Weighted high leverage collinear robust ridge estimator in logistic regression model

Article *in* Pakistan Journal of Statistics · January 2018

CITATIONS
0

READS
149

2 authors:

Syaiba Balqish Ariffin
University Malaysia of Computer Science and Engineering
7 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Habshah Midi
Universiti Putra Malaysia
224 PUBLICATIONS **2,296** CITATIONS
SEE PROFILE

Pak. J. Statist.
2018 Vol. 34(1), 55-75

WEIGHTED HIGH LEVERAGE COLLINEAR ROBUST RIDGE ESTIMATOR IN LOGISTIC REGRESSION MODEL

Syaiba Balqish Ariffin[1,§] and **Habshah Midi**[1,2]

[1] Faculty of Science, Universiti Putra Malaysia, Serdang Selangor, Malaysia

[2] Institute of Mathematical Research, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

[§] Corresponding author Email: syaibabalqish@gmail.com

ABSTRACT

The combination of high leverage points and multicollinearity problem occurs frequently in logistic regression model. Methods that successfully address these problems separately are not effective for the combined problems. A robust logistic ridge regression (RLR) which incorporates the weighted Bianco and Yohai (WBY) robust estimator with fully iterated logistic ridge regression (LR) is proposed to rectify the combined problems of high leverage points and multicollinearity in a data. A numerical example and simulation study are presented to compare the performance of the RLR with the ML, the WBY, and the LR estimators. Results of the study indicate that the RLR outperforms the established estimators for the combined problems.

KEY WORDS

Logistic ridge regression, Robust estimator, Diagnostic, High leverage point, Multicollinearity

1. INTRODUCTION

Two commonly occurring problems in logistic regression model are highly correlated predictor variables and the presence of extreme observations. Although the maximum likelihood (ML) estimator is fairly resistant to these problems in the moderate level, it is possible that only a single extreme observation (Syaiba and Habshah, 2010) or high correlation between two predictor variables (Schaefer et al., 1984) can render the ML coefficient estimates meaningless. For instance, a single extreme observation that departs significantly from the fit of good observations can pull the ML estimates toward itself, resulting biases in coefficient estimates (Habshah and Syaiba, 2012). Similarly, two highly correlated predictor variables can alter the ML coefficient estimates, so that they are not only differing from their true values by an order of magnitude, but they can actually switch sign (Weissfeld and Sereika, 1991; Lesaffre and Marx, 1993).

The high leverage point is referred as extreme or outlying observation in X-space and it may cause more problems to highly correlated predictor variables.Moreover, in the presence of multiple high leverage points, the more extreme cases could mask the effect

of another or swamp good observation as high leverage points (Syaiba and Habshah, 2010).

Multicollinearity is defined as a strong correlation between two or more continuous predictor variables. The existence of multicollinearity inflates the variance of coefficient estimates (Månsson and Shukur, 2011; Kibria et al., 2012). Another effect of multicollinearity is having a lack of statistical significance of individual predictor's Wald test while the overall model may be strongly significant (Lesaffre and Marx, 1993, Duzan and Shariff, 2015). Gunst (1983) identified misspecification error, sampling deficiencies, over-parameterized and the presence of high leverage point as the main factor of causing inter-correlation problem. In the last decades, some discussions appeared in the literature on the multicollinearity problem in the logistic regression model (Schaefer et al., 1984; Schaefer, 1986).Other researchers explored this problem in a generalized linear model (Marx and Smith, 1990; Weissfeld and Sereika, 1991; Segerstedt and Nyquist, 1992).

In order to accommodate the high correlation among predictors and to improve estimation, a ridge regression sacrifices small amounts of bias for large reductions in variance (Månsson and Shukur, 2011). Holland (1973) was the first to propose a solution for the combined problem of multicollinearity and outliers in linear regression framework by suggesting a ridge method applied to a robust estimator. Askin and Montgomery (1980) proposed an augmented robust method allowing for the ridge regression methods to be combined with the M-estimator. Godínez-Jaimeset al. (2012) investigated the relative effect of collinearity and separation on logistic regression by simulations. The simulation results illustrated that the values MSE and biases strongly affected by a low degree of overlapped cases and a high degree of correlation, particularly in small sample size. The work of Al-Aabdi and Al–Shaibani (2014) considered combining robust estimator and ridge estimator for the estimation of regression coefficient in the presence of multicollinearity and outlier in a data. In order to reduce the effect of outliers, robust ridge parameter for k is computed, and this parameter is used to obtain smaller MSE compared to the MSE of Ridge estimator. However, no specific robust estimator is stated in their paper. For classifying samples and features selection, Park and Konishi (2015) employed weighting scheme on log-likelihood function based on the Principal Component estimator to robustify penalized logistic regression model. Further, this study derived selection criterion for choosing the tuning parameter. Monte Carlo simulation results illustrated that the proposed robust penalized logistic regression outperforms sparse logistic regression, in terms of having smaller variances and bias, in combination problems of collinearities-outliers.

We expect that the standard errors of the LR estimates would be larger when multicollinear data is contaminated with high leverage point. Although the LR estimator serves as a better alternative in dealing with multicollinearity (Barker and Brown 2001; Vago and Kemeny, 2006; Park and Hastie, 2008; *Godínez*-Jaime et al., 2012; Shahmandi et al., 2013) there is uncertainty that the LR estimator performs equally good when multicollinearity and high leverage point occur together. In order to improve the performance of LR estimator, we propose incorporating a robust weighted Bianco and Yohai (WBY) estimator by Croux and Haesbroeck (2003) in the LR algorithm.

In Section 2, we provide a description of the LR estimator followed by a discussion in Section 3 of the proposed RLR estimator. A brief explanation on diagnostic approaches is reviewed in Section 4. Section 5 contains a simulation experiment. Applications on artificial data and real example data are introduced in Section 6 and Section 7. Section 8 offers some conclusions.

2. LOGISTIC RIDGE REGRESSION

The ridge regression was proposed to estimate regression coefficients with smaller mean squared error than their least squares when predictor variables are correlated (Hoerl and Kennard, 1970). In logistic regression, similar problems arise when estimating regression coefficients in collinear data. Therefore, the linear ridge regression has been extended to the logistic ridge regression and demonstrated coefficient estimates with smaller mean squared error than the ML estimator when the predictor variables are highly correlated (Schaefer et al., 1984; Schaefer, 1986). Månsson and Shukur (2011) and Kibria et al. (2012) improved the LR estimator by introducing a selection of ridge parameters, while Kibria and Salleh (2012) and Locking et al. (2013) applied it in a probit regression model. The LR requires the specification of a penalty parameter (or ridge parameter) that controls the degree of shrinkage of the coefficient estimates (Cule and De Iorio, 2012). A number of methods have been proposed to estimate the ridge parameter in the LR based on a simulation study (Månsson and Shukur, 2011; Kibria et al., 2012), but no consensus methods provide a universally optimum choice. In this section, we describe the LR estimator with the best option of ridge parameter as suggested by Månsson and Shukur (2011) and Kibria et al. (2012), and how this method can be improved by implementing the robust WBY estimator in formulating the algorithm of RLR estimator.

In this section, we describe the algorithm which we use to compute the RLR coefficient estimates. In the logistic regression model, let X be a $n \times p$ matrix of predictors with rows$x_i = (x_{i1}, \dots x_{ip})$. Let $y = (Y_1, \dots Y_n)$ be a vector of observed binary outcomes, $Y_i \in \{0,1\}$. The i^{th} response $Y_i$is a Bernoulli variable with probability of success $\pi(x_i)$. The logistic regression model related to the probability of success, $\pi(x_i)$ can be written as:

$$Pr(Y_i = 1|x_i) = \pi(x_i) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \tag{1}$$

where β is a vector of estimated parameter. The RLR estimates are obtained by maximizing the log-likelihood of the parameter vector, subject to ridge parameter. The penalized log-likelihood to be maximized is:

$$l(\beta) = \sum_{i=1}^{n} Y_i log\big(\pi(x_i)\big) + \sum_{i=1}^{n} (1 - Y_i) log\big(1 - \pi(x_i)\big) - k\|\beta^2\|. \tag{2}$$

One of the drawbacks of using the ML estimator is that the asymptotic variance becomes inflated when the predictor variables are highly correlated because some of the eigenvalues will be small (Månsson and Shukur, 2011). As a remedy to multicollinearity, Schaefer et al. (1984) suggested using the iterative LR coefficient estimates:

$$\hat{\beta}_{LR} = \big(X^T \widehat{W}_{ML} X + kI\big)\big(X^T \widehat{W}_{ML} X \hat{\beta}_{ML}\big) \tag{3}$$

where $\widehat{W}_{ML}$, $\hat{\beta}_{ML}$ and k are estimated by the ML estimator. The asymptotic of Mean Squared Error (MSE) of the LR equals:

$$E(L_{LR}^2) = \sum_{j=1}^{J} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \beta^T (X^T W X + kI)^{-2} \beta \quad (4)$$

where the first term is the asymptotic variance and the second term is the squared bias. By replacing λ_j with $\lambda_j + k$ in the denominator, the asymptotic variance is no longer inflated. Earlier recommendation for the ridge parameter is $k = p/\hat{\beta}^T \beta$ where k is a diagonal matrix of non-negative constants, $k > 0$and I is an identity matrix (Schaefer et al., 1984).This ridge parameter is quite conservative (relatively small) under extreme correlation. But, if k increases, squared bias also increases. Therefore, the choice of k is based on a logical balance between the decrease in the variance that should be larger than the increase of the squared bias (Månsson and Shukur, 2011).

In this paper, we consider the best option, k for the RLR estimates as suggested by Månsson and Shukur (2011) and Kibria et al. (2012) when the degree of correlation is high. The recommended, k is as follows:

$$k = max\left(\frac{1}{m_j}\right) \quad (5)$$

where $m_j = \sqrt{\hat{\sigma}^2/\hat{\alpha}_{max}^2}$ and $\hat{\sigma}^2 = (y - \hat{\pi}_i)^T (y - \hat{\pi}_i)/n - p - 1$. Given $\hat{\alpha}_j^2$ is an element of $\left(\gamma \hat{\beta}_{ML}\right)^2$ where λ_j are eigenvalues and γ an eigenvector of $X^T W X$. It is important to note that the computation of ridge parameter, k does not include the intercept term.

3. ROBUST LOGISTIC RIDGE REGRESSION

The ML estimator is the most efficient estimator, but it may behave poorly in the presence of high leverage point. Therefore, alternative estimator needs to be constructed. Several robust estimators have been introduced in logistic regression, and some of these estimators being standard available in statistical software packages (Croux and Haesbroeck, 2003; Habshah and Syaiba, 2012). So far, the robust WBY estimator serves the best estimates when dealing with high leverage points compared to other existing robust estimators in the literature. The advantage of weighting scheme in WBY estimator where it down weights the effect of high leverage points is extensively investigated by Habshah and Syaiba (2012).

In this paper, the focus is on the modification of Eq. (3) which consists of replacing the ML estimates by the WBY estimates. The WBY estimator is a weighted version of the Bianco and Yohai (BY) estimator. The BY estimator is found to be much more resistant than the ML estimator, but it may not has a bounded influence function. In order to obtain an overall bounded influential function, one can add a weighting step to down weight the high leverage point (Croux and Haesbroeck, 2003). Weighting function has already been used by several authors to render estimators more robust (Carroll and Pederson, 1993). Initially, high leverage points need to be identified before parameter estimation. Hence, the WBY estimator with an overall bounded influence function is

obtained where the effect of high leverage points is reduced by assigning a proper weights to each of observation (Croux and Haesbroeck, 2003). The WBY estimator can be defined as:

$$\hat{\gamma}_n = \underset{\gamma}{argmin} \sum_{i=1}^{n} \omega_i \varphi_{BY}(X_i^T \beta; Y_i) \tag{6}$$

where the deviance functions is given by

$$d(X_i^T \beta; Y_i) = -Y_i log(\pi_i) - (1 - Y_i) log(1 - \pi_i). \tag{7}$$

The function of φ_{BY} is a positive and continuously differentiable function for the BY estimator. It needs to satisfy $\varphi(s;0) = \varphi(-s;1)$ for any score s where a score value $s_i = x_i^T \beta$ is obtained as a linear combination of a given β. The weights are computed using:

$$\omega_i = \begin{cases} 1 & if \quad RMD_i^2 \leq \chi^2_{p,0.975} \\ 0 & else \quad RMD_i^2 > \chi^2_{p,0.975} \end{cases}. \tag{8}$$

Identification of high leverage point is determined by computing the Robust Mahalanobis Distance (RMD) where the estimation subset is determined by minimum covariance determinant (MCD) by Rousseeuw and Van Driessen (1999). The MCD estimator selects a subset of h observations out of n which minimizes the determinant of the covariance matrix corresponding to these h points. The h is set at $3n/4$ yielding 25% breakdown point estimator. The weighting is only based on X matrix without constant term. Since the weight function corresponding to the weighting scheme equals zero for large leverage points, the influence function of the WBY is bounded. For detailed information on algorithm and formula, we refer to Croux and Haesbroeck (2003). Then, we define iterative coefficients update for the RLR as follows:

$$\hat{\beta}_{RLR} = \left(X^T \widehat{W}_{WBY} X + k^* I\right)^{-1} \left(X^T \widehat{W}_{WBY} X \hat{\beta}_{WBY}\right) \tag{9}$$

where $\widehat{W}_{WBY}$, $\hat{\beta}_{WBY}$ and $k^*$are estimated by the WBY estimator.

4. DIAGNOSTIC APPROACH

In linear regression models, the condition of the information matrix $X^T X$ is directly affected by collinearity among the predictors. However, in the logistic regression model, the information matrix is $X^T W$X where W is a diagonal matrix of weights which is determined by the fitting algorithm at each iteration. It is collinearity in the weighted predictors $W^{1/2}X$ which directly affects the condition of the information matrix. The dissimilarity between collinearity among regressor and the ML collinearity due to ill-conditioned information matrix is well explained by Lessafre and Marx (1993). Likewise, the linear regression, identification of multicollinearity in logistic regression is detected using condition indices (CI) and condition number (CN) by computing eigenvalues from $X^T X$ and $X^T WX$ (Lesaffre and Marx, 1993). We summarize their algorithm as below:

Step 1.

Scaling the columns of X (including the intercept term) to unit length $x_{ij}^* = x_{ij} / \left\| X_j \right\|, i = 1, \ldots, n\ ;\ j = 1, \ldots, p.$

Step 2.

Compute eigenvalues $\hat{\lambda}_0, \ldots, \hat{\lambda}_p$ from information matrix, $\widehat{W} = (X^T W X)$ and $\hat{X} = (X^T X)$in decreasing order.

Step 3.

Define the condition indexes of $\kappa_{W_j} = \left(\hat{\lambda}_0/\hat{\lambda}_j\right)^{1/2}$ and $\kappa_{X_j} = \left(\hat{\lambda}_0/\hat{\lambda}_j\right)^{1/2}$.

Step 4.

Define the condition numbers of $\kappa_W = \hat{\lambda}_0/\hat{\lambda}_p$ and $\kappa_X = \hat{\lambda}_0/\hat{\lambda}_p$ and ratio $r_{WX} = \kappa_W/\kappa_X$.

Step 5.

Determine whether there is an ill-conditioning in$X$ and ML. According to Lesaffre and Marx (1993), if $\kappa_X \geq 30$, there is collinearity in X , if $\kappa_W \geq 30$ and κ_X is not high, there is ML-collinearity. If both (≥ 30) and ratio is $r_{WX} \geq 1$, there are collinearity exist in both X and ML.

Step 6.

Calculate the variance decomposition proportion table of $\hat{X}$ and $\widehat{W}$ to show determine which predictors and weighted predictors that highly correlated

Statistical practitioners also rely on diagnostic approaches when dealing with multicollinearity or high leverage points. They tend to delete influential predictors in model before making inference, leaving only a few overlapped cases between $Y = 0$ and $Y = 1$ which can also cause large estimated regression coefficients and standard errors (Segerstedt and Nyquist, 1992; Christmann and Rousseeuw, 2001; Rousseeuw and Christmann, 2003).

Meanwhile, for detection of high leverage points, we recommended using robust logistic diagnostic (RLGD) method by Syaiba and Habshah (2010) which is capable of detecting high leverage points correctly. The algorithm for RLGD method is described as follow:

Step 1.

Identify suspected high leverage point, d using the RMD-MCD. Perform two sets, set D (suspected high leverage points) and set R (good observations).

Step 2.

Estimate $\hat{\beta}^{(-D)}$ using the ML method.

Step 3.

Compute $\hat{\pi}_i^{(-D)} = exp\left(x_i^T \hat{\beta}^{(-D)}\right)/1 + exp\left(x_i^T \hat{\beta}^{(-D)}\right)$ and

Step 4.

Compute $v_i^{(-D)} = \hat{\pi}_i^{(-D)}\left(1 - \hat{\pi}_i^{(-D)}\right)$

Step 5.

Define $\tilde{X} = V^{1/2} X$

Step 6.

Compute

$$b_i^{(-D)} = x_i^T \left(\tilde{X}_R^T \tilde{X}_R\right)^{-1} x_i$$

and

$$p_{ii}^{*(-D)} = \begin{cases} b_i^{(-D)}/1 + b_i^{(-D)} & ; i \in R \\ b_i^{(-D)} & ; i \in D \end{cases}$$

Step 7.

Compare$p_{ii}^{*(-D)} > Med\left(p_{ii}^{*(-D)}\right) + cMAD(p_{ii}^{*(-D)})$. Any group deleted points with $p_{ii}^{*(-D)}$ exceeds the cut-off point, are finalized and declares as high leverage points

5. SIMULATION EXPERIMENT

It is a demand to conduct a Monte Carlo simulation when exact theoretical solutions are not available to support our expectation. In this section, the performance of RLR estimator is investigated to see whether we can rely on its parameter estimates when the predictors are highly correlated in the presence of high leverage points. Most authors considered various degrees of correlation, the number of sample size and the number of predictors as important factors that may affect the properties of the different estimates by the ML and LR estimators (Månsson and Shukur, 2011; Kibria et al., 2012). In this study, we consider another factor (the number of high leverage points) besides those already mentioned, when evaluating the performance of the ML, LR, WBY and RLR estimators.

The X matrix with multicollinearity is generated with various degrees of correlation using the following equation:

$$X_{ij} = (1 - \rho^2)Z_{ij} + \rho Z_{ip} \tag{10}$$

$i = 1, \dots, n$ and $j = 1, \dots, p$ where ρ represent the correlation between predictors with $Z_{ij} \sim N(0,1)$ which are randomly generated from the normal distribution. The original idea of generating multicollinear data was given by Lawrence and Arthur (1990), which extensively been used by Bagheri et al. (2010), Bagheri and Midi (2012, 2015) in multicollinearity problem in linear regression framework. As an alternative, Månsson and Shukur (2011) generated the X matrix according to $X_{ij} = \sqrt{(1 - \rho^2)}Z_{ij} + \rho Z_{ip}$ instead of Eq. (10) for conducting a simulation study in linear regression as well as in logistic regression. Meanwhile, the responses y_i are generated using the Bernoulli distribution $y_i \sim Bern(\pi_i)$ where we define the probabilities as:

$$\pi_i = exp(x_i^T \beta)/1 + \left(exp(x_i^T \beta)\right). \tag{11}$$

Here, the true values for regressors are fixed as $\beta_j = 1/\sqrt{p}$ and $\beta_0 = 0$. The value of the intercept reflects the average value of the log odds ratio. Hence, when the intercept equals zero, then there is an equal average probability of obtaining one and zero for responses. Three different degrees of correlation are considered, $\rho = (0.75, 0.85, 0.95)$ for a number of predictors, $p = (2,3,4)$ and sample sizes, $n = (60,80,100,150,200)$. In the contamination set, $h = (1,3,5)$ are the number high leverage points that are generated

using the uniform distribution, $h_{kj} \sim U(10,11)$ with a response y_k which is fixed at $Y = 0$. Then, the contaminated observations h_{kj} are plugged on the last rows of X_{ij}.

The best estimator is determined based on a comparison of lowest MSE and bias among the estimators (Månsson and Shukur, 2011; Kibria et al., 2012). It is important to note that the estimator with the lowest MSE may have a larger bias since the reduction of variance in the LR estimates is larger than the increase in squared bias depending on the number of sample size and magnitude of ridge parameter. The estimated coefficients of all estimators are computed over $R = 1000$ replications and contain summary measures of MSE and bias combining the individual results for the estimated coefficients including intercept term. Bias and MSE measures are computed in Euclidean norm as follows:

$$Bias = \left\| \frac{1}{R} \sum_{i=1}^{R} \hat{\beta}_i - \beta \right\| \text{ and } MSE = \frac{1}{R} \sum_{i=1}^{R} \left\| \hat{\beta}_i - \beta \right\|^2.$$

Referring to Schaefer et al. (1984), their preliminary findings from the empirical study indicated that the LR estimator has smaller MSE than MSE from the ML estimator when sample size, $n \geq 250$ with the degree of correlation, $\rho \geq 0.90$. As pointed out by Victoria-Feser (2002), the ML estimates are unstable for small sample size even without contamination. Since the LR estimator is using $\hat{\beta}_{ML}$, the LR estimates may fluctuate for small sample size. According to Victoria-Feser (2002), sample size, $n = 50$ is considerably very small for $p = 2$. The choice for sample size starting with $n = 100$ and $p = 2$ is recommended by Victoria-Feser (2002) to ensure the stability of the estimators. Peduzzi et al. (1996) emphasized that adjusting the number of sample size is more crucial than correcting for the degrees of freedom to obtain meaningful estimates in logistic regression. In our simulation, we are successfully in showing the MSE for all estimators reduce when the number of sample sizes increases starting with the smallest sample size, $n = 60$.

It is generally believed that the LR estimator should always be preferred when dealing with multicollinear data. However, the RLR estimator is expected to be better, since we are not able to ensure that multicollinear data is free from the high leverage points. In this section, we would like to discuss the advantages of incorporating the robust WBY estimator in logistic ridge regression.

Tables 1-4 present the results of simulation experiments concerning the MSEs and biases of the ML, the LR, the WBY and the RLR estimators. We will discuss on how these estimators are related to the degrees of correlation, to the number of sample sizes, to the number of high leverage points and to the number of predictors.

Table 1: Comparison of MSEs and Biases for Uncontaminated Data

n		ML	LR	WBY	RLR	ML	LR	WBY	RLR	ML	LR	WBY	RLR
		rho=0.75, p=2				rho=0.75, p=3				rho=0.75, p=4			
60	MSE	0.769	0.142	0.915	0.170	1.609	0.198	2.253	0.219	2.587	0.279	3.962	0.300
	Bias	0.113	0.229	0.100	0.254	0.148	0.235	0.171	0.262	0.156	0.272	0.236	0.311
80	MSE	0.513	0.121	0.647	0.141	1.057	0.189	1.253	0.208	1.674	0.276	1.912	0.298
	Bias	0.064	0.153	0.071	0.169	0.108	0.150	0.112	0.165	0.120	0.170	0.222	0.191
100	MSE	0.362	0.118	0.433	0.132	0.723	0.188	0.873	0.207	1.263	0.256	1.517	0.268
	Bias	0.060	0.117	0.062	0.125	0.070	0.102	0.083	0.110	0.101	0.105	0.114	0.118
150	MSE	0.255	0.115	0.297	0.120	0.463	0.185	0.541	0.206	0.742	0.230	0.881	0.250
	Bias	0.035	0.073	0.037	0.074	0.043	0.057	0.049	0.058	0.060	0.055	0.068	0.053
200	MSE	0.178	0.099	0.214	0.107	0.355	0.179	0.410	0.198	0.547	0.218	0.623	0.246
	Bias	0.026	0.051	0.034	0.057	0.040	0.033	0.040	0.032	0.042	0.023	0.047	0.023
		rho=0.85, p=2				rho=0.85, p=3				rho=0.85, p=4			
60	MSE	1.686	0.118	2.269	0.134	3.766	0.198	5.392	0.212	7.300	0.258	9.845	0.270
	Bias	0.117	0.249	0.148	0.270	0.145	0.286	0.191	0.331	0.197	0.333	0.259	0.416
80	MSE	1.202	0.114	1.397	0.131	2.519	0.172	3.211	0.188	4.261	0.209	4.802	0.251
	Bias	0.085	0.175	0.092	0.189	0.099	0.184	0.129	0.206	0.136	0.229	0.212	0.256
100	MSE	0.857	0.110	1.031	0.120	1.774	0.152	2.199	0.185	3.079	0.180	3.863	0.219
	Bias	0.081	0.125	0.082	0.135	0.089	0.128	0.081	0.142	0.098	0.154	0.123	0.175
150	MSE	0.575	0.103	0.659	0.115	1.160	0.136	1.302	0.145	1.918	0.152	2.268	0.167
	Bias	0.038	0.075	0.048	0.078	0.065	0.068	0.079	0.074	0.067	0.082	0.080	0.087
200	MSE	0.394	0.100	0.489	0.109	0.811	0.125	0.977	0.137	1.386	0.150	1.608	0.163
	Bias	0.034	0.045	0.042	0.046	0.053	0.045	0.057	0.046	0.061	0.042	0.071	0.046
		rho=0.95, p=2				rho=0.95, p=3				rho=0.95, p=4			
60	MSE	11.922	0.176	17.820	0.201	30.449	0.264	43.095	0.287	54.907	0.329	79.674	0.368
	Bias	0.188	0.371	0.195	0.398	0.296	0.481	0.367	0.497	0.230	0.548	0.676	0.579
80	MSE	8.833	0.108	11.792	0.123	21.335	0.155	25.649	0.180	35.828	0.217	51.269	0.254
	Bias	0.130	0.275	0.092	0.291	0.257	0.349	0.229	0.383	0.176	0.437	0.523	0.469
100	MSE	6.427	0.071	7.883	0.083	14.552	0.098	17.775	0.113	26.479	0.134	32.873	0.158
	Bias	0.083	0.199	0.095	0.218	0.158	0.267	0.148	0.290	0.166	0.333	0.221	0.361
150	MSE	4.728	0.045	5.437	0.049	9.749	0.052	11.733	0.059	16.072	0.065	19.555	0.075
	Bias	0.070	0.128	0.086	0.134	0.139	0.161	0.121	0.176	0.142	0.202	0.139	0.224
200	MSE	3.176	0.037	3.896	0.040	6.894	0.041	8.053	0.044	11.196	0.043	13.387	0.048
	Bias	0.049	0.076	0.072	0.082	0.131	0.106	0.118	0.114	0.080	0.131	0.095	0.142

Table 2: Comparison of MSEs and Biases for Contaminated Data with Two Predictors

n		ML	LR	WBY	RLR	ML	LR	WBY	RLR	ML	LR	WBY	RLR
		rho=0.75, p=2, h=1				rho=0.75, p=2, h=3				rho=0.75, p=2, h=5			
60	MSE	1.379	0.966	0.909	0.156	1.475	1.206	0.931	0.160	1.487	1.278	0.975	0.166
	Bias	0.889	0.904	0.104	0.242	1.044	1.042	0.126	0.239	1.087	1.083	0.105	0.235
80	MSE	1.104	0.835	0.620	0.140	1.308	1.163	0.666	0.140	1.366	1.240	0.602	0.146
	Bias	0.809	0.822	0.092	0.166	1.017	1.018	0.085	0.164	1.064	1.063	0.084	0.157
100	MSE	0.887	0.714	0.487	0.142	1.207	1.119	0.463	0.138	1.249	1.197	0.486	0.141
	Bias	0.724	0.736	0.059	0.124	0.996	0.998	0.052	0.127	1.044	1.043	0.081	0.124
150	MSE	0.495	0.438	0.280	0.125	1.058	1.012	0.285	0.129	1.151	1.111	0.307	0.134
	Bias	0.538	0.551	0.035	0.075	0.945	0.947	0.041	0.082	1.008	1.009	0.040	0.077
200	MSE	0.337	0.307	0.200	0.113	0.927	0.899	0.203	0.112	1.071	1.041	0.204	0.119
	Bias	0.424	0.437	0.032	0.054	0.890	0.892	0.029	0.051	0.976	0.978	0.024	0.048
		rho=0.85, p=2, h=1				rho=0.85, p=2, h=3				rho=0.85, p=2, h=5			
60	MSE	2.806	0.892	2.144	0.136	1.837	1.157	2.107	0.134	1.636	1.233	2.072	0.142
	Bias	0.826	0.890	0.116	0.271	1.024	1.031	0.124	0.261	1.074	1.073	0.115	0.267
80	MSE	2.144	0.765	1.377	0.128	1.574	1.126	1.447	0.126	1.538	1.203	1.579	0.132
	Bias	0.759	0.794	0.091	0.185	0.998	1.004	0.078	0.170	1.049	1.051	0.114	0.175
100	MSE	1.726	0.652	1.119	0.122	1.528	1.105	1.063	0.121	1.411	1.182	1.075	0.120
	Bias	0.665	0.702	0.069	0.137	0.972	0.981	0.055	0.135	1.029	1.031	0.069	0.134
150	MSE	0.925	0.435	0.607	0.115	1.358	1.031	0.644	0.111	1.268	1.121	0.635	0.119
	Bias	0.493	0.516	0.039	0.073	0.913	0.923	0.043	0.072	0.991	0.995	0.059	0.064
200	MSE	0.643	0.343	0.461	0.111	1.204	0.934	0.454	0.100	1.193	1.074	0.478	0.115
	Bias	0.390	0.409	0.031	0.049	0.846	0.856	0.029	0.049	0.958	0.961	0.048	0.045
		rho=0.95, p=2, h=1				rho=0.95, p=2, h=3				rho=0.95, p=2, h=5			
60	MSE	33.448	0.818	17.969	0.204	8.381	1.053	17.298	0.214	3.700	1.155	18.992	0.214
	Bias	0.648	0.899	0.204	0.399	0.966	1.010	0.111	0.403	1.050	1.056	0.190	0.399
80	MSE	28.023	0.702	11.216	0.126	8.078	1.011	10.941	0.132	3.292	1.118	11.155	0.130
	Bias	0.612	0.827	0.148	0.293	0.933	0.985	0.086	0.293	1.023	1.034	0.081	0.282
100	MSE	23.890	0.593	8.558	0.089	7.943	0.973	8.354	0.088	3.240	1.093	8.107	0.089
	Bias	0.530	0.751	0.102	0.226	0.908	0.962	0.055	0.227	1.003	1.016	0.072	0.221
150	MSE	16.054	0.393	4.873	0.050	7.868	0.880	4.829	0.050	2.812	1.037	4.956	0.055
	Bias	0.401	0.590	0.094	0.130	0.808	0.900	0.047	0.126	0.951	0.976	0.061	0.138
200	MSE	12.321	0.279	3.393	0.039	7.127	0.788	3.672	0.041	2.753	0.987	3.642	0.042
	Bias	0.324	0.473	0.050	0.080	0.726	0.829	0.042	0.084	0.900	0.939	0.032	0.085

Table 3: Comparison of MSEs and Biases for Contaminated Data with Three Predictors

n		ML	LR	WBY	RLR	ML	LR	WBY	RLR	ML	LR	WBY	RLR
		rho=0.75, p=3, h=1				rho=0.75, p=3, h=3				rho=0.75, p=3, h=5			
60	MSE	2.124	1.068	1.924	0.228	1.861	1.272	2.184	0.229	1.749	1.347	2.161	0.232
	Bias	0.876	0.901	0.161	0.259	1.022	1.027	0.158	0.264	1.065	1.064	0.170	0.258
80	MSE	1.568	0.967	1.313	0.217	1.595	1.255	1.283	0.224	1.543	1.329	1.410	0.228
	Bias	0.790	0.819	0.128	0.152	1.002	1.005	0.109	0.156	1.044	1.044	0.110	0.159
100	MSE	1.198	0.820	0.919	0.214	1.476	1.228	0.991	0.221	1.660	1.292	0.956	0.226
	Bias	0.714	0.728	0.073	0.119	0.981	0.985	0.094	0.113	1.017	1.019	0.081	0.109
150	MSE	0.724	0.548	0.586	0.205	1.251	1.129	0.573	0.211	1.424	1.278	0.588	0.216
	Bias	0.530	0.544	0.046	0.055	0.935	0.939	0.050	0.057	0.987	0.990	0.060	0.055
200	MSE	0.475	0.393	0.388	0.198	1.079	1.007	0.406	0.204	1.306	1.207	0.403	0.209
	Bias	0.415	0.427	0.033	0.030	0.880	0.883	0.040	0.037	0.961	0.963	0.039	0.032
		rho=0.85, p=3, h=1				rho=0.85, p=3, h=3				rho=0.85, p=3, h=5			
60	MSE	5.339	0.928	4.947	0.271	2.961	1.207	4.902	0.276	2.423	1.267	5.037	0.232
	Bias	0.782	0.887	0.177	0.312	0.996	1.012	0.174	0.305	1.048	1.050	0.188	0.311
80	MSE	3.959	0.832	3.133	0.244	2.576	1.198	3.295	0.241	2.049	1.244	3.323	0.193
	Bias	0.723	0.794	0.123	0.203	0.966	0.988	0.121	0.202	1.025	1.031	0.124	0.207
100	MSE	3.101	0.723	2.332	0.214	2.393	1.185	2.408	0.218	1.871	1.237	2.268	0.181
	Bias	0.649	0.699	0.083	0.152	0.944	0.967	0.108	0.143	1.010	1.014	0.108	0.129
150	MSE	1.824	0.543	1.452	0.193	2.067	1.173	1.375	0.193	1.696	1.229	1.440	0.173
	Bias	0.483	0.516	0.055	0.074	0.892	0.911	0.055	0.071	0.971	0.980	0.086	0.070
200	MSE	1.171	0.454	0.918	0.149	1.808	1.113	0.972	0.154	1.560	1.216	0.965	0.152
	Bias	0.384	0.409	0.045	0.037	0.822	0.840	0.053	0.040	0.940	0.946	0.047	0.042
		rho=0.95, p=3, h=1				rho=0.95, p=3, h=3				rho=0.95, p=3, h=5			
60	MSE	66.629	0.853	44.180	0.304	29.908	1.010	43.368	0.293	11.853	1.105	43.347	0.304
	Bias	0.507	0.918	0.203	0.515	0.867	0.994	0.212	0.502	0.998	1.031	0.235	0.505
80	MSE	56.513	0.765	25.294	0.184	26.203	0.973	26.291	0.186	11.659	1.087	27.227	0.190
	Bias	0.443	0.864	0.163	0.385	0.834	0.970	0.169	0.382	0.979	1.015	0.230	0.389
100	MSE	47.478	0.682	19.199	0.122	23.855	0.945	20.180	0.128	10.964	1.066	20.085	0.126
	Bias	0.408	0.809	0.146	0.299	0.798	0.951	0.150	0.299	0.954	0.998	0.189	0.299
150	MSE	34.994	0.495	11.958	0.061	21.375	0.875	11.483	0.060	9.311	1.027	10.953	0.062
	Bias	0.342	0.666	0.094	0.178	0.765	0.895	0.132	0.171	0.897	0.962	0.116	0.168
200	MSE	27.209	0.370	7.641	0.043	20.717	0.806	8.176	0.046	9.034	0.996	8.470	0.045
	Bias	0.311	0.548	0.047	0.106	0.653	0.830	0.059	0.111	0.850	0.926	0.073	0.111

Table 4: Comparison of MSEs and Biases for Contaminated Data with Four Predictors

n		ML	LR	WBY	RLR	ML	LR	WBY	RLR	ML	LR	WBY	RLR
		rho=0.75, p=4, h=1				rho=0.75, p=4, h=3				rho=0.75, p=4, h=5			
60	MSE	2.854	1.136	6.817	0.350	2.292	1.332	4.695	0.360	2.171	1.374	4.010	0.367
	Bias	0.857	0.898	0.273	0.300	1.010	1.015	0.251	0.299	1.051	1.050	0.275	0.301
80	MSE	2.142	1.056	2.241	0.315	1.987	1.325	2.182	0.350	1.780	1.362	2.572	0.362
	Bias	0.792	0.818	0.181	0.192	0.990	0.995	0.225	0.185	1.034	1.032	0.262	0.175
100	MSE	1.585	0.915	1.450	0.306	1.725	1.316	1.521	0.313	1.651	1.345	1.670	0.323
	Bias	0.702	0.721	0.106	0.108	0.970	0.977	0.130	0.106	1.017	1.020	0.118	0.113
150	MSE	0.917	0.639	0.833	0.299	1.424	1.230	0.916	0.310	1.392	1.283	0.913	0.321
	Bias	0.531	0.546	0.061	0.052	0.929	0.934	0.072	0.047	0.988	0.991	0.072	0.049
200	MSE	0.638	0.487	0.628	0.274	1.253	1.117	0.636	0.270	1.293	1.212	0.634	0.290
	Bias	0.419	0.433	0.052	0.026	0.872	0.876	0.049	0.022	0.961	0.964	0.055	0.019
		rho=0.82, p=4, h=1				rho=0.85, p=4, h=3				rho=0.85, p=4, h=5			
60	MSE	8.758	0.959	10.581	0.278	4.781	1.177	15.181	0.284	3.327	1.271	10.390	0.282
	Bias	0.754	0.891	0.261	0.379	0.962	1.002	0.627	0.381	1.030	1.036	0.358	0.383
80	MSE	6.527	0.873	6.089	0.222	4.010	1.219	6.241	0.226	2.758	1.240	5.774	0.230
	Bias	0.693	0.798	0.214	0.252	0.943	0.980	0.319	0.252	1.010	1.018	0.288	0.246
100	MSE	4.978	0.775	3.834	0.218	3.384	1.304	3.924	0.222	2.594	1.360	4.230	0.229
	Bias	0.629	0.703	0.127	0.164	0.925	0.958	0.152	0.168	0.992	1.003	0.136	0.167
150	MSE	2.725	0.622	2.142	0.166	2.799	1.274	2.388	0.169	2.259	1.339	2.427	0.172
	Bias	0.486	0.527	0.089	0.084	0.874	0.904	0.089	0.079	0.957	0.970	0.090	0.081
200	MSE	1.780	0.543	1.560	0.163	2.539	1.258	1.652	0.165	2.057	1.291	1.599	0.168
	Bias	0.387	0.417	0.057	0.044	0.810	0.833	0.058	0.048	0.928	0.938	0.055	0.037
		rho=0.95, p=4, h=1				rho=0.95, p=4, h=3				rho=0.95, p=4, h=5			
60	MSE	103.368	0.881	86.503	0.361	59.963	1.000	107.554	0.409	27.256	1.071	76.004	0.401
	Bias	0.544	0.934	1.571	0.578	0.792	0.994	1.226	0.612	0.998	1.020	0.296	0.603
80	MSE	82.451	0.818	48.156	0.242	55.477	0.962	50.336	0.262	22.021	1.045	52.623	0.250
	Bias	0.431	0.895	0.940	0.459	0.729	0.972	0.447	0.475	0.924	1.002	0.259	0.458
100	MSE	68.863	0.755	33.054	0.155	48.577	0.934	34.356	0.158	21.906	1.029	36.912	0.157
	Bias	0.401	0.854	0.219	0.357	0.704	0.952	0.273	0.358	0.928	0.989	0.245	0.353
150	MSE	51.680	0.595	18.507	0.071	38.164	0.879	20.323	0.073	21.140	0.998	19.510	0.078
	Bias	0.384	0.735	0.190	0.216	0.696	0.899	0.124	0.214	0.858	0.956	0.166	0.221
200	MSE	40.738	0.457	13.020	0.049	35.224	0.832	14.115	0.051	20.792	0.984	14.243	0.051
	Bias	0.318	0.616	0.119	0.140	0.620	0.845	0.096	0.148	0.789	0.924	0.093	0.147

Analyses of Simulated Data

Let us first focus on the simulation result for multicollinearity as shown in Table 1. The estimated MSEs and biases for the ML are severely affected by the presence of multicollinearity evident by larger MSEs and biases as the degree of correlation and number of predictor increases. We do not expect that the WBY estimator to be better than the LR since the WBY is not robust to multicollinearity. Unexpectedly, the WBY estimates are even worse than the ML estimates, but getting closer to the ML estimates as the sample size increases. The LR estimator gives the best estimates evident by the smallest MSEs and biases, followed by the RLR estimates which are fairly close the LR estimates. All of the estimators show reduction of MSEs and biases with the increment of sample size.

High leverage point has a clear negative impact on the ML estimates (see Table 2). As can be expected, the ML estimator performs even worst with both contaminations of multicollinearity and high leverage points, as it gives the largest MSEs and biases for mild correlation, $\rho = 0.75$. We observe that the high leverage point have an effect on reducing the MSEs of ML, especially for higher degree of correlation, $\rho = 0.95$. This could mislead our interpretation since it is become a general understanding that high leverage point induces larger MSEs as the number of contamination increases. It is interesting to observe that in the presence of high leverage point in multicollinear data, the LR estimator does not perform well by looking at the estimated MSEs and biases that are larger compared to the LR estimator having the only multicollinearity for every degree of correlation increases. The WBY estimates get affected as well, particularly for smaller sample size at higher degree of correlation. On the other hand, the RLR estimates give the smallest MSEs and biases.

The final factor that we vary is the number of predictor variables (see Table 3 and Table 4). It is very easy to compare directly the MSEs and biases when we increase the number of predictors since the number of sample sizes are fixed. The ML is the most affected followed by the LR, as we increase the number of high leverage point from h=1 to h=5 which becomes evident by the increased in MSEs and biases. The WBY estimates are also affected. There is some loss in precision (increased MSE) for the estimators based on weighting, certainly as the number of predictor increases. As a referee pointed out, this may be due to instabilities when computing the MCD estimator in higher dimensions. Conversely, the RLR estimator still produces the smallest MSEs and biases. Our results are consistent throughout the simulation experiments.

6. ARTIFICIAL DATA

As in the case in linear regression, model fitting via logistic regression is also sensitive to multicollinearity. Most software packages have diagnostic procedure, like variance Inflation factor (VIF) and tolerance test to identify correlated variables. Nevertheless it is possible for variables to pass these tests and have the program run, but yields output that is clearly nonsense.

As a simple example, we fit logistic regression model to the artificial data with $x_1 \sim N(0,1)$ and the outcome variable was generated by comparing $u \sim U(0,1)$ to the true

probability $\pi(x_1) = e^{x_1}/(1 + e^{x_1})$ and if $u < \pi(x_1)$ then $y = 1$, otherwise $y = 0$. The correlated variables were generated from x_1 and the constant as follows: $x_2 = x_1 + U(0,0.01)$ and $x_3 = 1 + U(0,0.1)$. Thus, x_1 and x_2 are highly correlated and $x_3$is nearly collinear with the constant term. The results of fitting logistic regression model is using R language with setting a seed value is 123 with sample size of $n = 50$. The artificial data are contaminated by four high leverage points allocated at $(x_{1,2}, x_{2,2}, x_{3,3}, x_{4,3}) = (10,11,10,11)$. Tables 5-7 show the results on multicollinearity diagnostic for artificial data followed by the parameter estimates, as displayed in Table 8.

Table 5
Multicollinearity Diagnostic for Uncontaminated Artificial Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	X1	X2	X3
X'X					
1.386	1	0.001	0.236	0.166	0.265
1.006	1.174	0.952	0.014	0.02	0.004
0.885	1.252	0.033	0.245	0.757	0.041
0.723	1.384	0.014	0.506	0.058	0.689
X'WX					
0.310	1	0.004	0.173	0.151	0.345
0.234	1.151	0.944	0.02	0.02	0.006
0.205	1.229	0.052	0.278	0.683	0.014
0.157	1.403	0	0.529	0.146	0.634

Table 6
Multicollinearity Diagnostic for High Correlated Artificial Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	X1	X2	X3
X'X					
2.085	1	0	0	0	0
1.915	1.043	0	0	0	0
3.22E-04	80.471	0.988	0	0	0.997
3.13E-06	815.883	0.012	1	1	0.003
X'WX					
0.448	1	0	0	0	0
0.406	1.051	0	0	0	0
6.93E-05	80.414	0.997	0	0	1
7.00E-07	799.759	0.003	1	1	0

Table 7
Multicollinearity Diagnostic for High Correlated with High Leverage Point for Artificial Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	X1	X2	X3
X'X					
1.780	1	0.123	0.049	0.08	0.128
1.159	1.239	0.078	0.341	0.216	0.056
0.696	1.599	0.012	0.505	0.635	0.060
0.364	2.211	0.787	0.105	0.070	0.756
X'WX					
0.416	1	0.13	0.028	0.083	0.137
0.264	1.254	0.069	0.279	0.321	0.048
0.168	1.575	0.017	0.559	0.529	0.063
0.085	2.211	0.784	0.134	0.068	0.752

Table 8
Parameter Estimation of Artificial Data

	ESTIMATOR			
	ML	LR	WBY	RLR
Uncontamination				
Intercept	0.072 (0.293)	0.065 (0.269)	0.186 (0.312)	0.173 (0.273)
X1	0.538 (0.346)	0.476 (0.306)	0.417 (0.332)	0.381 (0.310)
X2	-0.306 (0.365)	-0.268 (0.319)	-0.314 (0.437)	-0.284 (0.326)
X3	-0.133 (0.282)	-0.106 (0.258)	-0.016 (0.330)	-0.004 (0.260)
Multicollinearity				
Intercept	-18.345 (12.022)	0.004 (0.018)	-18.611 (11.681)	0.004 (0.018)
X1	57.226 (130.215)	0.024 (0.016)	58.054 (144.113)	0.024 (0.016)
X2	-56.772 (130.228)	0.024 (0.016)	-57.593 (144.051)	0.024 (0.016)
X3	17.800 (11.426)	0.005 (0.019)	18.057 (11.202)	0.005 (0.019)
Multicollinearity and High Leverage Points				
Intercept	0.148 (0.376)	0.125 (0.335)	-15.357 (12.176)	0.004 (0.018)
X1	0.531 (0.353)	0.482 (0.320)	60.057 (148.500)	0.021 (0.016)
X2	0.018 (0.134)	0.024 (0.130)	-59.641 (148.407)	0.021 (0.016)
X3	- 0.062 (0.160)	-0.053 (0.150)	14.996 (11.737)	0.005 (0.018)

Analyses of Artificial Data

In this section, we further discuss on the drawback of ML estimates for multicollinearity and the LR estimates for multicollinearity in the presence of high leverage points. Referring to Table 5 for uncontaminated artificial data, the condition number of $\kappa_x = 1.384$ and $\kappa_W = 1.403$ and ratio $r_{WX} = 1.013$ indicating no collinearity problems in both information matrix. Regarding to collinearity diagnostic by Lesaffree and Marx (1993), the condition indices , $CI < 30$ are not problematic.

Meanwhile, for correlated artificial data, the condition number of $\kappa_x = 815.883$ and $\kappa_W = 799.759$ are massive with very small eigenvalues show that there is serious multicollinearity occur in the data. The variance decomposition proportion in Table 6 clearly confirms that x_1 and x_2 are severely correlated, while mild correlation can be found in between x_3 and the intercept.

In the simulation results, we mentioned that the high leverage point potentially masked the effect of multicollinearity. We observe similar findings in artificial data as we contaminated multicollinear artificial data with two high leverage points in x_2 and another two in x_3. The condition number of κ_x and $\kappa_W$are drastically reduce from 815.883 to 2.211 and 799.759 to 2.211. One may think that there is no multicollinearity exist in this data by looking at results of the variance decomposition of proportion, as shown in Table 7. In the parameter estimate, we show that the diagnostic result in Table 7 mislead our conclusion about the existence of multicollinearity in the presence of high leverage point.

We do not include the complete result of RLGD in identifying the high leverage points due to space constraint. The RMD-MCD identified cases (1,2,3,4) as suspected high leverage points with RMD-MCD values (4523.668, 4810.753, 331.138, 368.411) and these cases are finally confirmed as high leverage points with$p_{ii}^{*(-D)}$ values (1986812.449, 2246852.671, 10926.407, 13517.088).

Table 8 presents parameter estimates for all the estimators. Under free contamination, the ML estimator should be referred as the best estimator. All the other estimates are fairly closer to the ML estimates.

Then, the model includes the highly correlated variables $x_1$and x_2 and mild correlated variable x_3 with the constant 1. The ML fails to provide a good estimate when multicollinearity exists in the artificial data. The WBY also not exempted from this problem. Both variables $x_1$and $x_2$have very large estimated slope coefficients and estimated standard errors. For variable x_3 and the intercept, we see that the estimated coefficients are of reasonable magnitude but the estimated standard errors are much larger than we would expect. Moreover, the multicollinearity actually switches sign for x_3 and the intercept. Under this type of contamination, one could refer to the LR as the best estimator. The RLR performs equally good as the LR.

Multicollinearity pattern changes when the high leverage points are plugged to correlated variables. Under both contaminations, the best estimator is the one that gives closer estimates to the LR in multicollinear data. It can be observe that the RLR estimates are the closest to the LR estimates in multicollinear data. Moreover, the standard errors of the RLR estimates are the smallest compared to other estimates. There are sharp declines of the ML estimates in both contaminations. In fact, there are sign different for intercept, x_2 and x_3 compared to a single contamination (multicollinearity). There are not many changes in the WBY estimates, but we could not rely on its estimate. The LR estimate also affected in both contaminations as its provide slightly bigger coefficient estimated and standard error with sign different for x_3.

7. REAL EXAMPLE

Our real data is cancer remission which is taken to illustrate severe multicollinearity in logistic regression (Lesaffre and Marx, 1993). The cancer remission data is a benchmark data with severe multicollinearity problem. The continuous risk factors associated with cancer remission are cell index, temperature, and li index. The binary response is 1 if the patient experiences a complete cancer remission and 0 otherwise. There were 27 patients involved and 9 of which experienced a complete cancer remission. Three extreme high leverage points are plugged on a temperature with $(x_{25,3}, x_{26,3}) = (10,11)$.

Tables 9-10 show the results on multicollinearity diagnostic for cancer remission data, followed by the parameter estimates, as displayed in Table 11.

Table 9
Multicollinearity Diagnostic for High Correlated Cancer Remission Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	LI	TEMP	CELL
X'X					
3.843	1	0	0.010	0	0.003
0.129	5.448	0	0.979	0	0.020
0.028	11.799	0.001	0.003	0.001	0.969
1.06E-04	190.776	0.999	0.008	0.999	0.008
X'WX					
0.576	1	0	0.005	0	0
0.015	6.165	0	0.454	0	0.007
5.52E-04	32.287	0.005	0.097	0.003	0.816
5.29E-06	329.954	0.995	0.444	0.997	0.176

Table 10
Multicollinearity Diagnostic for High Correlated with High Leverage Points Cancer Remission Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	LI	TEMP	CELL
X'X					
3.293	1	0.003	0.014	0.031	0.003
0.572	2.400	0.004	0.012	0.952	0.003
0.115	5.346	0.049	0.970	0.002	0.057
2.06E-02	12.652	0.943	0.004	0.016	0.936
X'WX					
0.419	1	0.001	0.015	0	0.001
0.013	5.644	0.009	0.897	0.001	0.022
1.07E-03	19.785	0.005	0.006	0.861	0.158
5.89E-04	26.668	0.985	0.082	0.137	0.819

Table 11
Parameter Estimation of Cancer Remission Data

	ESTIMATOR			
	ML	**LR**	**WBY**	**RLR**
Multicollinearity				
Intercept	67.634 (56.888)	-2.99E-03 (6.76E-03)	66.745 (73.569)	-3.06E-03 (6.87E-03)
LI	3.867 (1.778)	1.34E-03 (8.43E-03)	3.818 (2.238)	1.34E-03 (8.56E-03)
TEMP	-82.074 (61.712)	-2.94E-03 (6.72E-03)	-80.967 (81.163)	-3.02E-03 (6.82E-03)
CELL	9.652 (7.751)	-2.71E-03 (6.44E-03)	9.494 (6.497)	-2.77E-03 (6.54E-03)
Multicollinearity and High Leverage Points				
Intercept	-8.752 (6.054)	-7.07E-02 (4.90E-02)	51.425 (82.316)	-6.593E-03 (1.00E-02)
LI	2.862 (1.298)	-1.07E-02 (6.08E-02)	3.631 (2.331)	4.800E-04 (1.253E-02)
TEMP	0.716 (1.963)	-6.22E-02 (5.22E-02)	-62.963 (89.662)	-6.501E-03 (9.986E-03)
CELL	4.408 (5.724)	-6.26E-02 (4.62E-02)	6.977 (6.705)	-6.039-03 (9.483E-03)

Analyses of Real Data

RLGD identified two high leverage points in the original data. Cases 14 and 19 have RMD-MCD values (7.108, 4.217) while$p_{ii}^{*(-D)}$ values are given as (37.851, 16.399).Meanwhile, the modified data identified cases (14, 19, 25, 26) as high leverage points. Their RMD-MCD values and $p_{ii}^{*(-D)}$ are computed as (6.902, 4.079, 574.714, 638.440) and (36.496, 15.629, 303918.591, 376423.356).

The condition number of $\kappa_x = 190.776$ and$\kappa_W = 329.954$ with ratio $r_{wx} = 1.73$ determined the ill-conditioning in matrix X and information matrix of ML (see Table 9). The variance decomposition proportion table shows high correlation between temperature variable and the intercept term. In the presence of high leverage points (see Table 10), the condition numbers reduce to $\kappa_x = 12.652$ and $\kappa_W = 26.668$. It is quite difficult to judge as to which variables are correlated from the variance decomposition proportion table.

The LR is always expected to give the best estimates in multicollinear data. Referring to Table 11, the RLR estimates are fairly closer to the LR estimates. On the other hand, the ML and the WBY fail to provide good estimates as they have larger values for both estimated coefficients and standard errors, while the estimated coefficient of Intercept and cell variable change sign.

A good estimator for multicollinear high leverage points is the one that has smallest standard errors and estimated coefficients which are closest to estimated coefficient for the LR in multicollinear data. As to be expected, the RLR outperforms other estimators in both contamination scenarios. Even though the RLR standard errors are slightly larger compared to the LR standard errors in multicollinearity, the RLR estimated coefficients are not strayed too far and or changed sign. The WBY estimated coefficients and standard errors do not changes much from its previous estimated coefficients and standard errors in multicollinear data, but they give faulty inference. Meanwhile, there are reductions on the ML standard errors and sign different for estimated coefficients of temperature and cell variables. The LR also affected by the presence of high leverage

points in correlated data, evident by having larger standard errors and different sign for estimated coefficient of li variable compared to the RLR standard errors and estimated coefficients in multicollinearity.

8. CONCLUSION

Many circumstances in logistic regression model encountered a problem of having a severe multicollinearity and high leverage points. The proposed RLR technique offers substantial improvement over the ML, the WBY, and the LR estimation methods for the combined problems of multicollinearity and high leverage points. The findings obtained from the simulation study, artificial data and real example indicate that the RLR method is the overall best-performing technique. The fully iterated of the robust WBY estimator which protects against huge high leverage points in the RLR for severe multicollinearity is really pays-off. Using a particularly challenging data with high correlation, the RLR properly shrinks the parameter coefficients using ridge parameter, reduce the standard errors and uses its robust capability to correctly identify high leverage points.

As a general remark, we emphasize on the advantages of each estimator. In free contamination, the ML is the best estimator, while we suggest using the LR when multicollinearity exists in a data. The WBY estimator is designed to tackle the issue of the high leverage points by assigning a proper weight to each observation. Meanwhile, the RLR estimator provides reliable estimates for multicollinear-high leverage data. As we dealing with the logistic regression, all estimators have convergence issue due to a higher degree of correlations ($\rho \geq 0.975$), a number of high leverage points ($h \geq 10$), a number of predictor variables ($p \geq 5$), a number of sample sizes ($n \leq 100$) and a number of overlapped cases in ($Y = 0$) and ($Y = 1$).

REFERENCES

1. Al-Aabdi, F.A.A. and Al–Shaibani, R.M.A. (2014). Robust estimators of logistic regression with problems multicollinearity or outliers values. *Journal of Kufa for Mathematics and Computer*, 2(2), 64-71.
2. Askin, R.G. and Montgomery, D.C. (1980). Augmented robust estimators. *Technometrics*, 22(3), 333-341.
3. Bagheri, A. and Midi, H. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions*, 39(1), 51-70.
4. Bagheri, A., Midi, H. and Imon, A.H.M.R. (2010). The effect of collinearity-influential observations on collinear data set: A monte carlo simulation study. *Journal of Applied Sciences*, 10(18), 2086-2093.
5. Bagheri, A. and Midi, H. (2012). On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations, *Mathematical Problems in Engineering*, vol. 2012, Article ID 531607, 16 pages. doi: 10.1155/2012/531607.
6. Barker, L. and Brown, C. (2001). Logistic regression when binary variables are highly correlated. *Statistics in Medicine*, 20(9-10), 1431-1442.
7. Carroll, R.J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of Royal Statistics Society B*, 55(3), 693-706.

8. Christmann, A. and Rousseeuw, P.J. (2001). Measuring overlap in binary regression. *Computational Statistics and Data Analysis*, 37(1), 65-75.
9. Collet, D. and Jemain, A.A. (1985). Residuals outliers and influential observations in regression analysis. *Sains Malaysiana*, 14, 493-511.
10. Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*. 44(1-2), 273-295.
11. Cule, E. and De Iorio, M. (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. *The Annals of Applied Statistics*. 1(2), 302-332.
12. Duzan, H. and Shariff, N.S.B.M. (2015). Ridge regression for solving the multicollinearity problem: Review of methods and models. *Journal of Applied Sciences*, 15(3), 392.
13. Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*, Publication 5, University Institute of Economics, Oslo, 5-8.
14. *Godínez*-Jaimes, F., Ramírez-Valverde, G., Reyes-Carreto, R., Ariza-Hernandez, F.J., Barrera-Rodriguez, E. (2012). Collinearity and separated data in the Logistic regression model. *Journal Agrociencia*, 46(4), 411-425.
15. Gunst, R.F. (1983). Regression analysis with multicollinear predictor variables: Definition, detection and effects. *Communications in Statistics*, 12(19), 2217-2260.
16. Habshah, M., Syaiba, B.A. (2012). The performance of classical and robust logistic estimators in the presence of outliers. *Pertanika Journal of Science and Technology*, 20(2), 313-325.
17. Hoerl, A.E and Kennard, R.W. (1970). Ridge Regression: Applications to non-orthogonal problems. *Technometrics*, 12(1), 69-82.
18. Holland, P.W. (1973). Weighted ridge regression: Combining ridge and robust regression methods, *NBER Working* paper series, 11, 1-19.
19. Kibria, B.M.G., Månsson, K. and Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics*, 40(4), 401-414.
20. Kibria, B.M.G. and Salleh, A.K.E. (2012). Improved the estimators of the parameters of a probit regression model: A ridge regression approach. *Journal of Statistical Planning and Inference*, 142, 1421-1435.
21. Lawrence, K.D. and Arthur, J.L. (Eds.). (1990). Robust regression: analysis and applications. New York: Marcel Dekker.
22. Lesaffre, E. and Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods*, 22(7), 1933-1952.
23. Locking, H., Månsson, K. and Shukur, G. (2013). Performance of some ridge parameters for probit regression: with application on Swedish job search data. *Communications in Statistics Simulation and Computation*, 42(3), 698-710.
24. Månsson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. *Communications in Statistics – Theory and Methods*, 40(18), 3366-3381.
25. Marx, B.D. and Smith, E.P. (1990). Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(6), 1128-1135.
26. Park, M.Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1), 30-50.

27. Park, H. and Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*, 86(7), 1450-1461.
28. Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373-1379.
29. Piegorsch, W.W. (1992). Complementary log regression for generalized linear models. *The American Statistician*, 46(2), 94-99.
30. Rousseeuw, P.J and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223.
31. Rousseeuw, P.J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, 43(3), 315-332.
32. Schaefer, R.L., Roi, L.D. and Wolfe, R.A. (1984). A ridge logistic estimator. *Communications in Statistics – Theory and Methods*, 13(1), 99-113.
33. Schaefer, R.L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1-2), 75-91.
34. Shahmandi, M, Farmanesh, F. Gharahbeigi, M.M. and Shahmandi, L. (2013). Data analyzing by attention to weighted multicollinearity in logistic regression applicable in industrial data. *British Journal of Applied Science and Technology*, 3(4), 748-763.
35. Segerstedt, B. and Nyquist, H. (1992). On the conditioning problem in generalized linear models. *Journal of Applied Statistics*, 19(4), 513-526.
36. 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), 3042-3050.
37. Vago, E. and Kemeny, S. (2006). Logistic ridge regression for clinical data analysis (a case study). *Applied Ecology and Environmental Research*, 4(2), 171-179.
38. Victoria-Feser, M-P. (2002). Robust inference with binary data. *Psychometrika*, 67, 21-32.
39. Weissfeld, L.A. and Sereika, S.M. (1991). A multicollinearity diagnostic for generalized linear model. *Communications in Statistics – Theory and Methods*, 20(4), 1183-1198.