Comparison of Unconditional Asymptotic, Exact Conditional and Robust Logistic Regression Methods for Binary Contaminated Data Sets a Simulation Study

İkili Bozulmuş Veri Yapılarında Genel, Sağlam ve Kesin Lojistik Regresyon Yöntemlerinin Karşılaştırılması: Bir Simülasyon Çalışması

Muzaffer Bilgin, Ertugrul Colak
Department of Biostatistics, School of Medicine, Eskisehir, Turkey

Abstract: In clinical research outliers occurs in spite of careful study design, and implementation of error-prevention techniques. Exact conditional and Robust logistic regression techniques are alternatives to the unconditional asymptotic logistic regression analysis when the dataset is contaminated with outliers. Our specific objectives were to compare the performance of exact conditional and robust logistic regression methods by Monte-Carlo simulation study when the data is skewed with outliers. Robust logistic regression method had less biased parameter estimates even at the 1% contamination level. We proposed using robust logistic regression method rather than exact conditional method when the data set is contaminated.
Keywords: contaminated dataset, logistic regression, robust logistic regression, exact logistic regression

Bilgin M, Colak E. 2018, Comparison of Unconditional Asymptotic, Exact Conditional and Robust Logistic Regression Methods for Binary Contaminated Data Sets a Simulation Study, *Osmangazi Journal of Medicine*, 40 (2):53-59, **Doi:** 10.20515/otd.409043

Özet: Klinik araştırmalarda, ölçüm hatasını yok etmek için dikkatli çalışma düzen tasarımları kullanılmasına rağmen aykırı değerler ile karşılaşılabilir. Veri setlerinde aykırı değerler bulunduğu durumda, koşullu olmayan asimptotik lojistik regresyon yöntemlerine alternatif olarak kesin koşullu ve sağlam lojistik regresyon yöntemleri kullanılmaktadır. Bu çalışmanın temel amacı, veri setinde aykırı değerler ile çarpık bir yapı olduğu durumda, kesin ve sağlam lojistik regresyon yöntemlerinin performanslarını Monte-Carlo simülasyon çalışmaları ile karşılaştırmaktır. Sağlam lojistik regresyon yöntemi, % 1 kontaminasyon seviyesinde bile daha az yanlı parametre tahminleri yaptığı bulundu. Veri setlerinde aykırı değerler ile bozulma durumu söz konusu olduğunda sağlam lojistik regresyon yöntemi kullanılmasını öneriyoruz.
Anahtar Kelimeler: bozulmuş ikili veri, lojistik regresyon, sağlam lojistik regresyon, kesin lojistik regresyon

Bilgin M, Çolak E. 2018, İkili Bozulmuş Veri Yapılarında Genel, Sağlam ve Kesin Lojistik Regresyon Yöntemlerinin Karşılaştırılması: Bir Simülasyon Çalışması, *Osmangazi Tıp Dergisi,* **Doi:** 10.20515/otd.409043

Geliş Tarihi / Received
23.03.2018

Kabul Tarihi / Accepted:
12.04.2018

Yazışma Adresi / Correspondence Address
Muzaffer BİLGİN
mail: mbilgin@ogu.edu.tr

1. Introduction

Many medical studies deal with the binary outcomes such as presence or absence of a disease, dead or alive of an individual, etc. (1, 2). Logistic regression is a form of statistical technique that is often appropriate for categorical outcome variables. It defines the relationship between a categorical response variable and a set of explanatory variables (3, 4). Statistical inference for such modeling requires large-sample approximations, and fitting logistic regression models to such data is performed through the unconditional likelihood function. However, unconditional asymptotic logistic regression (UALR) method may be inadequate when the sample size is small or the dataset is sparse, skewed and especially heavily contaminated with outliers. An outlier is an observation whose response dependent-variable value is unusual given its value on the explanatory variables (5). In clinical and epidemiological research outliers occurs in spite of careful study design, and implementation of error-prevention techniques. In addition, sometimes an outlier observed because of the biological variation (6).

In such situations, statistical inference based on exact conditional may be more reliable (7, 8). Exact conditional logistic regression (ECLR) analysis was theorically described by Cox and the computational method is developed with efficient algorithm for generating the required conditional distributions by Hirji, Mehta, and Patel (9, 10).

Robust logistic regression (RLR) is the logical alternative to UALR and ECLR when the dataset contaminated with outliers (11). RLR was originally laid out by Bianco and Yohai (1996). Croux and Haesbroeck (2003) developed RLR analysis, having faster and more stable algorithm.

In this study, our specific objectives were to compare the performance of three individual logistic regression methods by Monte-Carlo simulation study. We generated simulated random datasets of various sample sizes and contamination levels. For comparison, we used the estimates of the regression coefficients, the standard errors, and the bias of these estimates.

2. Materials and Methods

Let y denote a random variable representing a discrete dichotomous response variable. The vector of observed measurements of the response variable for the sample size n is $\boldsymbol{y}' = (y_1, \dots, y_n)$. The vector of explanatory variables for the i[th] individual is $\boldsymbol{x_i}' = (1, x_{1i}, \dots, x_{pi})$ and the matrix of explanatory variables is $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})'$ where p is the number of explanatory variables,$i = 1, \dots, n$. Let $\pi(\boldsymbol{x_i}) = \mathrm{P}(Y_i = 1|\boldsymbol{x_i})$ be event probability for the i[th] individual.

The specific form of the logistic regression form is described as follows.

$$P(Y_i = 1|\boldsymbol{x_i}) = \frac{e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_i x_{pi}}}{1 + e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_i x_{pi}}}$$

$$= F(\boldsymbol{x_i'\beta})$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ coefficient vector of explanatory variables.

2.1. Unconditional Asymptotic Logistic Regression

The most widely used logistic regression method in clinical researches is UALR. Estimation of the unknown parameters β in the model is made by using unconditional likelihood and log-likelihood functions described as follows, respectively (12).

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} F(\boldsymbol{x_i'\beta})^{y_i} \, [1 - F(\boldsymbol{x_i'\beta})]^{1-y_i}$$

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \, ln[F(\boldsymbol{x_i'\beta})]$$

$$+(1 - y_i) \, ln[1 - F(\boldsymbol{x_i'\beta})]\}$$

The estimates of the coefficients $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p]$ are obtained by solving the following derivatives according to β's.

$$\frac{\partial \ln[L(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} x_{ji}[y_i - F(\boldsymbol{x_i'\beta})] = 0$$

$$j = 1, \dots, p$$

$$\widehat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}}\{L(\boldsymbol{\beta})\} = \min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n} d(\boldsymbol{x_i'\beta}; y_i)\right\}$$

where

$$d(\boldsymbol{x_i'\beta}; y_i) = -y_i\, ln[F(\boldsymbol{x_i'\beta})] - (1 - y_i)\, ln[1 - F(\boldsymbol{x_i'\beta})]$$

Statistical inference for the coefficient estimates is based on maximizing this unconditional likelihood function, and several asymptotic statistics such as likelihood ratio, score, and Wald can be used to perform hypothesis tests (7).

2.2.Exact Conditional Logistic Regression

The conditional likelihood function is used to estimate the parameter estimation in exact logistic regression. However, to perform conditional inference, the sufficient statistics fort he β_j in the unconditional likelihood function are calculated as $T_j = \sum_{i=1}^{n} y_i x_{ji}$. Then, the vector of sufficient statistics for all coefficients is $\boldsymbol{T} = (T_0, T_1, \dots, T_P)'$. The probability density function for $\boldsymbol{T}$ is obtained as follows by summing over all binary sequences $\boldsymbol{y}$ (0 for nonevent, 1 for event) that generate observable t.

$$P(\boldsymbol{T} = \boldsymbol{t}) = \frac{C(\boldsymbol{t}) \exp(\boldsymbol{t'\beta})}{\prod_{i=1}^{n}[1 + \exp(\boldsymbol{x_i'\beta})]}$$

where $C(\boldsymbol{t}) = \|\{\boldsymbol{y} : \boldsymbol{y'X} = \boldsymbol{t}\}\|$ is the number of the sequences y that generate t. To obtain conditional likelihood function β_0 is accepted as a nuisance parameter. The corresponding sufficient statistic for the nuisance parameter is T_0. The conditional likelihood function can be described as follows by removing the nuisance parameter from the analysis and conditioning on its sufficient statistics.

$$P(\boldsymbol{T}_p = \boldsymbol{t}_p | \boldsymbol{T}_0 = \boldsymbol{t}_0) = \frac{P(\boldsymbol{T} = \boldsymbol{t})}{P(\boldsymbol{T}_0 = \boldsymbol{t}_0)} = \frac{C(\boldsymbol{t}) \exp(\boldsymbol{t}_p'\boldsymbol{\beta}_p)}{\sum_u C(\boldsymbol{u}, \boldsymbol{t}_0) \exp(\boldsymbol{u'\beta}_p)} = f_{\boldsymbol{\beta}_p}(\boldsymbol{t}_p | \boldsymbol{t}_0)$$

where $C(\boldsymbol{u}, \boldsymbol{t}_0)$ is the number of vectors y such that $\boldsymbol{y'X}_p = \boldsymbol{u}$ and $\boldsymbol{y'X}_0 = \boldsymbol{t}_0$.

The conditional exact inference for the parameters of interest can be made by generating conditional distribution which is called permutation or exact conditional distribution. The exact *p*-values for the hypothesis test about the parameter β_j can be obtained as follows (7, 10, 13, 14).

$H_0: \beta_j = 0$ vs. $H_A: \beta_j < 0$ the p value is $P_L(t_j; 0) = \sum_{u \le t_j} f_0(u|t_0)$

$H_0: \beta_j = 0$ vs. $H_A: \beta_j > 0$ the p value is $P_G(t_j; 0) = \sum_{u \ge t_j} f_0(u|t_0)$

$H_0: \beta_j = 0$ vs. $H_A: \beta_j \neq 0$ the p value is $P(t_j; 0) = 2\, min\{P_L(t_j; 0), P_G(t_j; 0)\}$

2.3.Robust Logistic Regression

Croux and Haesbroeck modified the RLR method that was proposed by Bianco and Yohai and developed an algorithm providing fast and stable estimation results compared to other RLR methods (13). The robust estimates of the coefficients obtained by solving the following equation.

$$\hat{\beta} = \min_{\beta} \sum_{\mathrm{i}=1}^{\mathrm{n}} \{\mathrm{H}(\mathrm{d}(\boldsymbol{x_i'\beta};\, y_i)) + Q(\boldsymbol{x_i'\beta})\}$$

with the Q function is defined as the bias-correction term given by

$$\mathrm{Q}(\boldsymbol{x_i'\beta}) = \mathrm{G}(\mathrm{F}(\boldsymbol{x_i'\beta})) + \mathrm{G}(1 - \mathrm{F}(\boldsymbol{x_i'\beta})) - \mathrm{G}(1)$$

H and G functions are described as follows, respectively.

$$H(t) = \begin{cases} t\, e^{-\sqrt{k}} & if\ t \le k \\ -2\, e^{-\sqrt{t}}\left(1+\sqrt{t}\right) + e^{-\sqrt{k}}\left(2\left(1+\sqrt{k}\right)+k\right) & otherwise \end{cases}$$

$$G(t) = \begin{cases} te^{-\sqrt{-\ln(t)}} + e^{1/4}\sqrt{\pi}\ \phi\left(\sqrt{2}\left(1/2+\sqrt{-\ln(t)}\right)\right) - e^{-1/4}\sqrt{\pi} & if\ t \le e^{-k} \\ e^{-\sqrt{k}}t - e^{-1/4}\sqrt{\pi}\ \phi\left(\sqrt{2}\left(1/2+\sqrt{k}\right)\right) & otherwise \end{cases}$$

where Φ is the normal cumulative distribution function and k is constant. The constant k should be determined to achieve a compromise between robustness and efficiency. It was suggested that using $d = 0.5$.

2.4. Simulation Algorithm

In simulation study, we used the following logistic regression model involving one explanatory variable

$$Logit[\pi(x_i)] = \beta_0 + \beta_1 x_i$$

where β_0, β_1 are the model parameters, and e_i is the random effect (error) having logistic distribution with 0 location and 1 scale parameters.

To compare the performance of the methods, the following simulation steps were applied.

1. We setup parameter values as $\beta_0 = 0$ and $\beta_1 = 2$.
2. Sample size was described as $n = n_1 + n_2$ and $n_2 = c * n$ where c is the contamination percent to obtain contaminated data set.
3. The explanatory variable x_i with n_1 size was generated from standard normal distribution and n_2 size was generated from uniform distribution with minimum and maximum value 4.5 and 5.5, respectively.
4. The random effect e_i with n size was generated from logistic distribution with 0 location and 1 scale parameters.
5. The binary response variable y with n_1 size was generated as

$$y_i = \begin{cases} 0 & if\ \ \beta_0 + \beta_1 x_i + e_i \le 0 \\ 1 & if\ \ \beta_0 + \beta_1 x_i + e_i > 0 \end{cases} \quad i = 1, \dots, n_1$$

and the contamination case with n_2 size was generated as adding misclassified observations on a hyperplane parallel to the true discriminating hyperplane $y = \beta_0 + \beta_1 x_i$, the shift between the two hyperplanes being equal to $3 \times \sqrt{p}$, where p is the number of parameters (β_0, β_1) (15).

6. UALR, ECLR, and RLR methods were performed by using the explanatory variable achieved in step 3 and the binary response variable obtained in step 5. Then the coefficient estimates and their standard errors for $\beta_0 = 0$ and $\beta_1 = 2$ were obtained from each model.

7. The first six steps were independently replicated 10000 times. Thus, 1000 different parameter estimates and their standard errors were obtained for each method. Then, the means of the 10000 different parameter estimates and standard errors calculated.

The three logistic regression methods were compared by evaluating how close the means of the parameter estimates were to the values determined for β_0 and β_1. The various sample sizes and contamination percent were determined to be n =100, 200, 300, 400 and 500; c =0%, 1%, 2%, 3%, 4% and 5% respectively.

3. Results

The parameter estimates, standard errors, and their biases obtained from the simulation study were displayed in Table 1 according to varying sample sizes and contamination percent. With homogeneous data sets where the contamination ratio c =0%, the three logistic regression methods provided parameter estimates with negligible bias for all sample sizes. In addition, the biases of the estimates are getting smaller when the sample

size increases. However, The UALR and ECLR methods provided more biased estimates than the RLR method with the contaminated data set where the contamination percent $c > 0\%$.

The standard errors of the estimates for β_0 were found higher in the UALR method than in the RLR method and do not show any significant change according to the contamination percent. However, the standard errors of the estimates for β_1 show differences in both the UALR and RLR methods. The standard errors obtained from the UALR is significantly reduce as the contamination percent increase. However, this leads to an inverse result for the standard errors obtained from the RLR, especially for $3\% \leq c \leq 5$.

4. Discussion

It is fact that there is more than one method that can be used for the same purpose in the analysis of data for clinical studies. The choice of the correct method ensures that unbiased, consistent, efficient, and sufficient parameter estimates with minimum variance.

Logistic regression methods are widely used in the analysis of clinical researches involving categorical response variable. Especially, UALR method is conventional logistic regression method that is used to determine the effects of risk factors on the response variable. However, the UALR method provides unreliable parameter estimates for contaminated data sets with outliers (16). On the other hand, uncontaminated data set are not always obtained from the clinical researches despite carefully designed study, error-prevention methods (6).

ECLR and RLR techniques have begun to be used as alternative methods to the UALR analysis when the dataset is small, sparse, skewed and especially heavily contaminated with outliers (7, 8, 13-15, 17-19). Therefore, in this study, the performance of UALR, ECLR, and RLR methods were compared with homogeneous and contaminated data set involving outliers in order to guide the clinical researchers.

As a result of the simulation studies, we found that the contamination rate of the data set is a crucial criterion to select the correct method. However, when there is no contamination in the data set, all the methods yield reliable estimates. In comparing the methods according to the parameter estimates and their biases, all the methods provided biased estimates for contaminated data sets. However, RLR method had less biased parameter estimates even at the 1% contamination level. Croux and Haesbroeck (2003) compared the performance of the RLR and the UALR method at 5% contamination, and showed that the RLR method gives less biased estimates.

When the method was compared according to the standard errors of the parameter estimates, it was observed that the standard errors of the parameter estimate obtained from both the UALR and the RLR methods decreased as sample size increased. The ECLR method does not calculate the standard errors of parameter estimates. This is because, in the determination of the significance of the parameter estimates in the ECLR method, the p value is calculated using conditional distribution instead of the test statistic (20). The standard errors for the β_0 estimate obtained from RLR method were observed to be smaller at each contamination levels. On the other hand, in the UALR method the standard errors for the β_1 estimate were significantly reduced as the contamination rate increased. However, it was observed that the standard errors for β_1 estimate obtained from the RLR method increased significantly as the contamination percent increased. This indicates that the RLR method makes a correction on the standard errors of the parameter estimates by modeling the effect of the contamination. Otherwise, the reduction of standard errors significantly influences the test statistics used in the significance of the parameter estimates (14, 16, 20, 21).

In conclusion, for the analysis of binary data, we proposed using RLR method rather than UALR and ECLR to obtain reliable estimates when the data set involve at least 1% contamination. Thus, it is not necessary to exclude the outliers from the study.

Table 1: *Means of the parameter estimates, biases and mean square error obtained from UALR, RLR and ECLR Methods using various sample size from 10000 simulated data sets*

n	C	Parameters $\beta_0 = 0, \beta_1 = 2$														
		UALR						RLR						ECLR		
		$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	MSE($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	MSE($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	MSE($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	MSE($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	MSE($\hat{\beta}_1$)
100	0 %	0.0017	0.0017	0.07342	2.1081	0.1081	0.21138	-0.0012	-0.0012	0.07720	2.1064	0.1064	0.23124	2.0729	0.0729	0.20922
	1 %	-0.0598	-0.0598	0.05176	1.4107	-0.5893	0.41407	-0.0112	-0.0112	0.07089	1.9807	-0.0193	0.22903	1.3909	-0.6091	0.43164
	2 %	-0.0975	-0.0975	0.04905	0.9807	-1.0193	1.07233	-0.0151	-0.0151	0.06388	1.8173	-0.1827	0.24524	0.9699	-1.0301	1.09070
	3 %	-0.1317	-0.1317	0.05490	0.6871	-1.3129	1.74376	-0.0253	-0.0253	0.05805	1.6505	-0.3495	0.32938	0.6801	-1.3199	1.76021
	4 %	-0.1614	-0.1614	0.06227	0.4853	-1.5147	2.30625	-0.0396	-0.0396	0.05372	1.4588	-0.5412	0.50770	0.4812	-1.5188	2.31774
	5 %	-0.1811	-0.1811	0.07064	0.3516	-1.6484	2.72375	-0.0623	-0.0623	0.05123	1.1084	-0.8916	0.84791	0.3469	-1.6531	2.73979
200	0 %	-0.0016	-0.0016	0.03486	2.0477	0.0477	0.09532	-0.0002	-0.0002	0.03648	2.0519	0.0519	0.10479	2.0359	0.0359	0.09206
	1 %	-0.0572	-0.0572	0.02648	1.3900	-0.6100	0.40389	-0.0096	-0.0096	0.03310	1.9158	0.0842	0.10830	1.3817	-0.6183	0.41147
	2 %	-0.0975	-0.0975	0.03014	0.9737	-1.0263	1.06758	-0.0165	-0.0165	0.03114	1.7635	-0.2365	0.15061	0.9674	-1.0326	1.08137
	3 %	-0.1323	-0.1323	0.03669	0.6847	-1.3153	1.74079	-0.0270	-0.0270	0.02853	1.6029	-0.3971	0.24601	0.6824	-1.3176	1.74516
	4 %	-0.1618	-0.1618	0.04458	0.4859	-1.5141	2.30021	-0.0431	-0.0431	0.02627	1.4198	-0.5802	0.42533	0.4832	-1.5168	2.30626
	5 %	-0.1787	-0.1787	0.05075	0.3509	-1.6491	2.72389	-0.0590	-0.0590	0.02529	1.1923	-0.8077	0.75667	0.3496	-1.6504	2.72737
300	0 %	0.0003	0.0003	0.02315	2.0347	0.0347	0.05838	$-1.786e^{-05}$	$-1.786e^{-05}$	0.02421	2.0370	0.0370	0.06410	2.0276	0.0276	0.05992
	1 %	-0.0560	-0.0560	0.01877	1.3829	-0.6171	0.39813	-0.0088	-0.0088	0.02165	1.8977	-0.1023	0.07063	1.3765	-0.6235	0.40823
	2 %	-0.0975	-0.0975	0.02244	0.9733	-1.0267	1.07016	-0.0162	-0.0162	0.01994	1.7529	-0.2471	0.12344	0.9697	-1.0303	1.07185
	3 %	-0.1335	-0.1335	0.02966	0.6828	-1.3172	1.73683	-0.0289	-0.0289	0.01856	1.5851	-0.4149	0.22533	0.6822	-1.3178	1.74255
	4 %	-0.1598	-0.1598	0.03751	0.4854	-1.5146	2.29815	-0.0414	-0.0414	0.01730	1.4103	-0.5897	0.40735	0.4839	-1.5161	2.30229
	5 %	-0.1811	-0.1811	0.04522	0.3524	-1.6476	2.72067	-0.0596	-0.0596	0.01796	1.1927	-0.8073	0.72494	0.3505	-1.6495	2.72323
400	0 %	0.00084	0.00084	0.01741	2.0251	0.0251	0.04348	0.0012	0.0012	0.01813	2.0257	0.0257	0.04731	2.0177	0.0177	0.04337
	1 %	-0.0563	-0.0563	0.01488	1.3811	-0.6189	0.39822	-0.008	-0.008	0.01617	1.8899	-0.1101	0.05822	1.3774	-0.6226	0.40235
	2 %	-0.0973	-0.0973	0.01911	0.9711	-1.0289	1.06657	-0.0169	-0.0169	0.01472	1.7408	-0.2592	0.11022	0.9688	-1.0312	1.07087
	3 %	-0.1323	-0.1323	0.02696	0.6829	-1.3171	1.73574	-0.0281	-0.0281	0.01392	1.5807	-0.4193	0.21478	0.6839	-1.3161	1.73682
	4 %	-0.1598	-0.1598	0.03460	0.4854	-1.5146	2.29890	-0.0403	-0.0403	0.01335	1.4033	-0.5967	0.40111	0.4845	-1.5155	2.29965
	5 %	-0.1801	-0.1801	0.04257	0.3509	-1.6491	2.72011	-0.0576	-0.0576	0.01437	1.1832	-0.8168	0.71131	0.3511	-1.6489	2.72048
500	0 %	-0.0003	-0.0003	0.01336	2.0164	0.0164	0.03513	-0.00023	-0.00023	0.01397	2.0187	0.0187	0.03814	2.0199	0.0199	0.03437
	1 %	-0.0544	-0.0544	0.01255	1.3803	-0.6197	0.39801	-0.0072	-0.0072	0.01303	1.8852	-0.1148	0.04919	1.3771	-0.6229	0.39976
	2 %	-0.0968	-0.0968	0.01745	0.9701	-1.0299	1.06680	-0.0161	-0.0161	0.01215	1.7368	-0.2632	0.10365	0.9689	-1.0311	1.06925
	3 %	-0.1318	-0.1318	0.02508	0.6827	-1.3173	1.73405	-0.0275	-0.0275	0.01130	1.5734	-0.4266	0.20831	0.6821	-1.3179	1.74046
	4 %	-0.1594	-0.1594	0.03284	0.4847	-1.5153	2.29702	-0.0406	-0.0406	0.01114	1.3978	-0.6022	0.39555	0.4843	-1.5157	2.29963
	5 %	-0.1806	-0.1806	0.04045	0.3518	-1.6482	2.71973	-0.0579	-0.0579	0.01220	1.1860	-0.8140	0.70537	0.3506	-1.6494	2.72176

Limitations

This study has some limitations. The main issue was that the simulations were performed by using large data set. The comparison of the methods can be extended including spare, skewed data set with small sample size.

REFERENCES

1.Ryan TP. Some issues in logistic regression. Commun Stat-Theor M. 2000;29(9-10):2019-32.

2.Rush S. Logistic regression: The standard method of analysis in medical research. Technical Report Mathematics; 2001.

3.Czepiel SA. Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep net/stat/mlelr pdf. 2002.

4.Dominguez-Almendros S, Benitez-Parejo N, Gonzalez-Ramirez AR. Logistic regression models. Allergol Immunopathol (Madr). 2011;39(5):295-305.

5.Chen C, editor Paper 265-27 Robust Regression and Outlier Detection with the ROBUSTREG Procedure. Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference; 2002.

6.Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005;2(10):e267.

7.Derr RE, editor Performing Exact Logistic Regression with the SAS System-Revised 2009. Proceedings of the Twenty-fifth Annual SAS Users Group International Conference; Cary, NC; 2009: Citeseer.

8.Gervini D. Robust adaptive estimators for binary regression models. Journal of Statistical Planning and Inference. 2005;131(2):297-311.

9.Cox DR, Snell EJ. Analysis of binary data: CRC Press; 1989.

10.Hirji KF, Mehta CR, Patel NR. Computing Distributions for Exact Logistic-Regression. Journal of the American Statistical Association. 1987;82(400):1110-7.

11.King EN, Ryan TP. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. Am Stat. 2002;56(3):163-70.

12.Pampel FC. Logistic regression: A primer: Sage; 2000.

13.Croux C, Haesbroeck G. Implementing the Bianco and Yohai estimator for logistic regression. Computational Statistics & Data Analysis. 2003;44(1-2):273-95.

14.Mehta CR, Patel NR. Exact logistic regression: theory and examples. Stat Med. 1995;14(19):2143-60.

15.Bianco AM, Martinez E. Robust testing in the logistic regression model. Computational Statistics & Data Analysis. 2009;53(12):4095-105.

16.Hosmer Jr DW, Lemeshow S. Applied logistic regression: John Wiley & Sons; 2004.

17.Hosseinian S, Morgenthaler S. Robust binary regression. Journal of Statistical Planning and Inference [Internet]. 2011 Apr; 141(4):[1497-509 pp.]. Available from: <Go to ISI>://WOS:000286960500013.

18.Komarek P, Moore AW, editors. Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs. AISTATS; 2003.

19.Kordzakhia N, Mishra GD, Reiersolmoen L. Robust estimation in the logistic regression model. Journal of Statistical Planning and Inference. 2001;98(1-2):211-23.

20.Mehta CR, Patel NR, Senchaudhuri P. Efficient Monte Carlo methods for conditional logistic regression. Journal of the American Statistical Association. 2000;95(449):99-108.

21.Agresti A, Kateri M. Categorical data analysis: Springer; 2011.