Journal of New Theory
ISSN: 2149-1402

36 (2021) 49-63
Journal of New Theory
https://dergipark.org.tr/en/pub/jnt
Open Access

Robust Logistic Modelling for Datasets with Unusual Points

Kumru Urgancı Tekin[1], Burcu Mestav[2], Neslihan İyit[3]

Article History
Received: 13 Jul 2021
Accepted: 22 Sep 2021
Published: 30 Sep 2021
10.53570/jnt.971062
Research Article

Abstract – Unusual Points (UPs) occur for different reasons, such as an observational error or the presence of a phenomenon with unknown cause. Influential Points (IPs), one of the UPs, have a negative effect on parameter estimation in the Logistic Regression model. Many researchers in fisheries sciences face this problem and have recourse to some manipulations to overcome this problem. The limitations of these manipulations have prompted researchers to use more suitable and innovative estimation techniques to deal with the problem. In this study, we examine the classification accuracies and parameter estimation performances of the Maximum Likelihood (ML) estimator and robust estimators through modified real datasets and simulation experiments. Besides, we discuss the potential applicability of the assessed robust estimators to the estimation models when the IPs are kept in the dataset. The obtained results show that the Weighted Maximum Likelihood (WML) and Weighted Bianco-Yohai (WBY) estimators of robust estimators outperform the others.

Keywords *– Influential point, robust estimators, unusual point, logistic regression*

Mathematics Subject Classification (2020) – 62G32, 65C60

1. Introduction

The most frequently adopted statistical method to obtain parameter estimates of the explanatory variables relationship with the binary outcome (0 and 1) is Logistic Regression. Binary Logistic Regression (BLR) models the functional relationship between the binary response variable and one/more explanatory variable [1-4]. Maximum Likelihood Estimator (MLE), which has the optimal properties under proper circumstances, is utilized to estimate the parameters in BLR; however, it is considerably affected by the presence of an unusual data point(s) in the dataset and may cause misleading inferences and misinterpretations in parameter estimates [5-11].

The unusual data point(s) (UP(s)) is generally defined as point(s) that are relatively far from the central tendency compared to all values [12-13]. These types of point(s) may derive from errors existing during the recording of observations, sampling errors, and experimental errors or may originate from an unknown phenomenon in a study area (e.g., economy, applied science, health, engineering).

The UP(s) are differently named as an outlier(s), influential point(s), or leverage point(s) according to their locations in the two-dimensional space. Among these definitions, influential point(s) (IPs) can be described as the product of dangerous outliers and bad leverage points and significantly affect the fit of the

[1]kumruurganci@comu.edu.tr (Corresponding Author); [2]burcumestav@comu.edu.tr; [3]niyit@selcuk.edu.tr
[1,2]Department of Statistics, Faculty of Arts and Sciences, Çanakkale Onsekiz Mart University, Çanakkale, Turkey
[3]Department of Statistics, Faculty of Sciences, Selcuk University, Konya, Turkey

model or the estimation of the parameters compared to the others [14-17]. If the variable on the x-axis is continuous and the one on the y-axis is binary, unusual points can only occur as a transposition $0 \rightarrow 1$ or $1 \rightarrow 0$ in the y-axis direction [6]. This type of UP(s) is also recognized as a residual outlier or misclassification-type error [16].

These point(s) can be observed in datasets of numerous studies conducted in applied areas, and most researchers have been confused about what to do with them and how to manage them. To manage IP(s), researchers generally have had to decide among such strategies as keeping them, removing them, or recoding them [12]. [18] reported as a result of their research on the frequency of these points in different scientific disciplines that there is no overarching explanation and the frequency varies according to the study area and sample size; and they have claimed that if these outliers occur in about 1-10% of the dataset, it is normal. Although this decision differs according to the scientific field studied, it is recommended to estimate parameters using a more robust estimator instead of MLE by [18] if these points are to be kept in the data. Robust estimator instead of MLE has become the focus of many research fields in statistics [19].

In the BLR model, [20] is the first to indicate the problem of parameter estimation in the presence of IPs. After that, several robust alternative parameter estimation methods much less influenced by these points are suggested in the literature (i.e., [5,6,18,21-31]). Besides, many researchers have also studied to compare the performances of the estimators to examine the robustness of these proposed estimators on simulation experiments [11,16,27,32,33]. These studies have shown that MLE can be influenced even by the presence of 1% IPs in the dataset, and therefore robust estimators were recommended [34]. However, there are very few studies examining these points in terms of their effects on parameter estimates as they move away from the centre. Our hypothesis in the present study is to display that IP(s) occurs in different levels of percentage amounts in the dataset, three standard deviations away from the centre influence parameter estimations and to illustrate to researchers the possibility of being anomaly as research questions. The present paper was built on two purposes within the framework of our hypothesis: (a) to examine the performance of MLE and some robust estimators in parameter estimation in extreme situations, such as different sample size data have different percentages of influential points (i.e., contamination rate) in simulation experiments, and to contribute to the literature by providing information on what kind of results researchers may obtain if they encounter such data points.

2. Material

In this study, we carried out comprehensive simulation experiments to examine commonly cited or recently proposed robust estimators for BLR.

In the simulation study, to examine the performance of the estimators in different situations, we generated specific datasets created in combinations that vary according to different percentages of IPs occurring farther from the centre of the dataset in three different sample sizes (100, 250, and 500). We generated datasets with IPs, which we called a contaminated dataset, by adding IPs that fall 1.5, 3, and 5 whiskers away from the centre of the dataset that constituted 1%, 5%, 10%, and 15% of the dataset in each sample size and IP-free datasets (0% contaminated), which we called a clean dataset in each sample size for control purposes. The simulated datasets contain a response and two explanatory variables. We first generated a design matrix of explanatory variables of size $n \times p$ by drawing each observation from a bivariate normal distribution $\left(x_i \sim N(\mu, \Sigma)\right)$. Where μ is a mean vector of length $p = 2$ and Σ is a 2×2 non-singular covariance matrix. The considered true values of the BLR model parameters are set to be $\beta = (\beta_0, \beta_1, \beta_2)' = (0, 2, 2)'$. Then, we produced the binary response variable according to the BLR model as follows:

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

where the error terms were generated according to a logistic distribution, $\varepsilon_i \sim logistic(0,1)$. We added the contaminants (IPs) to the dataset by inflating the covariance matrix and deriving it in the % amounts denoted in the simulation scenario. In the simulation study, we obtained the design matrix of the contaminated and uncontaminated explanatory variables with the configuration denoted below by [35].

$$(1-\gamma)N_{ss}(\mu,\Sigma)+\gamma N_{ss}\,(\mu,k\times\Sigma)$$

where N_{ss} is the sample size (100, 250 and 500), γ represents the percentage of contaminants (γ = 1%, 5%, 10%, and 15% contamination rate) in a dataset, and k represents a scalar which determines the separation of the contaminants from the rest of the data (k = 1.5, 3, and 5 whiskers), for any amount of contamination.

The aforesaid processes were applied to all the estimators used in this study. Each simulation study was replicated 1000 times by using the Monte Carlo simulation.

3. Method

The logistic regression model is a special case of GLMs, especially for a binary response variable y_i, with the assigned values 1 (success) and 0 (failure). The explanatory variables ($x_i \in R, i = 1,2,\dots,n$) and the probability of response variable $p(Y_i = 1|X_i = x_i)$ are linked to explanatory variables by the mean of a link function $g(\pi) = X\beta$, such that $g^{-1}(X\beta)$ is the logit link function, which transforms the covariate values in the internal (0,1). The BLR model can be defined by:

$$p(Y_i = 1|X_i = x_i) = F(x_i'\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\,, \quad i = 1,2,\dots,n \tag{2}$$

where $X = (1, x_1, \dots, x_p)$ is an $n \times k$ matrix of explanatory variables with $k = p + 1$ and $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ is the vector of the unknown regression coefficient. The BLR model can be defined by:

$$\eta_i = x_i'\beta \tag{3}$$

where η_i is a linear predictor known as transformation function and $\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$. Suppose that the response variable y_i has Bernoulli distribution and the joint probability density function for the i^{th} observation is,

$$f(y_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}, \;\; i = 1,2,\dots,n \tag{4}$$

and each y_i observation takes the value 1 or 0. The likelihood function is given by:

$$l(\beta; y_i) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{5}$$

Then, we take a logarithm of the likelihood function (log-likelihood), which can be written as:

$$\begin{aligned} l(\beta; y_i) &= ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} l\,(y_i, \beta) \\ &= \sum_{i=1}^{n}\left[y_i ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)\right] + \sum_{i=1}^{n} ln\left(1-\pi(x_i)\right) \end{aligned} \tag{6}$$

To estimate the parameters in BLR, Maximum Likelihood Estimator (MLE) is used. The likelihood function is produced by maximizing the logarithm of and is defined as:

$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} l\,(y_i, \beta) \tag{7}$$

As an alternative, MLE deviation statistics are minimized according to β [16], and it is defined as:

$$d_i = \left[-y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) - (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$
$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} d_i \tag{8}$$

It is known that MLE is the most efficient estimator, but it may behave very inadequately in the presence of outlying observations in terms of their location and impact. Many robust estimators have been proposed in the literature to replace MLE in order to solve this problem, but in this study, we aspired to evaluate the performances of the most cited and most recommended estimators. These robust estimators are briefly discussed in the subsequent sections.

3.1. The Mallows Type Leverage Dependent Weights Estimator (MALLOWS)

MALLOWS type estimator, introduced by [22] and intensively examined by [26], was obtained by minimizing log-likelihood function using weights dependent on explanatory variables. A robust estimate of β can be obtained by the solution of the following function [23]:

$$\sum_{i=1}^{n} w_i\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \tag{9}$$

where $w_i = W\big(h_n(x_i)\big)$ are the weights function. W is bounded by depending on $W(u)$ and a non-increasing function. $W(u)$ is dependent on a parameter $c > 0$, and $W(u) = \left(1 - \frac{u^2}{c^2}\right)^3 I(|u| \le c)$. If $w_i \equiv 1$ and $c(x_i, \beta) \equiv 0$, then Eq. (8) supplies the usual BLR model parameter estimate. If $w_i = w(x_i, x_i'\beta)$, $c(x_i, x_i'\beta) \equiv 0$, and the weights depend only on the design, this estimate is called Weighted Maximum Likelihood (MALLOWS type estimator).

3.2. Weighted Maximum Likelihood Estimator (WMLE)

This estimator is obtained in a similar way to the strategy used in constructing the MALLOWS type estimator. That is, it detects unusual values and makes the parameter estimation by equalizing the weights of these values to zero. WMLEs for BLR can be obtained with a solution in (Eq. 8). However, in this study, parameters were estimated by equalizing the weights obtained by the weighting function introduced by [36] and proposed by [33]. First, the square of the Mahalanobis distances of the explanatory variables is calculated according to the computed $\hat{\mu}^{(0)}$ and $\hat{\Sigma}^{(0)}$ values. The square of the Mahalanobis distances (m^2) is calculated by:

$$m^2 = \left(x_i - \hat{\mu}^{(0)}\right)' \left(\hat{\Sigma}^{(0)}\right)^{-1} \left(x_i - \hat{\mu}^{(0)}\right)$$

The weight function proposed by [33] is defined as:

$$w_i = (0.8 * m^2 + 0.2)$$

Then, WMLEs for BLR can be obtained by the solution of the following:

$$\sum_{i=1}^{n} w_i\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \tag{10}$$

3.3. The Conditionally Unbiased Bounded Influence Function (CUBIF) Estimator

In CUBIF estimator, introduced by [22], the weights depend on the response variables besides the explanatory variables. This method minimises a measure of efficiency based on the asymptotic (co)variance matrix to bound the measures of infinitesimal sensitivity. The M-estimators are the solution of the form of $\sum \psi(y_i, x_i, \beta) = 0$, where ψ is a known function. Its optimal function is written by:

$$\psi(y, x, \beta, B) = W(\beta, y, x, b, B)\left[y - g(\beta' x) - c\left(\beta' x, \frac{b}{h(x, B)}\right)x\right] \tag{11}$$

where B is a (co)variance matrix, b is bounded infinitesimal sensitively and $h(x, B) = (x' B^{-1} x)^{1/2}$ is a leverage measure. $c\left(\beta' x, \frac{b}{h(x,B)}\right)$ is a bias correction with corrected residual

$$\left(r(y, x, \beta, b, B) = y_i - g(\beta' x) - c\left(\beta' x, \frac{b}{h(x, B)}\right)\right)$$

The weight function $W(\beta, y, x, b, B) = W_b r\big((y, x, \beta)h(x, B)\big)$ downweighs observations with high leverage points and largely corrected residuals making M-estimator have bonded influence.

3.4. Consistent Misclassification Estimator (CME)

It is a known fact that unusual points in the dataset cause misclassification, and this issue has been studied by many researchers under different assumptions [37]. Misclassification is a stand-alone issue, and there are estimators developed for parameter estimation in case of misclassification. In this study, we used the Consistent Misclassification estimator (CME), proposed by [6], since we consider the parameter estimation in contaminated datasets. If $P(Y = 1|x_i) = F(x_i'|\beta_L)$ considered robust estimation in the BLR model, a misclassification model in which each response is misclassified with probability γ, so that [23]:

$$P(Y = 1|x_i) = F\big(x_i{}'\beta_{Mc}\big) + \gamma\{1 - 2F\big(x_i{}'\beta_{Mc}\big)\} = G\big(x_i{}'\beta_{Mc}, \gamma\big) \tag{12}$$

where β_L is the true regression parameter for the conventional BLR model and β_{Mc} is the true regression parameter under the misclassification model. [6] has investigated small values of γ and the use of (Eq. 11) in generating robust estimators and diagnostics and suggested a bias-corrected version that is suitable for small γ.

3.5. Robust Quasi-Likelihood Estimator (RQL)

The quasi-likelihood estimator, proposed by [38], is defined as solutions of the following equation:

$$\sum_{i=1}^{n} \frac{y_i - \mu(\beta' x_i)}{V(\beta' x_i)} \mu'^{(\beta' x_i)} x_i = 0$$

Then, the quasi-likelihood approach to parameter estimation was robustified by [26] by bounding and centring the quasi-likelihood score function [39].

$$\psi(y, \beta) = \frac{y_i - \mu(\beta' x)}{V(\beta' x)} \mu'(\beta' x) x,$$

To deal with high leverage points, they suggest putting weight on each point [39].

3.6. Bianco Yohai Estimator (BYE) and Weighted Bianco Yohai Estimator (WBYE)

[25] have found that Pregibon's estimator based on deviation statistics (Eq. 7) does not reduce the weight of high leverage points and is inconsistent. They have improved the consistent and more robust Bianco and Yohai Estimator (BYE) by shrinking Pregibon's estimator as follows:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\rho\left(\left[d(x_i'\beta; y_i) + G\left(F(x_i'\beta)\right) + G\left(1 - F(x_i'\beta)\right)\right]\right)\right] \tag{13}$$

where $\rho(x) = \left(x - x^2/(2c)\right) I_{(-\infty,c)}(x) + (c/2) I_{(c,\infty)}(x)$ is Huber's loss function and c is a tuning parameter,

$$G(x) = \int_0^x \rho'\left(-ln(u)\right) du$$

and I_A stands for the usual indicator function. $G\left(F(x_i'\beta)\right) + G\left(1 - F(x_i'\beta)\right)$ is a bias correction term [40].

[25] have also stressed that other choices of the bounded function ρ are possible. To reduce the effect of unusual points in the covariate space, [27] have proposed to include an extra weight to downweigh the high leverage points in (Eq. 10). Weighted Bianco and Yohai (WBY) estimator can be defined as follows [27-40]:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w(x_i) \left[\rho\left(\left[d(x_i'\beta; y_i) + G\left(F(x_i'\beta)\right) + G\left(1 - F(x_i'\beta)\right)\right]\right)\right] \tag{14}$$

where

$$w(x_i) = \begin{cases} 1 & if\ (RMD_i)^2 \le \chi^2_{m,0.975} \\ 0 & otherwise \end{cases}$$

are the weights for a decreasing function of Robust Mahalanobis Distances, and distances are computed by using the Minimum Covariance Determinant (MCD) estimator [41].

WBYE remains consistent because the weighting is merely applied to the explanatory variables. Unfortunately, the above weighting procedure also decreases the weights of the good leverage points, which is not required, and can lead to a loss of efficiency [11-16].

To test the performance of the estimators, we conducted computational experiments on Monte Carlo simulation and modified real datasets. The evaluations focused on the magnitude and severity of the IPs and the number of observations by adding outliers to the uncontaminated data. In the study, we used R 3.0.2. [42-44] to set up the Monte Carlo simulation and to examine the performance of the estimators via BLR analysis procedure.

The performances of the estimators are evaluated in view of each predicted beta parameter based on the bias and MSE (mean-squared errors):

$$\text{Bias} = \left\| \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i - \beta_i \right\|,$$

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left\| \hat{\beta}_i - \beta_i \right\|^2$$

where $\|.\|$ indicates the Euclidean norm.

4. Results

The values for the bias and the MSE of the MLE, and the seven robust estimators are given in this section in Table 1-4 for the simulation study. A "good estimator" is one that has the values of the bias, and MSE is relatively small or close to zero. The bias and MSE of the eight estimators are shown in Table 1. In the uncontaminated dataset, it can be seen that the biases and MSEs of all the estimators are considerably close to each other and also will reduce when the number of observations is increased.

Table 1. Bias, variance, and MSE values of ML, WMLE, and robust estimators for uncontaminated dataset

Sample Size	Output	MLE	WMLE	CUBIF	CME	MALLOWS	RQL	BYE	WBYE
$n = 100$	Bias	0.256	0.261	0.255	0.294	0.253	0.280	0.286	0.292
	MSE	0.746	0.772	0.747	0.960	0.744	0.839	0.849	0.882
$n = 250$	Bias	0.107	0.106	0.106	0.110	0.105	0.130	0.112	0.111
	MSE	0.224	0.230	0.224	0.234	0.224	0.274	0.240	0.246
$n = 500$	Bias	0.038	0.037	0.038	0.038	0.037	0.046	0.039	0.038
	MSE	0.099	0.102	0.100	0.103	0.099	0.121	0.107	0.110

MLE: Maximum Likelihood Estimator, **WMLE:** Weighted Maximum likelihood estimator, **CUBIF:** The Conditionally Unbiased Bounded Influence Function, **CME**: Consistent Misclassification Estimator, **MALLOWS:** The Mallows Type Leverage Dependent Weights Estimator**, RQL**: Robust Quasi-Likelihood Estimator, **BYE**: Bianco Yohai Estimator, **WBYE**: Weighted Bianco Yohai Estimator, **MSE**: Mean square error

In Table 2-4, the Bias and MSE outputs of the simulation derived from examining the estimator's behaviour under different conditions are given. As seen in the tables, the MLE method was quickly affected by the 1% degradation rate (percentage of IPs) that occurred, and outputs are the same in other studies. The presence of moderate and extreme IPs (5%, 10%, 15%) changes the results dramatically. Whereas the WMLE performs best in terms of Bias and MSE as the percentage of IP (degradation rate) increases, MLE appears to behave very poorly. The closest values to WMLE in terms of MSE and bias were observed in WBY and MALLOWS, respectively. The weighting process in the WML and WBY estimators becomes more advantageous in extreme contamination. It can be observed that the CUBIF, CME and RQL estimators do not perform well even at 5% contamination. The robustness performance of MLE dramatically decreases as the rate of contamination increases as IPs move away from the centre. At a distance of 3 whiskers, WMLE, WBY, and MALLOWS show the best performance at medium and high fouling rates, respectively, while MALLOWS, WBY, and WMLE estimators at a distance of 5 whiskers have the best performance, respectively, in terms of biases and MSEs. Meanwhile, it can be observed that bias and MSE decrease when the sample size is increased. WMLE, WBY, and Mallows have the overall best performance among all the compared estimators for different sample sizes. CUBIF, CME, and RQL estimators did not perform as well as WMLE, WBY and Mallows, even as the sample size was increased. This situation is thought to be due to the location of the unusual points. Finally, the WMLE, WBY, and Mallows estimators exhibited reasonable perform in the contaminated dataset.

Table 2. Bias, Variance, and MSE values of MLE and robust estimators over m = 1000 replication for 100 sample sizes in all cases

γ	k	Output	MLE	WMLE	CUBIF	CME	MALLOWS	RQL	BYE	WBYE
1%	1.5	Bias	1.460	0.229	0.798	0.287	1.024	0.644	0.129	0.259
		MSE	2.258	0.846	0.851	2.016	1.229	1.884	0.957	0.851
	3	Bias	2.132	0.246	0.814	0.251	0.099	0.334	0.094	0.266
		MSE	4.644	0.745	0.875	0.954	0.565	1.122	0.768	0.862
	5	Bias	2.569	0.204	0.803	0.543	0.234	0.390	0.184	0.253
		MSE	6.686	0.749	0.855	2.953	0.742	3.340	0.841	0.896
5%	1.5	Bias	2.358	0.250	1.748	2.253	2.024	0.229	1.073	0.253
		MSE	5.650	0.768	3.164	5.528	4.197	1.867	1.681	0.921
	3	Bias	2.692	0.249	1.745	2.692	0.664	2.616	2.692	0.280
		MSE	7.336	0.755	3.148	7.338	0.937	7.448	7.614	0.923
	5	Bias	2.753	0.251	1.742	2.753	0.228	2.751	2.807	0.274
		MSE	7.672	0.758	3.136	7.671	0.749	7.669	7.934	0.891
10%	1.5	Bias	2.624	0.239	2.414	2.625	2.455	2.621	2.623	0.039
		MSE	6.983	0.733	5.924	6.988	6.119	7.015	6.983	1.607
	3	Bias	2.785	0.281	2.411	2.785	1.352	2.780	2.780	0.258
		MSE	7.862	0.802	5.912	7.859	2.333	7.841	7.840	0.876
	5	Bias	2.802	0.253	2.407	2.799	0.236	2.785	2.788	0.056
		MSE	7.949	0.778	5.887	7.940	0.727	7.863	7.881	1.569
15%	1.5	Bias	2.740	0.284	2.710	2.741	2.623	2.746	2.746	0.262
		MSE	7.607	0.898	7.439	7.612	6.977	7.652	7.647	0.931
	3	Bias	2.835	0.209	2.725	2.833	1.932	2.814	2.819	0.243
		MSE	8.139	0.719	7.524	8.128	4.136	8.033	8.061	0.813
	5	Bias	2.830	0.223	2.725	2.826	0.216	2.804	2.806	0.243
		MSE	8.113	0.721	7.525	8.095	0.712	7.978	7.987	0.814

γ: contamination rate, k: whiskers distance from the centre of the data

Table 3. Bias, Variance, and MSE values of MLE and robust estimators over m = 1000 replication for 250 sample sizes in all cases

γ	k	Output	MLE	WMLE	CUBIF	CME	MALLOWS	RQL	BYE	WBYE
1%	1.5	Bias	1.071	0.081	0.549	0.088	0.720	0.101	0.142	0.073
		MSE	1.212	0.221	0.408	0.240	0.609	0.300	0.224	0.237
	3	Bias	1.643	0.084	0.558	0.068	0.077	0.089	0.022	0.084
		MSE	2.743	0.225	0.419	0.235	0.200	0.276	0.228	0.243
	5	Bias	2.127	0.082	0.551	0.086	0.078	0.106	0.055	0.054
		MSE	4.560	0.241	0.414	0.249	0.233	0.299	0.260	0.260
5%	1.5	Bias	2.267	0.101	1.604	1.920	1.898	0.096	0.934	0.088
		MSE	5.175	0.240	2.617	4.515	3.640	0.318	1.016	0.241
	3	Bias	2.658	0.085	1.609	2.658	0.635	2.221	1.435	0.086
		MSE	7.100	0.236	2.632	7.100	0.593	6.163	3.610	0.251
	5	Bias	2.742	0.087	1.609	2.742	0.086	2.742	2.741	0.076
		MSE	7.556	0.230	2.634	7.556	0.247	7.561	7.555	0.282
10%	1.5	Bias	2.603	0.104	2.335	2.604	2.431	2.602	2.598	0.101
		MSE	6.811	0.244	5.487	6.814	5.946	6.838	6.790	0.250
	3	Bias	2.777	0.087	2.334	2.777	1.422	2.775	2.773	0.099
		MSE	7.750	0.240	5.482	7.748	2.183	7.738	7.732	0.262
	5	Bias	2.799	0.093	2.333	2.797	0.078	2.783	2.783	0.107
		MSE	7.872	0.241	5.477	7.863	0.219	7.788	7.786	0.271
15%	1.5	Bias	2.725	0.077	2.700	2.725	2.607	2.731	2.732	0.088
		MSE	7.460	0.227	7.325	7.463	6.829	7.496	7.503	0.253
	3	Bias	2.826	0.090	2.719	2.825	2.033	2.808	2.813	0.104
		MSE	8.029	0.235	7.430	8.019	4.260	7.930	7.953	0.260
	5	Bias	2.826	0.067	2.721	2.823	0.069	2.801	2.802	0.083
		MSE	8.022	0.215	7.441	8.006	0.224	7.888	7.895	0.242

$\boldsymbol{\gamma}$: contamination rate, $\boldsymbol{k}$: whiskers distance from the centre of the data

Table 4. Bias, Variance, and MSE values of MLE and robust estimators over m = 1000 replication for 500 sample sizes in all cases

γ	k	Output	MLE	WMLE	CUBIF	CME	MALLOWS	RQL	BYE	WBYE
1%	1.5	Bias	1.08	0.053	0.569	0.038	0.738	0.035	0.154	0.056
		MSE	1.201	0.109	0.378	0.111	0.591	0.129	0.128	0.122
	3	Bias	1.642	0.051	0.567	0.047	0.091	0.064	0.044	0.054
		MSE	2.717	0.108	0.374	0.111	0.102	0.135	0.113	0.119
	5	Bias	2.126	0.049	0.567	0.046	0.042	0.058	0.020	0.055
		MSE	4.538	0.111	0.376	0.114	0.111	0.133	0.115	0.120
5%	1.5	Bias	2.268	0.039	1.614	2.131	1.909	0.005	0.955	0.043
		MSE	5.160	0.110	2.627	4.771	3.663	0.131	0.986	0.121
	3	Bias	2.658	0.038	1.615	2.658	0.67	2.408	0.546	0.039
		MSE	7.08	0.102	2.629	7.08	0.534	6.509	0.655	0.110
	5	Bias	2.742	0.04	1.613	2.742	0.043	2.743	2.741	0.044
		MSE	7.537	0.112	2.62	7.537	0.100	7.541	7.532	0.122
10%	1.5	Bias	2.596	0.044	2.312	2.597	2.422	2.605	2.585	0.051
		MSE	6.756	0.100	5.363	6.76	5.883	6.804	6.703	0.112
	3	Bias	2.774	0.046	2.31	2.774	0.413	2.772	2.771	0.048
		MSE	7.715	0.110	5.353	7.713	0.250	7.706	7.701	0.118
	5	Bias	2.797	0.033	2.307	2.795	0.052	2.782	2.784	0.037
		MSE	7.841	0.111	5.337	7.832	0.109	7.758	7.768	0.119
15%	1.5	Bias	2.724	0.036	2.701	2.724	2.608	2.729	2.732	0.040
		MSE	7.438	0.104	7.315	7.442	6.821	7.469	7.483	0.112
	3	Bias	2.795	0.792	2.626	2.794	0.632	2.783	2.785	0.791
		MSE	7.831	0.767	6.913	7.825	0.522	7.766	7.778	0.782
	5	Bias	2.826	0.046	2.722	2.823	0.038	2.802	2.802	0.048
		MSE	8.007	0.107	7.427	7.991	0.103	7.872	7.873	0.119

γ: contamination rate, k: whiskers distance from the centre of the data

Figure 2 shows the changes in the performance of the eight estimators concerning MSE in their respective datasets have IP(s) under different conditions. As clear from the plots, the WMLE, WBY, and MALLOWS estimators had reasonable perform in the contaminated dataset.

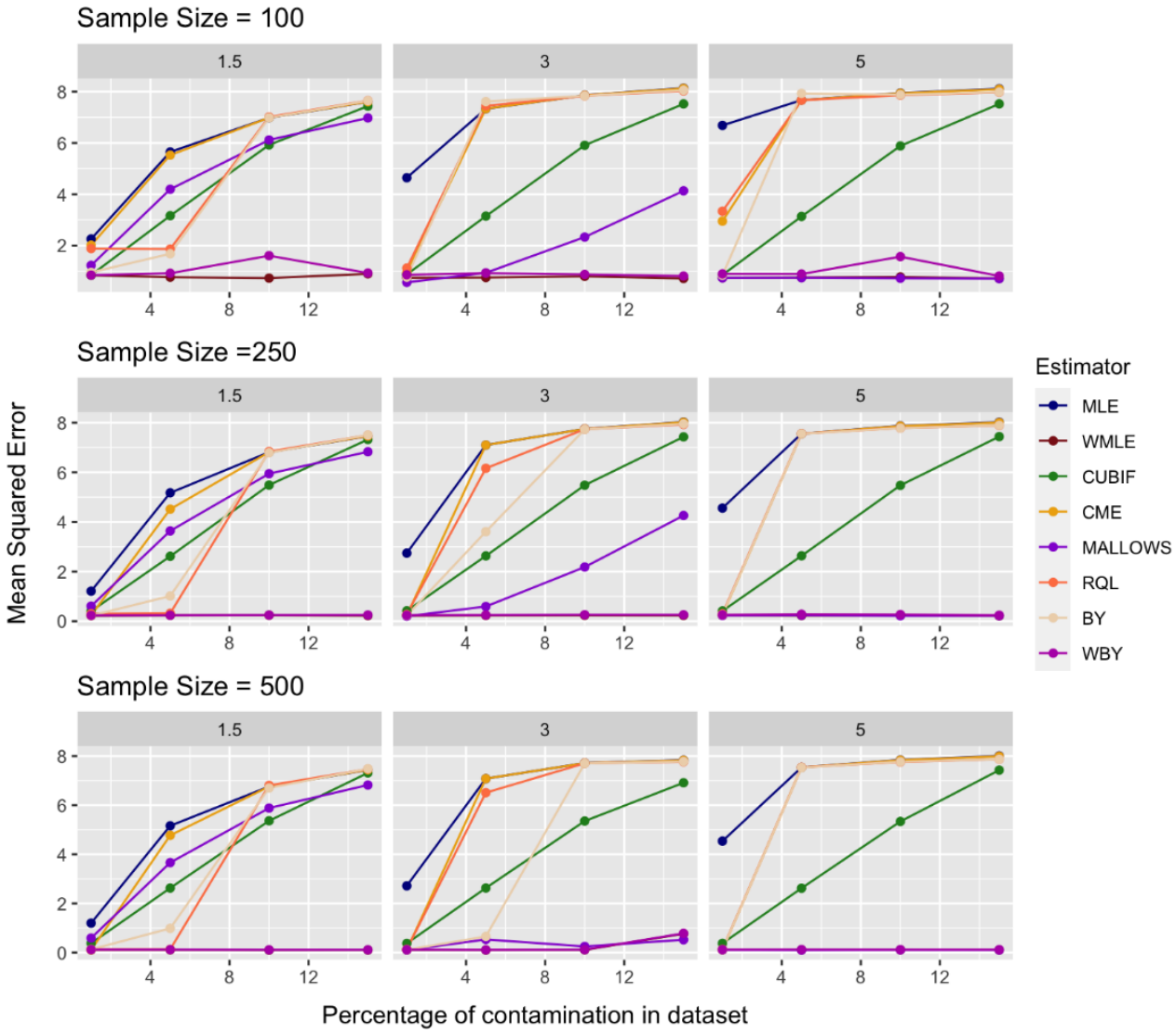

Fig. 2. Plots of change of estimator performance in terms of MSE under different conditions

MLE: Maximum Likelihood Estimator, **WMLE:** Weighted Maximum likelihood estimator, **CUBIF:** The Conditionally Unbiased Bounded Influence Function, **CME**: Consistent Misclassification Estimator, **MALLOWS:** The Mallows Type Leverage Dependent Weights Estimator**, RQL**: Robust Quasi-Likelihood Estimator, **BYE**: Bianco Yohai Estimator, **WBYE**: Weighted Bianco Yohai Estimator

5. Conclusion

Since datasets containing Influential Points (IP(s)), one of the Unusual Points (UP(s)), are possible to be encountered in every field, it becomes essential to use estimators that give more robust results than the MLE estimator to make parameter estimation. Since the size of these kinds of points is as important as their distances to the centre, the estimators designed are approaches developed with lessening the weight depending on both the location of the points and their sample size. Therefore, the simulation scenario in this study is developed considering the modelling principles of robust estimators focused on weighting.

The first of these approaches is the WML estimator, which was developed for weighting the likelihood function, and then [23] expanded the estimators by adding weights to reduce the effect of unusual points and developed new estimators. Parameter estimates (CUBIF, CME and MALLOWS) are made by using IP(s) the effect of which is reduced by this extended method, according to their position to the centre and their values.

Another approach is extended estimators (RQL), with less weight given to IP(s) and minimizing deviations from the estimated parameters. [25] developed another robust method, the Bianco and Yohai (BY) estimator, adding a function limited, differentiable, and decreasing. However, since this approach was also ineffective in reducing the weight of IP(s) with an increasing amount, [27] extended the estimator by adding a different weight to the Bianco and Yohai (BY) estimator, they defined Bianco and Yohai (WBY) estimator obtaining more consistent results.

In this study, we evaluated the MLE and seven robust estimators from the contaminated dataset with IP(s) whether it is feasible to obtain consistent parameter estimates. We conducted simulation experiments under different scenarios to examine the performance of MLE and robust estimators under contaminated and uncontaminated datasets. According to the simulation results, turned out that the uncontaminated dataset MLE and robust estimators exhibited performances similar to each other, the classical ML estimates lacked robustness and could be biased when IPs were present, while robust estimators gave better results. Among the robust estimators, the WML WBY and MALLOWS estimators, respectively, produced the smallest BIAS and MSE in the contaminated data. With the increase of the contamination at five whiskers, the MALLOWS, WBY, and WML estimators, respectively, produced the smallest bias and MSE. The results demonstrated that there might be frequent and significant differences in the case of IPs in the dataset and, therefore, should be taken into account as an example of how results can differ in different research areas. It can, thus, be concluded that the WML, WBY, and MALLOWS estimators outperformed the ML estimator and the rest of the robust estimators in the presence of IP(s).

In addition to examining the performances of robust estimators, we evaluated the problem regarding what percentage of UP(s) should be kept in the dataset. The results of our study showed that, like other studies comparing predictors, the traditional ML estimator deteriorated even at 1% contamination; for this reason, if the datasets contain approximately 1-10% or more unusual points, we recommend that they should be examined carefully. A robust method is needed, especially when there is an UP(s) at 1.5 or more whisker distance from the centre. These data should be treated from an objective perspective, and they should then be examined specifically. After being examined in detail with as different analytical methods as possible, it should be kept in the dataset, or one of the other strategies (removed or transformed) should be opted for. If the distance and amount of contamination are high, these points, determined by analytical and graphical methods, may be the possibility of anomaly, depending on the field of study. Anomaly detection methods are different from the detection of IP(s), and in such cases, these points should be considered a separate research subject without treating them as IP(s) or outliers. More studies are needed to develop and research more suitable robust methods that can be used to detect unusual points and anomalies in BLR and for parameter estimation in these types of datasets.

The first point to be considered on which estimator should be used for performance in further studies is the location of the IP(s) and the amount of the IP(s) in the dataset.

As the distance of IP(s) to the centre increases, it can be said that WML and MALLOWS estimators, in which weighting is performed according to the location to lessen the effect of the points, are better. On the other hand, the WBY estimator is a better alternative in case the number of IP(s) is high (1-10% and/or more).

Author Contributions

All the authors contributed equally to this work. They all read and approved the last version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, J. S. S. White, *Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution*, Trends in Ecology and Evolution 24 (2009) 127–135.

[2] O. Komori, S. Eguchi, S. Ikeda, H. Okamura, M. Ichinokawa, S. Nakayama, *An Asymmetric Logistic Regression Model for Ecological Data*, Methods in Ecology and Evolution 7 (2016) 249–260.

[3] F. O. Adenkule, *A Binary Logistic Regression Model for Prediction of Feed Conversion Ratio of Clarias gariepinus from Feed Composition Data*, Mar. Sci. Tech. Bull 10(2) (2021) 134–141.

[4] M. U. S. Nunes, O. R. Cardoso, M. Soeth, R. A. M. Silvano, L. F. Fa´varo, *Fishers' Ecological Knowledge on the Reproduction of Fish and Shrimp in a Subtropical Coastal Ecosystem*, Hydrobiologia 848 (2021) 929–942.

[5] D. Pregibon, *Resistant Fits for Some Commonly Used Logistic Models with Medical Applications*, Biometrics 38(2) (1982) 485–498.

[6] J. Copas, *Binary Regression Models for Contaminated Data*, Journal of the Royal Statistical Society Series B (Methodological) 50(2) (1988) 225–265.

[7] M. Pia, V. Feser, *Robust Inference with Binary Data*, Psychometrika 67(1) (2002) 21–32.

[8] A. H. M. Rahmatullah Imon, A. S. Hadi, *Identification of Multiple Outliers in Logistic Regression*, Communications in Statistics - Theory and Methods 37(11) (2008) 1697–1709.

[9] A. A. M. Nurunnabi, A. H. M. Rahmatullah Imon, M. Nasser, *Identification of Multiple Influential Observations in Logistic Regression*, Journal of Applied Statistics 37(10) (2009) 1605–1624.

[10] S. K. Sarkar, M. Habshah, S. Rana, *Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study*, Journal of Applied Sciences 11 (2011) 315–332.

[11] M. Habshah, S. B. Ariffin, *The Performance of Classical and Robust Logistic Regression Estimators in the Presence of Outliers*, Pertanika Journal of Science and Technology 20(2) (2012) 313–325.

[12] C. Leys, M. Delacre, Y. L. Mora, D. Lakens, C. Ley, *How to Classify, Detect, and Manage Univariate and Multivariate Outliers, with Emphasis on pre-registration*, International Review of Social Psychology 32(1) (2019) 1–10.

[13] L. Xu, M. Mazur, X. Chen, Y. Chen, *Improving the Robustness of Fisheries Stock Assessment Models to Outliers in Input Data*, Fisheries Research 230 (2020).

[14] S. Nargis, *Robust Methods in Logistic Regression*, Unpublished Master Thesis, University of Canberra, (2005) Bruce ACT, Australia.

[15] C. Croux, C. Flandre, G. Haesbroeck, *The Breakdown Behavior of the Maximum Likelihood Estimator in the Logistic Regression Model*, Statistics & Probability Letters 60(4) (2002) 377–386.

[16] S. Ahmad, M. Norazan, H. Midi, *Robust Estimators in Logistic Regression: A Comparative Simulation Study*, Journal of Modern Applied Statistical Methods 9(2) (2010) 502–511.

[17] H. Aguinis, R. K. Gottfredson, H. Joo, *Best-Practice Recommendations for Defining, Identifying, and Handling Outliers*, Organizational Research Methods 16(2) (2013) 270–301.

[18] F. R. Hampel, E. M. Ronchetti, P. J. Rousseuw, W. A. Stahel, Robust statistics. The Approach Based on Influence Functions, John Wiley & Sons, New York, NY, 1986.

[19] H. Midi, S. B. Ariffin, *Modified Standardized Pearson Residual for the Identification of Outliers in Logistic Regression Model*, Journal of Applied Sciences 13 (2013) 828–836.

[20] D. Pregibon, *Logistic Regression Diagnostics*, The Annals of Statistics 9(4) (1981) 705–724.

[21] L. A. Stefanski, R. J. Carroll, D. Ruppert, *Optimally Bounded Score Functions for Generalized Linear Models with Applications to Logistic Regression*, Biometrika 73(2) (1986) 413–424.

[22] H. R. Künsch, L. A. Stefanski, 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) (1989) 460–466.

[23] R. Carroll, S. Pederson, *On Robust Estimation in the Logistic Regression Model*, Journal of the Royal Statistical Society Series B (Methodological) 55(3) (1993) 693–706.

[24] A. Christmann, *Least Median of Weighted Squares in Logistic Regression with Large Strata*, Biometrika 81(2) (1994) 413–417.

[25] A. Bianco, V. J. Yohai, *Robust Estimation in the Logistic Regression Model*, Robust Statistics, Data Analysis, and Computer Intensive Methods (1996) 17–34.

[26] E. Cantoni, E. Ronchetti, *Robust Inference for Generalized Linear Models*, Journal of the American Statistical Association 96(455) (2001) 1022–1030.

[27] C. Croux, G. Haesbroeck, *Implementing the Bianco and Yohai estimator for Logistic Regression*, Computational Statistics & Data Analysis 44(1-2) (2003) 273–295.

[28] P. J. Rousseeuw, A. Christmann, *Robustness Against Separation and Outliers in Logistic Regression*, Computational Statistics & Data Analysis 43(3) (2003) 315–332.

[29] H. Bondel, *Minimum Distance Estimation for the Logistic Regression Model*, Biometrika 92(3) (2005) 724–731.

[30] P. Čížek, *Robust and Efficient Adaptive Estimation of Binary-Choice Regression Models*, Journal of the American Statistical Association 103(482) (2008) 687–696.

[31] M. Valdora, V. J. Yohai, *Robust Estimators for Generalized Linear Models*, Journal of Statistical Planning and Inference 146 (2014) 31–48.

[32] G. Adimari, L. Ventura, *Robust Inference for Generalized Linear Models with Application to Logistic Regression*, Statistics & Probability 55(4) (2001) 413–419.

[33] I. A. I. Ahmed, W. Cheng, *The Performance of Robust Methods in Logistic Regression Model*, Scientific Research Publishing 10 (2020) 127–138.

[34] T. Parlak, *Lojistik Regresyonda Robust Tahmin Yöntemlerinin Kullanılması*, Yüksek Lisans Tezi, Ankara Üniversitesi (2019), Ankara, Türkiye.

[35] K. I. Penny, I. T. Jolliffe, *A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data*, Journal of the Royal Statistical Society: Series D (The Statistician) 50(3) (2001) 295–308.

[36] M. Šimecková, *Maximum Weighted Likelihood Estimator in Logistic Regression*, WDS'05 Proceedings of Contributed Papers Part I (2005) 144–148.

[37] B. D. Meyer, N. Mittag, *Misclassification in Binary Choice Models*, Journal of Econometrics 200(2) (2017) 295–311.

[38] R. W. M. Wedderburn, *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton method*, Biometrika 61(3) (1974) 439–447.

[39] R. A. Maronna, R. D. Martin, V. J. Yohai, M. Salibián-Barrera, Robust Statistics: Theory and Methods with R, John Wiley & Sons, New York, NY, 2019.

[40] M. Krzyśko, Ł. Smaga, *Selected Robust Logistic Regression Specification for Classification of Multi-dimensional Functional Data in presence of Outlier*, Folia Oeconomica 2(334) (2018) 53–66.

[41] P. J. Rousseeuw, A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, NY, 1987.

[42] R Development Core Team, R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2008.

[43] J. Wang, R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, K. Konis, Package "robust". R-Project, March 8 2020.

[44] M. Maechler, P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. S. Barrera, T. Verbeke, M. Koller, E. L. T. Conceicao, M. A. di Palma, Package "robustbase", R-Project, March 23, 2020.