Journal of Statistics Applications & Probability

Volume 13
Issue 2 *Mar. 2024*

Article 7

3-1-2024

Robust Estimation of Multiple Logistic Regression Model

Ehab A. Mahmood
Department of Banking and Finance, College of Administration and Economics, University of Babylon, Babil, Iraq, ehab.mahmood@uobabylon.edu.iq

Recommended Citation

A. Mahmood, Ehab (2024) "Robust Estimation of Multiple Logistic Regression Model," *Journal of Statistics Applications & Probability*: Vol. 13: Iss. 2, Article 7.
DOI: https://dx.doi.org/10.18576/jsap/130205
Available at: https://digitalcommons.aaru.edu.jo/jsap/vol13/iss2/7

Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/130205

Robust Estimation of Multiple Logistic Regression Model

Ehab A. Mahmood

Department of Banking and Finance, College of Administration and Economics, University of Babylon, Babil, Iraq

Received: 27 May 2023, Revised: 29 Jul. 2023, Accepted: 5 Oct. 2023.
Published online: 1 Mar. 2024.

Abstract: The multiple logistic regression model is commonly used in scientific researches. It is a regression model with more than two categories of response variable and multi explanatory variables. The conventional maximum likelihood estimator (MLE) is widely used to determine parameter values. It is used because it is famous and easy to apply. However, this estimator is highly sensitive to leverage points and outliers. The main objective of this research is to get the best estimation of the multiple logistic regression parameters with problem of leverage points. Two robust estimators based on robust Mahalanobis distance (RMD) are established. They are named (MLERMD1 and MLERMD2). The proposed robust methods are compared with MLE and some other famous robust methods. The bias and mean square error are considered as measures for comparison. Simulation study is conducted with different sample sizes and percentages of leverage points. Besides, real example data are applied to compare among the methods. Results of simulation and real example show that the performance of the proposed methods (MLERMD1 and MLERMD2) is more efficient than those of MLE and the other robust methods. The MLERMD2 have the least values of bias and mean square error with different percentage of leverage points.

Keywords: Logistic regression model; Maximum likelihood estimator; Leverage points; Outliers, Robust estimation.

1 Introduction

The multiple logistic regression model is widely used in the field of medical and behavioral sciences. The model considers response variable with more than one explanatory variable. The logistic regression model has two types, either nominal logistic (that response variable has more than two categories) or binary logistic (that response variable has only two categories). The binary logistic regression model is assumed that the response variable (Y) follows Bernoulli distribution which takes (1) for occurrence and (0) for non-occurrence with unknown distribution of explanatory variables. The outcome of binary logistic regression is the probability of an occurrence (Pr. (y=1)) or non-occurrence (Pr. (y=0)), this is. The maximum likelihood estimator (MLE) is widely used to estimate the parameters of multiple logistic regression because it has good optimality properties and easy to apply. However, the MLE is extremely sensitive to outliers and leverage points [1, 2, 3]. The outliers are defined as the observations that are deviated from the others in the response variable. The observations that are deviated from the majority of data set in the explanatory variables are called leverage points. To solve this problem, several robust methods have been proposed. Pregibon [4] proposed robust formula of the logistic regression model. Unfortunately, this estimator did not down-weight influential observations successfully and it was not consistent. Johnson [5] suggested to identify influential observations to estimate parameters of logistic regression model. Künsch et al. [6] studied robust estimation in generalized linear regression model and logistic regression. They proposed to use conditionally unbiased bounded influence (CUBIF) method and they showed that the optimal estimator does not depend on the distribution of the explanatory variables.

Carroll and Pederson [7] suggested a method that down-weight the basis of leverage points and outliers. This estimator is known as the Mallows-class.

Bianco and Yohai [1] suggested class of robust and Fisher- consistent M estimates to estimate parameters of logistic regression model. The estimator is known as (BY) estimator, it was shown to be consistent, asymptotically normal and has good bias. However, this estimator may be effected by existing some leverage points in the covariate space. Therefore, Croux and Haesbroeck [8] suggested to add weight to (BY) estimator to down-weight the high leverage point. The weight is computed according to the value of robust Mahalanobis distance. This method is named (WBY) estimator.

Rousseeuw and Christmann [9] proposed to apply weighted maximum estimated likelihood estimator (WEMEL) to down-weight the leverage points, it is defined as

$$\sum_{i=1}^{n}[y_i - F(x_i'\beta)]w_i x_i = 0$$

*Corresponding author e-mail: ehab.mahmood@uobabylon.edu.iq

$$w_i = \frac{M}{\max[RMD^2(x_i'), M]}$$

where, RMD is robust Mahalanobis distance and M is the 75th percentile of RMD^2.

Plan and Vershynin [10] proposed a convex programming method for the spare parameters of logistic regression model. However, this is not robust outliers in covariate matrix. Feng et al. [3] proposed robust logistic regression (RoLR) to estimate the parameters when the explanatory variable follows normal distribution. They proved that RoLR is robust estimator when the data has outliers. However, this method is not considered that the explanatory variable may follow another distribution. Xu and Principe [11] proposed methods to deal with the problems of logistic regression with outliers and class imbalance. Hobza et al. [12] proposed robust method to estimate parameters of the logistic regression model based on the modified median estimator.

Ahmed and Cheng [13] suggested two techniques to estimate parameters of the logistic regression model. The robust techniques are weighted maximum likelihood estimator (WMLEw1, WMLEw2), they are considered two different weights of the Mallows class.

Kızılarslan and Camkıran [14] compared the performance of some robust estimators for the different distributions of the explanatory variables with high leverage points in the logistic regression model.

In this paper, we propose two robust methods to estimate parameters of multiple logistic regression model. The proposed methods consider the robust Mahalanobis distance (RMD) based on minimum covariance determinant (MCD) to identify leverage points.

2. Materials and Methods

2.1 Maximum Likelihood Estimator (MLE)

The multiple logistic regression model is defined as

$$Y = X\beta + e \quad (1)$$

where, Y is (n*1) vector of the response variable follows Bernoulli distribution which takes 1 for occurrence and 0 for non-occurrence, X is (n*k) matrix of explanatory variables, β is (k*1) vector of parameters model and e is (n*1) vector of random error.

A logarithm transformation of the binary logistic regression model is given as [13]:

$$\hat{y} = \log\left(\frac{P}{1-P}\right) = X\beta \quad (2)$$

$$\frac{P}{1-P} = e^{X\beta}$$

where, $p = Pr.(y = 1)$ and $1 - p = Pr.(y = 0)$

$$p = F(X\beta) = \frac{\exp(X\beta)}{1+\exp(X\beta)} \quad (3)$$

Because of that y follows Bernoulli distribution, the probability density function of y is defined as

$$g(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \qquad i = 1, 2, \ldots, n \quad (4)$$

So the joint probability density function is define as

$$l(y_1, y_2, .., y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

The MLE is defined by maximizing the algorithm of the likelihood function as

$$\log[l(y_1, y_2, .., y_n)] = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \log(1-p_i)$$

$$= \sum_{i=1}^{n} y_i(X\beta) - \sum_{i=1}^{n} \log(1+\exp(X\beta)) \quad (5)$$

By differentiating Eq. (5) with respect to (β), we can obtain MLE of β as

$$\sum_{i=1}^{n}[y_i - F(x_i'\beta)]x_i = 0 \quad (6)$$

Newton-Raphson method is applied to solve Eq. (6) [15].

2.2 Robust Estimation Methods

Robust methods have been proposed to remedy outliers and high leverage points by researchers. In this section, some of these methods have been explained.

2.2.1 Conditionally Unbiased Bounded Influence Function (CUBIF)

The CUBIF estimator depends on the asymptotic covariance matrix to abound on a measure of infinitesimal sensitivity. It is considered the following function [16] :

$$\psi_{cond}(y,x,\alpha,\beta) = d(y,x,\alpha,\beta)w_b(|d(y,x,\alpha,\beta)|(x'\beta^{-1}x)^{0.5})x$$

where,

$$d(y,x,\alpha,\beta) = y - g(x'\alpha) - c(x'\alpha, \frac{b}{(x'\beta^{-1}x)^{0.5}}) \quad (7)$$

$w_b(a) = H_b(a)/a$, where H_b is the Huber function $H_b(a) = \max(-b, \min(a,b))$.

The main difference between CUBIF results and Mallows class is that w_b in (7) factors into two parts: The first one depends on x ($(x'\beta^{-1}x)^{0.5}$) and the second one depends on $d(y,x,\beta) = y - g(x'\alpha) - c(x'\alpha, \frac{b}{(x'\beta^{-1}x)^{0.5}})$

Because of the distribution F of the (x) is unknown, it is replaced by empirical distribution, this means that we solve

$$\sum_{i=1}^{N} \psi_{cond}(y,x,\hat{\alpha}_N,\hat{\beta}_N) = 0$$

and

$$N^{-1}\sum_{i=1}^{N} x_i x_i' v\left(x_i'\hat{\alpha}_N, \frac{b}{(x_i'\hat{\beta}_N^{-1}x_i)^{0.5}}\right) = \hat{\beta}_N$$

where,

$$v(\beta,a) = \int \left(y - g(\beta) - c(\beta,a)\right)^2 w^2(y,\beta,a).\exp(y\beta - G(\beta) - S(y))\mu(dy)$$

and

$$w(y,\beta,a) = \min(1, \frac{a}{|y-g(\beta)-c(\beta,a)|})$$

2.2.2 Mallows Class

In general, the robust estimates is defined by finding the solution of the following equation:

$$\sum_{i=1}^{n} w_i x_i [y_i - F(x_i'\beta) - c(x_i,\beta)] = 0 \quad (8)$$

where, w_i are weights. If $w_i = 1$ and $c(x_i,\beta) = 0$, the solution of equation (8) yields MLE. If $w_i = w(x_i, x_i'\beta)$ and $c(x_i, x_i'\beta) = 0$, then the solution is called Mallows class [17, 18].

The Mallows-class aims to down-weight leverage points by consider function of predictors and the parameter β that can. So, the Mallows estimators have bias smaller than MLE estimators. The method can be defined as [7]

$$M_{nj}(\beta) = n^{-1}\sum_{i=1}^{n} w_i^j x_i x_i' F^{(1)}(x_i'\beta) \quad (9)$$

$$T_{nj}(\beta) = M_{n1}^{-1}(\beta)n^{-1}\sum_{i=1}^{n} w_i^{(2-j)} x_i x_i' M_{n1}^{-1}(\beta)M_{n2}(\beta)M_{n1}^{-1}(\beta)x_i F^{(j)}(x_i'\beta)$$

where, $w_i^{(2-j)}$ is the (2-j)th derivative of w(u,m), $u = x_i$ and $m = x_i'\beta$. The consistent solutions $\hat{\beta}$ to equation (8) are asymptotically normal distribution

$$n^{1/2}(\hat{\beta} - \beta) \approx N(0, M_{n1}^{-1}(\beta)M_{n2}(\beta)M_{n1}^{-1}(\beta))$$

The covariance matrix of $\hat{\beta}$ is estimated as

$n^{-1}M_{n1}^{-1}(\hat{\beta})M_{n2}(\hat{\beta})M_{n1}^{-1}(\hat{\beta})$

2.2.3 *Bianco and Yohai (BY)*

The robust formula of the logistic regression model was proposed by Pregibon [4] as

$$\hat{\beta} = argmin \sum_{i=1}^{n} \rho[D_i(x_i'\beta, y_i)] \quad (10)$$

where, ρ is Huber's type function given as

$$\rho(t) = \begin{cases} t & if \quad t \le c \\ 2(tc)^{1/2} - c & if \quad t > c \end{cases} \quad (11)$$

However, the estimators are not consistent, bias and not robust with high leverage points. Bianco and Yohai [1] improved these estimators by minimizing it as follows

$$\hat{\beta} = argmin \sum_{i=1}^{n} \rho[D_i(x_i'\beta, y_i) + G\big(F(x_i'\beta)\big) + G\big(1 - F(x_i'\beta)\big)] \quad (12)$$

The ρ is defined as

$$\rho(t) = \begin{cases} t - (\frac{t^2}{2c}) & if \quad t \le c \\ \frac{c}{2} & if \quad t > c \end{cases} \quad (13)$$

where, c is positive number.

$$G(t) = \int_0^t \acute{\rho}(-\ln u)du$$

2.2.4 *Weighted Bianco and Yohai (WBY)*

Croux and Haesbroeck [8] complemented the (BY) estimator by providing a fast and stable algorithm. They developed the (BY) estimator by adding weight to down-weight high leverage points. They considered robust Mahalanobis distance (RMD) as measure to identify leverage points. The estimator is given as

$$\hat{\beta} = argmin \sum_{i=1}^{n} w_i \rho[D_i(x_i'\beta, y_i) + G\big(F(x_i'\beta)\big) + G\big(1 - F(x_i'\beta)\big)] \quad (14)$$

$$w_i = \begin{cases} 1 & if \quad \mathrm{RMD}_i^2 \le \chi^2 \\ 0 & otherwise \end{cases} \quad (15)$$

2.2.5 *Robust Logistic Regression (RoLR)*

The (RoLR) is considered to remove the samples with very big size and then is maximized trimmed correlation of the remained samples. The sample is trimmed for all $\| x_i \| > T$, [3]

$$T = 4\sqrt{\frac{\log(p)+\log(n)}{p}} \quad (16)$$

where, p is No. of parameters.

The MLE is applied for the trimmed sample. They showed that this estimator is robust of existing outliers and high leverage points.

2.2.6 *Proposed Robust Methods*

The robust Mahalanobis distance (RMD) that is computed from explanatory variables is useful to identify leverage points in the regression model [19, 20]. It depends on the location and scatter parameters. In the literature, many methods have been proposed to find robust estimation of multivariate location and scatter. The Minimum Covariance Determinant (MCD) method is highly robust method to estimate multivariate location and scatter [21, 22, 23]. So, in this research, the robust Mahalanobis distance based on MCD method is considered to identify leverage points in the logistic regression model.

Let X(n*p) be a matrix of explanatory variables, n = sample size and p = number of parameters. The RMD is given as

$$RMD_i = \sqrt{(x_i - \mu_{MCD})' \Sigma_{mcd}^{-1} (x_i - \mu_{MCD})} \qquad i = 1, 2, \ldots, n \qquad (17)$$

where,

μ_{MCD} (n*1) is a vector of the MCD estimates of location.

Σ_{MCD} (n*p) is a matrix of the MCD estimates of scatter.

Midi et al. [24] and Alguraibawi et al. [19] proposed the following formula to compute the cut-off point:

$$cut_{point} = Median(RMD_i) + 3 * MAD(RMD_i) \qquad (18)$$

where, *MAD* is the median absolute deviation of RMD.

The main objective of the robust estimation methods is down-weight the influence of leverage points. So, two formulas are proposed to compute the weight as

$$w1_{(i)} = \begin{cases} 1 & \text{If } RMD_i \le cut_{point} \\ 0 & \text{If } RMD_i > cut_{point} \end{cases} \qquad (19)$$

$$w2_{(i)} = \begin{cases} 1 & \text{If } RMD_i \le cut_{point} \\ \frac{cut_{point}}{RMD_i} & \text{If } RMD_i > cut_{point} \end{cases} \qquad (20)$$

After reduce the effect of the leverage points, the MLE method can be applied successfully. The proposed estimators are named MLE_{RMD1} and MLE_{RMD2}, respectively.

3. Results and Discussion

3.1 Simulation Study

A simulation study was carried out by using R-program to examine the performance of the proposed robust methods. The performance of the proposed methods (MLE_{RMD1} and MLE_{RMD2}) were compared with MLE and five robust methods (CUBIF, Mallows, BY, WBY and RoLR). Two explanatory variables were generated, x1~N (0,1) and x2~N(0,1), parameter values of β_0= 0.5, β_1=0.1 and β_2=0.2 with three different sample sizes, n = (100, 200 and 300). Because of the response variable (y) is following binomial distribution, it was generated with parameter (p) that is defined in eq. (3). Four different scenarios are then examined, uncontaminated and contaminated data with 5%, 10% and 15%. Following by Midi and Syaiba [2], both of the explanatory variables are contaminated according to the following formulas:

$$x_i = x_i + 5 \qquad (21)$$

The simulation was repeated 2000 times for the four scenarios. The performance of the proposed methods is evaluated based on the bias and MSE. The bias and MSE are given as [2]:

$$Bias = \left| \frac{\sum_{i=1}^{2000} (\hat{\beta}_i - \beta)}{2000} \right| \qquad (22)$$

$$MSE = \frac{\sum_{i=1}^{2000} (\hat{\beta}_i - \beta)^2}{1999} \qquad (23)$$

The results of estimated parameters, bias and MSE of all the methods of uncontaminated and contaminated with sample sizes 100 and 200 are exhibited in Tables 1-8, respectively. Nevertheless, the complete tables of estimated parameters, bias and MSE could not be attached for the sample size (n=300) due to space limitation. Generally, the performance of MLE and the robust methods of estimated parameters are reasonably closer for the sample sizes (100 and 200).

Table 1: Estimated Parameters, bias and MSE for sample size (n=100) and uncontaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.513	0.013	0.047	0.109	0.009	0.050	0.216	0.016	0.050	0.007
CUBIF	0.513	0.013	0.047	0.109	0.009	0.050	0.216	0.016	0.050	0.007
Mallows	0.512	0.012	0.047	0.109	0.009	0.050	0.216	0.016	0.050	0.007
BY	0.514	0.014	0.047	0.109	0.009	0.051	0.216	0.016	0.049	0.007
WBY	0.514	0.014	0.047	0.109	0.009	0.051	0.216	0.016	0.049	0.007
RoLR	0.539	0.039	0.131	0.119	0.019	0.542	0.239	0.039	0.516	0.021
MLE_{RMD1}	0.514	0.014	0.048	0.108	0.008	0.054	0.216	0.016	0.052	0.009
MLE_{RMD2}	0.513	0.013	0.047	0.108	0.008	0.052	0.211	0.011	0.056	0.007

Table 2*:* Estimated Parameters, bias and MSE for sample size (n=100) with 5% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.496	0.004	0.046	0.055	0.045	0.030	0.209	0.009	0.053	0.008
CUBIF	0.499	0.001	0.046	0.068	0.032	0.032	0.209	0.009	0.053	0.008
Mallows	0.499	0.001	0.046	0.066	0.034	0.031	0.208	0.008	0.053	0.008
BY	0.497	0.003	0.046	0.058	0.042	0.031	0.209	0.009	0.053	0.008
WBY	-	-	-	-	-	-	-	-	-	-
RoLR	0.531	0.031	0.131	0.072	0.028	0.545	0.216	0.016	0.530	0.020
MLE_{RMD1}	0.509	0.009	0.048	0.100	0.000	0.059	0.209	0.009	0.059	0.013
MLE_{RMD2}	0.510	0.010	0.048	0.106	0.006	0.041	0.217	0.017	0.058	0.008

Table 3: Estimated Parameters, bias and MSE for sample size (n=100) with 10% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.472	0.028	0.048	0.060	0.040	0.027	0.115	0.085	0.037	0.009
CUBIF	0.481	0.019	0.049	0.071	0.029	0.029	0.139	0.061	0.037	0.009
Mallows	0.480	0.020	0.048	0.069	0.031	0.028	0.136	0.064	0.034	0.009
BY	0.475	0.025	0.049	0.063	0.037	0.029	0.124	0.076	0.039	0.009
WBY	-	-	-	-	-	-	-	-	-	-
RoLR	0.541	0.041	0.151	0.113	0.013	0.646	0.241	0.041	0.597	0.022
MLE_{RMD1}	0.511	0.011	0.053	0.110	0.010	0.056	0.213	0.013	0.056	0.013
MLE_{RMD2}	0.477	0.023	0.045	0.065	0.035	0.029	0.117	0.083	0.035	0.008

Table 4: Estimated Parameters, bias and MSE for sample size (n=100) with 15% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.480	0.020	0.053	0.039	0.061	0.021	0.107	0.093	0.037	0.009
CUBIF	0.487	0.013	0.053	0.045	0.055	0.022	0.126	0.074	0.037	0.009
Mallows	0.487	0.013	0.052	0.045	0.055	0.021	0.124	0.076	0.035	0.009
BY	0.483	0.017	0.053	0.040	0.060	0.022	0.115	0.085	0.039	0.009
WBY	-	-	-	-	-	-	-	-	-	-
RoLR	0.557	0.057	0.160	0.133	0.033	0.650	0.216	0.016	0.644	0.023
MLE_{RMD1}	0.527	0.027	0.060	0.105	0.005	0.060	0.210	0.010	0.061	0.012
MLE_{RMD2}	0.500	0.000	0.047	0.094	0.006	0.059	0.205	0.005	0.061	0.008

Table 5: Estimated Parameters, bias and MSE for sample size (n=200) and uncontaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.511	0.011	0.023	0.105	0.005	0.024	0.204	0.004	0.023	0.004
CUBIF	0.511	0.011	0.023	0.105	0.005	0.024	0.204	0.004	0.023	0.004
Mallows	0.511	0.011	0.023	0.105	0.005	0.024	0.203	0.003	0.023	0.004
BY	0.511	0.011	0.023	0.105	0.005	0.024	0.204	0.004	0.023	0.004
WBY	0.511	0.011	0.023	0.105	0.005	0.024	0.204	0.004	0.023	0.004
RoLR	0.532	0.032	0.099	0.136	0.036	0.622	0.216	0.016	0.593	0.016
MLE_{RMD1}	0.511	0.011	0.023	0.106	0.006	0.025	0.204	0.004	0.024	0.006
MLE_{RMD2}	0.511	0.011	0.022	0.105	0.005	0.024	0.209	0.009	0.024	0.004

Table 6: Estimated Parameters, bias and MSE for sample size (n=200) with 5% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.485	0.015	0.022	0.068	0.032	0.017	0.140	0.060	0.022	0.004
CUBIF	0.494	0.006	0.022	0.080	0.020	0.018	0.164	0.036	0.020	0.005
Mallows	0.493	0.007	0.022	0.078	0.022	0.017	0.160	0.040	0.020	0.004
BY	0.487	0.013	0.022	0.070	0.030	0.018	0.148	0.052	0.023	0.005
WBY	0.486	0.014	0.022	0.070	0.030	0.018	0.148	0.052	0.023	0.005
RoLR	0.534	0.034	0.106	0.091	0.009	0.639	0.205	0.005	0.662	0.017
MLE_{RMD1}	0.512	0.012	0.024	0.106	0.006	0.025	0.210	0.010	0.025	0.009
MLE_{RMD2}	0.502	0.002	0.021	0.104	0.004	0.025	0.205	0.005	0.024	0.004

Table 7: Estimated Parameters, bias and MSE for sample size (n=200) with 10% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.473	0.023	0.024	0.052	0.048	0.014	0.097	0.103	0.023	0.005
CUBIF	0.481	0.019	0.024	0.062	0.038	0.015	0.122	0.078	0.021	0.005
Mallows	0.481	0.019	0.024	0.061	0.039	0.014	0.118	0.082	0.020	0.005
BY	0.474	0.026	0.024	0.054	0.046	0.015	0.104	0.096	0.024	0.005
WBY	-	-	-	-	-	-	-	-	-	-
RoLR	0.535	0.035	0.112	0.137	0.037	0.712	0.211	0.011	0.691	0.018
MLE_{RMD1}	0.512	0.012	0.025	0.104	0.004	0.026	0.202	0.002	0.026	0.008
MLE_{RMD2}	0.500	0.000	0.022	0.099	0.001	0.025	0.204	0.004	0.025	0.004

Table 8: Estimated Parameters, bias and MSE for sample size (n=200) with 15% contaminated data

Method	β_0			β_1			β_2			MSE
	Est.	bias	MSE	Est.	bias	MSE	Est.	bias	MSE	
MLE	0.466	0.034	0.027	0.041	0.059	0.013	0.077	0.123	0.025	0.005
CUBIF	0.472	0.028	0.027	0.047	0.053	0.013	0.094	0.106	0.024	0.006
Mallows	0.473	0.027	0.027	0.048	0.052	0.013	0.093	0.107	0.022	0.005
BY	0.466	0.034	0.027	0.042	0.058	0.013	0.083	0.117	0.026	0.006
WBY	0.464	0.036	0.028	0.043	0.057	0.014	0.083	0.117	0.026	0.006
RoLR	0.552	0.052	0.122	0.098	0.002	0.768	0.222	0.022	0.775	0.019
MLE_{RMD1}	0.511	0.011	0.029	0.095	0.005	0.029	0.200	0.000	0.029	0.009
MLE_{RMD2}	0.492	0.002	0.022	0.094	0.004	0.028	0.193	0.007	0.026	0.004

Generally, it can be observed from the Table 1 that there is not much difference of estimated parameters, biases and

MSEs among all the methods for uncontaminated data for sample size (n=100). The biases and MSEs of the methods are close to each other to estimate parameters of the model. However, the estimated parameters, bias and MSE of RoLR and MLE are more affected by leverage points especially with high percentages of contamination. Besides, unidentifiable parameter estimates in some cases of WBY method because of there is not existence of overlapping, they are shown in Tables 2-4 for sample size (n=100). The results of bias of CUBIF, Mallows and BY are slightly less than the bias of MLE. The MLE_{RMD1} successfully estimated model parameters with less bias and MSE. Furthermore, the MLE_{RMD2} successfully estimated model parameters with the least values of bias and MSE at different percentages of contamination.

Similar results were obtained for the sample size (n=200), these results are given in Tables (5-8) for uncontaminated data and (5%, 10% and 15%) contaminated data, respectively.

In summary, the proposed methods MLE_{RMD1} and MLE_{RMD2} provide the better performance results than the others to estimate model parameters with different percentage of contamination. Besides, the MLE_{RMD2} method can be estimate the parameters with the least values of bias and MSE.

3.2 Numerical Example

A real data set of measurements of (30) leukemia patients is considered to represent the logistic regression model. The response variable is represented by (1) if the patient surviving at least 52 weeks and (0) otherwise. The model includes two explanatory variables: white blood cell count (WBC) and AG status (one for positive patient and zero for negative patient). Cook and Weisberg [25] identified observation number (15) as influential. Table 9 presents the estimated parameters and MSE of the model. It is noticed the robust methods give values of MSE smaller than MLE except RoLR method (without estimated values). However, the proposed method MLE_{RMD2} successfully estimated model parameters with the smallest value of MSE.

Table 9: Estimated parameters and MSE of leukemia data

Method	Est.			MSE
	β_0	β_1	β_2	
MLE	1.427	0.000	2.286	0.481
CUBIF	0.733	0.000	2.238	0.370
Mallows	0.187	0.000	2.534	0.272
BY	1.385	0.000	2.218	0.333
WBY	0.159	0.000	1.927	0.296
RoLR	-	-	-	-
MLE_{RMD1}	0.170	0.000	2.531	0.290
MLE_{RMD2}	0.194	0.000	3.217	0.259

For more illustration, the results of MSE are shown in Figure 1.

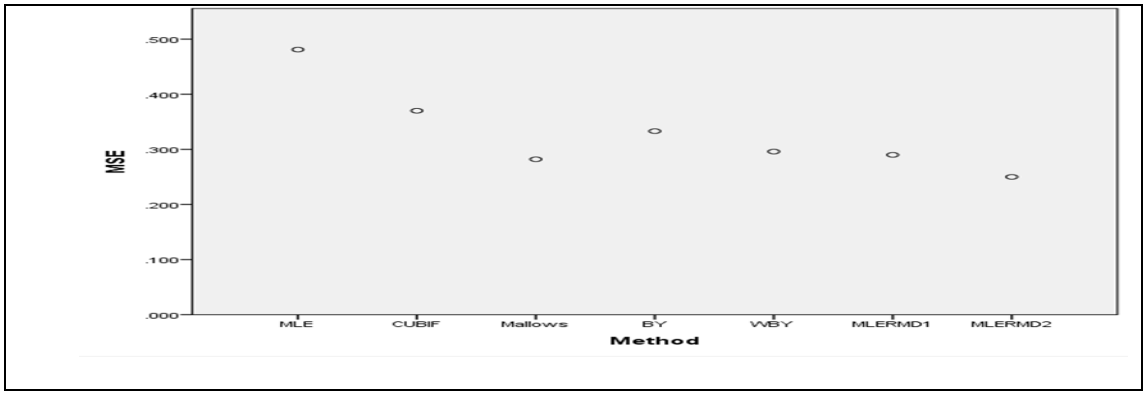

Fig.1. MSE of the estimated methods

4. Conclusions

The problem of existing leverage points in the multiple logistic regression model is considered. We proposed two robust methods namely MLE_{RMD1} and MLE_{RMD2} to estimate the parameters of the model. A comparison with MLE and five robust methods was made by depending on bias and MSE as measures for comparison. According to the simulation and example results, the MLE method has the largest values of bias and MSE with existing of leverage points. The robust methods were less affected by leverage points except WBY method. However, the proposed robust methods are very successful to estimate model parameters with different percentage of contamination. The MLE_{RMD2} gives the best result to estimate parameters of the multiple logistic regression model when the explanatory variables have some leverage points.

References

[1] A. M. Bianco and V. J. Yohai , *Robust estimation in the logistic regression Model*, in Robust Statistics, Data Analysis, and Computer Intensive Methods. Lecture Notes in Statistics, vol. 109. H. Rieder, Springer, New York, 17-34, (1996).

[2] H. Midi and B. A. Syaiba, The performance of classical and robust logistic regression estimators in the presence of outliers, *Editorial Board*, 313, (2012).

[3] J. Feng, H. Xu, S, Mannor, and S. Yan, *Robust logistic regression and classification*, in Advances in Neural Information Processing Systems, vol. 27, M. I. Jordan, Y. LeCun and S. A. Solla, MIT Press, USA, 253-261, (2014).

[4] D. Pregibon, Resistant fits for some commonly used logistic models with medical applications, *Biometrics*, 485-498, (1982).

[5] W. Johnson, Influence measures for logistic regression: another point of view, *Biometrika*, 72(1), 59-65, (1985).

[6] H. R. Künsch, L. A. Stefanski and R. J. Carroll, Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models, *Journal of the American Statistical Association*, 84(406), 460-466, (1989).

[7] R. J. Carroll and S. Pederson, On robustness in the logistic regression model, *Journal of the Royal Statistical Society: Series: B*, 55(3), 693-706, (1993).

[8] C. Croux and G. Haesbroeck, Implementing The Bianco and Yohai estimator for logistic regression, *Computational Statistics and Data Analysis*, 44(1-2), 273-295, (2003).

[9] P, J. Rousseeuw and A. Christmann, Robustness against separation and outliers in logistic regression, *Computational Statistics and Data Analysis*, 43(3), 315-332, (2003).

[10] Y. Plan and R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach, *IEEE Transactions on Information Theory*, 59(1), 482-494, (2012).

[11] G. Xu, B. G. Hu and J. C. Principe, *Robust bounded logistic regression in the class imbalance problem*, In 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, 1434-1441, (2016).

[12] T. Hobza, N. Martín and L. Pardo, A Wald-type test statistic based on robust modified median estimator in logistic regression models, *Journal of Statistical Computation and Simulation*, 87(12), 2309-2333, (2017.

[13] I. A. I. Ahmed and W. Cheng, The performance of robust methods in logistic regression model, *Open Journal of Statistics*, 10, 127-138, (2020).

[14] Ş. Kızılarslan and C. Camkıran, Comparison of robust logistic regression estimators for variables with generalized extreme value distributions, *Model Assisted Statistics and Applications*, 16(3), 177-187, (2021).

[15] S. Ahmad, N. M. Ramli and H. Midi, Robust estimators in logistic regression: a comparative simulation study, *Journal of Modern Applied Statistical Methods*, 9(2), 502-511, (2010).

[16] H. R. Künsch, L. A. Stefanski and R. J. Carroll, Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models, *Journal of the American Statistical Association*, 84(406), 460-466, (1989).

[17] C. L. Mallows, *On Some Topics in Robustness: Technical Memorandum*, Murray Hill Press, New Jersey USA, 140-141, (1975).

[18] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York USA, 153-159, (1986).

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

[20] H. A. Lim and H. Midi, Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) for the identification of high leverage points in linear model, *Computational Statistics*, 31(2), (2016).

[21] P. J. Rousseeuw, Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications*, 37(8), 283-297, (1985).

[22] P. J. Rousseeuw, and K. V. Driessen, A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223, (1999).

[23] M. Hubert, M. Debruyne and P. J. Rousseeuw, Minimum covariance determinant and extensions, *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3), (2018).

[24] H. Midi, M. R. Norazan and A. H. M. Imon, The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression, *Journal of Applied Statistics*, 36(5), 507-520, (2009).

[25] R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall Press, New York, (1982).