| X_j = x_j\right) \\ &= \exp(x_j^T \beta)/\{1 + \exp\left(x_j^T \beta\right)\} \quad \ldots 3 \end{aligned}$$

where x is an $n \times (k + 1)$ explanatory variables and β is a $(k + 1) \times 1$ of regression coefficients and the letter T refers to transpose. From Equation (2), it is clearly to see the values of π_j falls within the interval $(0 < \pi_j < 1)$. The β describes the relationship between x_j and π_j .When $\beta > 0$, there is a positive relationship between x_j and π_j and the curve will increase toward $\pi = 1$ as x increase as shown in Figure (1.a). Figure (1.b) shows that when $\beta < 0$, the relationship between x_j and π_j is negative and the curve will increase toward $\pi = 1$ as x decrease. The ML estimator is widely used to estimate the parameter of the model. The must of algorithms apply the Newton-Raphson approach to solve the ML estimator ([12], [13]). The ML estimator is identified by the following objective function

$$l(x; \beta) = \sum_{i=1}^{n} [y_i \, \gamma_i] \qquad \ldots 4$$

where $\gamma_i = ln(\pi_i) + (1 - y_i)\ ln(1 - \pi_i)$

The iterative reweighted least squares (IRLS) for ML estimator are defined as [3]:

$$\hat{\beta}_{ML} = (X^T W X)^{-1}(X^T W z) \qquad \ldots 5$$

where $W = \hat{\pi}_j(1-\hat{\pi}_j)$ is a diagonal matrix and $z = x^T\beta$.

Pregibon (1982) suggested a new robust approach by modify the objective function of ML estimator to be more resistant for outlying data in the LRM by minimize the following function [13]

$$\sum \rho(\gamma_i(\beta)) \qquad \ldots 6$$

where ρ is a bounded, differentiable and a decreasing function given as

$$\rho(x) = \begin{Bmatrix} x - \left(\frac{x^2}{2c}\right) & if \quad x \le c \\ \frac{c}{2} & if \quad x > c \end{Bmatrix} \qquad \ldots 7$$

where, c is a positive number [4].

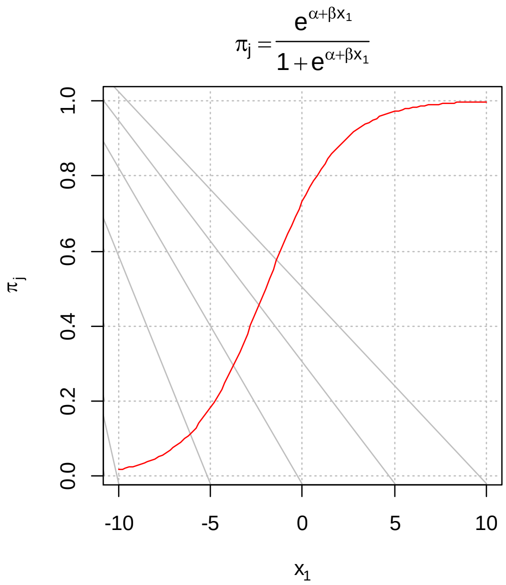


Figure (1.a): relationship between **x_j and π_j**, when $\boldsymbol{\alpha = 1,\ \beta = 0.5}$

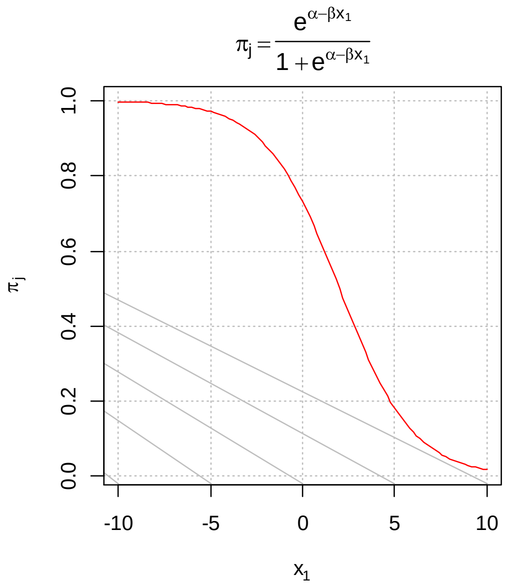


Figure (1.b): relationship between **x_j and π_j**, when $\boldsymbol{\alpha = 1,\ \beta = -0.5}$

3. Robust logistic regression method

Many works have been done to robustness the ML estimator in the binary logistic regression (see; [3], [6], [13], [17] and [18]). In this work we suggested a new robust logistic regression (RLR) based on robust ML estimator by using robust Mahalanobis distance (RMD). The RMD is computed by using the algorithm of stochastic gradient descent (SGD). The RMD depends on the reweighted minimum volume ellipsoid (MVE) estimator. The MVE approach is suggested by Rousseeuw and Van zameron in 1999 which known as a high robust approach for multivariate location and shape parameters [14]. The MVE estimator is based on the smallest volume ellipsoid that covers h out of n observations. It is known as low bias and an affine equivariant, high breakdown robust estimator of multivariate location and shape. When we estimate the coefficients of LRM, HLPs should have large values of mahalanobis distances (MD), given as (see; [1], [14] and [17]):

$$MD_i^{\ 2} = (x_i - \hat{\mu})^T \hat{\Sigma}^T (x_i - \hat{\mu}) \qquad \dots 8$$

where, $\hat{\mu}$ is a vector of sample mean, $\hat{\Sigma}$ is a sample variance- covariance matrix. To determine the cut-off point value for the MD, we usually assume that $\mu = 0$ and $\Sigma = I$, because MD are invariant under affine transformation [4]. The classical tolerance ellipsoid is given by;

$$MD_i \leq \sqrt{\chi^2{}_{d,0.975}} \quad \dots 9$$

The 97.5% quantile of the χ^2 distributed with d degree of freedom therefor, about 97.5% of observations belong to the ellipsoid. The classical MD is highly influenced by the presence of HLPs due to it depend on the classical estimates of location and covariance matrix [14]. In order to robustness the MD to be more resistant for HLPs, we consider robust estimators based on the MVE for mean and variance- covariance matrix. Then, the robust mahalanobis distance (RMD) is defined as following ([1], [4])

$$RMD_i^{\ 2} = (x_i - \hat{\mu}_{MVE})^T \hat{\Sigma}_{MVE}^T (x_i - \hat{\mu}_{MVE}) \quad \dots 10$$

where $\hat{\mu}_{MVE}$ and $\hat{\Sigma}_{MVE}$ are robust locations and shape estimates of the MVE, respectively.

Assume J is a set of h points where, $\{|\hat{\Sigma}_J| \leq |\hat{\Sigma}_K|$ for all subsets $K, K \neq h\}$, h=p+1, p is the number of explanatory variables, then (see; [1], [4] and [13]);

$$\hat{\mu}_{MVE} = \left[\sum_{i\in J} x_i\right]/h \qquad \dots 11$$

$$\hat{\Sigma}_{MVE} = \left[\sum_{i\in J}(x_i - \hat{\mu}_{MVE})^T (x_i - \hat{\mu}_{MVE})\right]/h \dots 12$$

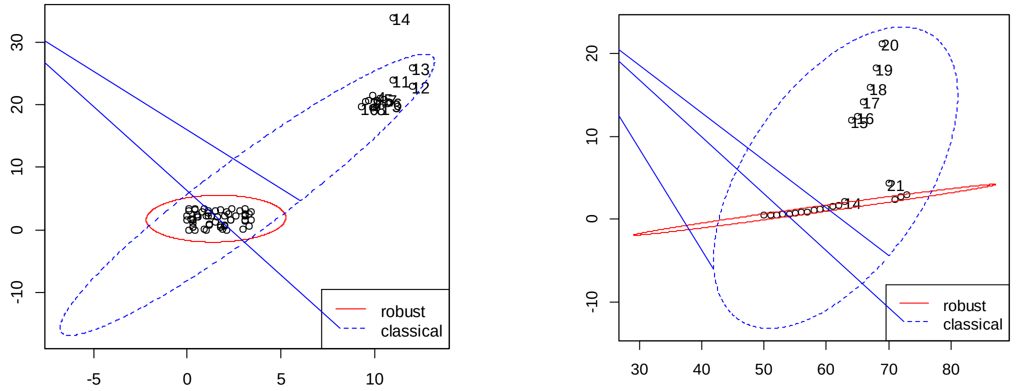
By following weighted Bianco and Yohai (WBY) which defined as [4]:

$$\hat{\gamma}_n = \underset{\gamma}{argmin} \quad \sum_{i=1}^{n} \omega_i \varphi_{BY}(z', \gamma; y) \quad \ldots 13$$

where φ_{BY} is the function of the Bianco and Yohai (BY) estimator and the ω_i is weighted function computed as following [4]:

$$\omega_i = \begin{Bmatrix} 1 & if & RMD \le \chi^2_{p,0.975} \\ 0 & & else \end{Bmatrix} \quad \ldots 14$$

The HLPs are determined by using *RMD* and the estimation subset is identified by *MVE*. Filzmoser et al. [2005] proposed a robust threshold based on the *MVE* robust estimator, adaptively from the data [8]. The procedure of *MVE* threshold is by finding the maximum positive between the theoretical distribution of chi-square and the empirical distribution of robust distance. Figure 2 demonstrate plots the 0.975 tolerance ellipse of the bivariate data set *x*. The ellipse is defined by those data points whose distance is equal to the square root of the 0.975 chi square quantile with 2 degrees of freedom. The robust ellipse is more resistant than classical ellipse for outliers in both types of datasets (Hawkins Brado Kass dataset and telephone dataset [2]).



Hawkins, Bradu, Kass data set **Telephone dataset**

Figure (2): Tolerance ellipse (97.5%) for Hawkins Bradu Kass and Telephone datasets

4. Monte Carlo Simulation Study

In this section, a Monte Carlo simulation study is applied to assess the performance of the proposed RLR method and compare it with some of the existed methods such as ML estimator and robust M-estimator in LRM. The LRM is generated by following Croux and Haesbroeck (2003) [6], with two predictors normally distributed, x_1 and $x_2 \sim N(0,1)$ and the response variable is defined as:

$$\mathrm{y_i} = \begin{Bmatrix} 0 \text{ if } & X^T\beta < 0 \\ 1 \text{ if } & X^T\beta \geq 0 \end{Bmatrix}, \mathrm{i} = 1,2, \dots , \mathrm{n} \quad \dots 15$$

We assumed the vector of true parameter is equal to (1, 1.5, 2) with sample size $n = 50, 100$ and 200. The error term ε_{i} is generated as logistic distribution $\varepsilon_{\mathrm{i}} \sim \Lambda(0,1)$.

To test the robustness of the methods, we contaminated the generated data by different percentage high leverage points in the explanatory variables such as τ: where $\tau = 0\%, 1\%, 5\%$ and 10%. The high leverage points denoted as x_i^* are generated according to following formula:

$x_{ij}^* = x_{ij} + \vartheta, i = 1,2, \dots , n$ and $j = 1,2$ ...16

Where, we supposed that the value of ϑ is equal to 10 to make a high leverage points in the explanatory variables. All of the simulation experiments are run included 1,000 replications for convergence. The bias and the mean squared error (MSE) were used to assess the performance of the methods. The bias and MSE are respectively defined as [1]:

$$Bias = \left\| \frac{1}{r} \sum_i \grave{\beta}_i - \beta \right\|, r = 1,2, \dots ,1000 \quad \dots 17$$

$$MSE = \left\| \frac{1}{r} \sum_i \grave{\beta}_i - \beta \right\|^2, r = 1,2, \dots ,1000 \quad \dots 18$$

where ‖ . ‖ is the Euclidean norm [6].

5. Results and Interpretation

Good estimators are having small values of bias and MSE or close to zero. The bias and the MSE for the methods of the study are demonstrated in Tables 1-4. Good estimators are having small values of bias and MSE or close to zero. From Table 1 with zero percentage of contaminated (clean data), we can see the Ml-estimator has relatively good performance in deferent size of samples compared with others due to there is no HLPs in the simulated data. Also we can see the M-estimator and RLR methods are fairly closed to Ml-estimator. Tables 2-4 show that the ML-estimator is destroyed because of the data set has HLPs. The bias and MSE of the ML- estimates were immediately affected by 1% percentage of contaminated of simulated data. Moreover, the ML-estimator becomes worst when the percentage of contaminated data increase. The RLR method has the best performance through it has the lowest values of bias and MSE.

Table1: *MSE* and *bias* values for all estimators with clean data

n	*Criteria*	ML	M-estimator	RLR
30	*Bias*	0.0862	0.0884	0.0878
	MSE	0.2644	0.2668	0.2651
50	*Bias*	0.0837	0.0860	0.0851
	MSE	0.2620	0.2636	0.2627
100	*Bias*	0.0825	0.0853	0.0838
	MSE	0.2590	0.2614	0.2605
200	*Bias*	0.0811	0.0842	0.0830
	MSE	0.2583	0.2598	0.2592

Table2: *MSE* and *bias* values for all estimators with 1% HLPs

n	*Criteria*	ML	M-estimator	RLR
30	*Bias*	0.3385	0.1015	0.0991
	MSE	0.5311	0.2809	0.2798
50	*Bias*	0.3331	0.0988	0.0970
	MSE	0.4889	0.2783	0.2777
100	*Bias*	0.3055	0.0950	0.0935
	MSE	0.3030	0.2763	0.2749
200	*Bias*	0.2985	0.0957	0.0889
	MSE	0.2797	0.2735	0.2724

Table3: *MSE* and *bias* values for all estimators with 5% HLPs

n	*Criteria*	ML	M-estimator	RLR
30	*Bias*	2.3277	0.8198	0.8167
	MSE	2.6395	0.8691	0.8674
50	*Bias*	2.2212	0.8176	0.8155
	MSE	2.6335	0.8484	0.8463
100	*Bias*	2.1015	0.8132	0.8117
	MSE	2.6100	0.8460	0.8441
200	*Bias*	1.9089	0.8124	0.8109
	MSE	1.9392	0.8458	0.8432

Table4: *MSE* and *bias* values for all estimators with 10% HLPs

n	*Criteria*	ML	M-estimator	RLR
30	*Bias*	6.3277	0.9211	0.9166
	MSE	4.6395	0.9750	0.9674
50	*Bias*	6.0012	0.9181	0.9154
	MSE	4.0335	0.9711	0.9460
100	*Bias*	4.1015	0.9159	0.9117
	MSE	3.6100	0.9683	0.9442
200	*Bias*	3.9089	0.9135	0.9108
	MSE	3.4392	0.9679	0.9432

6. Conclusions

In this study we proposed a new robust technique to tackle the problem of presence of high leverage points in the logistic regression model. The suggested method depends on the robust mahalanobis distance. The robust mahalanobis distance is computed by using *MVE* based on stochastic gradient descent. The simulation study in different size of samples and different percentage of contaminated by high leverage points were used to examine the performance of suggested method. The results of simulation experiment show that the proposed method has a supreme performance when the logistic regression model contains high leverage points.

7. References

1- Ahmad, S., Ramli, N.M. and Midi, H., 2010. Robust estimators in logistic regression: A comparative simulation study. *Journal of Modern Applied Statistical Methods*, *9*(2).

2- Alguraibawi, M., Midi, H. and Imon, A.H.M., 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*.

3- Ariffin, S.B. and Midi, H., 2014. Robust Logistic Ridge Regression Estimator in the Presence of High Leverage Multicollinear Observations. In *16th Int. Conf. Math. Comput. Methods Sci. Eng*.

4- Bianco, A.M. and Yohai, V.J., 1996. Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods* (pp. 17-34). Springer, New York, NY.

5- Bottou, L., 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg.

6- Croux, C. and Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. *Computational statistics & data analysis*, *44*(1-2), pp.273-295.

7- Cox, D.R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), pp.215-232.

8- Filzmoser, P., 2005. Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, *34*(2), pp.127-138.

9- Hobza, T., Pardo, L. and Vajda, I., 2008. Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference*, *138*(12), pp.3822-3840.

10- Hubert, M., Rousseeuw, P.J. and Van Aelst, S., 2008. High-breakdown robust multivariate methods. *Statistical science*, pp.92-119.

11- Kordzakhia, N., Mishra, G.D. and Reiersølmoen, L., 2001. Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, *98*(1-2), pp.211-223.

12- Pregibon, D., 1981. Logistic regression diagnostics. *The Annals of Statistics*, *9*(4), pp.705-724.

13- Pregibon, D., 1982. Resistant fits for some commonly used logistic models with medical application. *Biometrics*, *38*(2), pp.485-498.

14- Rousseeuw, P.J. and Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, *85*(411), pp.633-639.

15- Ruder, S. (2016). An overview of gradient descent optimization algorithms. *http:// ruder.io/optimizing-gradient-descent/*

16- Shalev-Shwartz, S. and Tewari, A., 2011. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, *12*(Jun), pp.1865-1892.

17- Syaiba, B.A. and Habshah, M., 2010. Robust logistic diagnostic for the identification of high leverage points in logistic regression model. *Journal of Applied Sciences*, *10*(23), pp.3042-3050.

18- Tabatabai, M.A., Li, H., Eby, W.M., Kengwoung-Keumo, J.J., Manne, U., Bae, S., Fouad, M. and Singh, K.P., 2014. Robust logistic and probit methods for binary and multinomial regression. *Journal of biometrics & biostatistics*, *5*(4).