Robust estimator using regression depth in logistic regression model

Mie Fujiki
Department of Informatics and Mathematical Science,
Graduate School of Engineering Science,
Osaka University
Toyonaka, Osaka 560-8531, Japan
E-mail: miemasa@sigmath.es.osaka-u.ac.jp

1. Introduction

In this paper, we consider a modified deepest regression estimator (DRE) in logistic regression model. We propose an estimator that takes the median of all candidate fits with maximal regression depth. We compare a modified DRE with the maximum likelihood estimator, Firth's method, and the original DRE in logistic regression model by the computer simulations. We show that a modified DRE is not affected by overlap or complete separation.

Rousseeuw and Hubert (1999) introduced the regression depth method for linear regression models. Regression depth is defined as the smallest number of residuals that need to change sign. DRE is defined as the fit which makes regression depth the maximum relative to the data. DRE is a robust regression estimator and it has high asymptotic efficiency. However, in simple regression, the performance of DRE is not high in case of small sample size with outliers. We showed that mean squared error of a modified DRE using median is smaller than that other estimators in small and large sample size with outliers (Fujiki and Shirahata, 2011).

Moreover, due to the monotone invariance property of regression depth, DRE is invariant to monotone transformations of the response, though this property does not hold for least squares or other estimators such as least trimmed squares or S-estimators. Thereby, it is possible to apply DRE to more general models. In general, the maximum likelihood method is used to estimate regression parameters in logistic regression model. However, a maximum likelihood estimator does not exist in case of complete separation or quasi-complete separation. Firth(1993) suggested the method to remove bias of a maximum likelihood estimator, but this method is not investigated under near separation. Though Ohkura and Kamakura(2007) discussed the method to approximate an estimator using Firth's method to an estimator using the exact logistic regression, DRE is not yet compared with Firth's method in logistic regression model. Therefore, we consider that we apply a modified DRE in logistic regression model, and we also investigate an estimator using Firth's method and our DRE under near separation or overlap by the computer simulations.

2. Regression depth

In multiple regression, we want to fit an affine hyperplane of the form $g((\boldsymbol{x}_i, 1)\boldsymbol{\theta}) = \theta_1 x_{i1} + \cdots + \theta_{p-1} x_{i,p-1} + \theta_p$ to a dataset $Z_n = \{(x_{i1}, \cdots, x_{i,p-1}, y_i) : i = 1, \cdots, n\} \subset \mathbb{R}^p$. We denote the x-part of each data point $\boldsymbol{z}_i$ by $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{i,p-1})^t \in \mathbb{R}^{p-1}$. A candidate fit is denoted by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^t \in \mathbb{R}^p$. The residuals are then denoted by $r_i(\boldsymbol{\theta}) = r_i = y_i - g((\boldsymbol{x}_i, 1)\boldsymbol{\theta})$.

Definition 1 *A candidate fit $\boldsymbol{\theta}$ to Z_n is called a nonfit iff there exists an affine hyperplane V in $\boldsymbol{x}$-space that no $\boldsymbol{x}_i$ belongs to V and such that* (1) or (2).

(1) $r_i(\boldsymbol{\theta}) = y_i - g((\boldsymbol{x}_i, 1)\boldsymbol{\theta}) > 0$ for all $\boldsymbol{x}_{\mathrm{i}}$ in one of its open halfspaces

(2) $r_i(\boldsymbol{\theta}) = y_i - g((\boldsymbol{x}_i, 1)\boldsymbol{\theta}) < 0$ for all $\boldsymbol{x}_{\mathrm{i}}$ in the other open halfspaces.

Definition 2 *The regression depth (rdepth) of a fit* $\boldsymbol{\theta} \in \mathbb{R}^p$ *relative to a dataset* $Z_n \subset \mathbb{R}^p$ *is the smallest number of observations that need to be removed to make* $\boldsymbol{\theta}$ *a nonfit, and is given by*

$$(3) \quad rdepth(\boldsymbol{\theta}, Z_n) = \min_{\boldsymbol{u},v} \{\sharp(r_i(\boldsymbol{\theta}) \geq 0 \ \& \ \boldsymbol{x}_i^t\boldsymbol{u} < v) \ + \ \sharp(r_i(\boldsymbol{\theta}) \leq 0 \ \& \ \boldsymbol{x}_i^t\boldsymbol{u} > v)\}$$

where the minimum is over all unit vectors $\boldsymbol{u} = (u_1, \cdots, u_{p-1})^t \in \mathbb{R}^{p-1}$ *and all* $v \in \mathbb{R}$ *with* $\boldsymbol{x}_i^t\boldsymbol{u} \neq v$ *for all* $(\boldsymbol{x}_i^t, y_i) \in Z_n$.

Christmann and Rousseeuw (2001) introduced the regression depth approach for logistic regression model. Data sets have the form $Z_n \subset \mathbb{R}^p$ where $y_i \in \{0, 1\}$ for $i = 1, \cdots, n$. For simplicity, we will assume that the design matrix has full column rank. Denote the cumulative distribution function fo the logistic distribution by $\Lambda(t) = 1/[1 + \exp(-t)]$, $z \in \mathbb{R}$. The regression depth of a fit invariant with respect to monotone transformations, though this invariance property does not hold for the objective function of most regression estimators, such as least squares, least trimmed squares, and S-estimators. Christmann and Rousseeuw (2001) defined regression depth in logistic regression model from Definition 2, as follows. Hence, the regression depth is invariant with respect to different codings of the binary response variable.

Definition 3 *The regression depth of fit* $\boldsymbol{\theta}$ *relative to* Z_n *is equal to the regression depth of* $-\boldsymbol{\theta}$ *relative to the data set* $Z_n^{'} = \{(x_{i1}, \cdots, x_{i,p-1}, 1 - y_i) : i = 1, \cdots, n\}$,

$$(4) \quad rdepth(\boldsymbol{\theta}, Z_n) = rdepth(-\boldsymbol{\theta}, Z_n^{'}).$$

3. Deepest regression estimator

Definition 4 *In p dimensions the deepest regression estimator* $DR(Z_n)$ *is defined as the fit* $\boldsymbol{\theta}$ *with maximal* $rdepth(\boldsymbol{\theta}, Z_n)$,

$$(5) \quad DR(Z_n) = \arg\max_{\boldsymbol{\theta}} \ rdepth(\boldsymbol{\theta}, Z_n)$$

where $\boldsymbol{\theta}$ *maximizing* $rdepth(\boldsymbol{\theta}, Z_n)$ *is not necessarily unique.*

It suffices to consider all fits through p data points in Definition 4. If several of these fits are tied in the sense that they have the same maximal regression depth, the deepest regression estimator $DR(Z_n)$ is obtained by taking their average. Note that the average does not necessarily have the maximal depth. The DRE is now uniquely defined, taking the average does not change the robustness (Aelst *et al.*, 2002). Moreover, no distributional assumptions are made to define the DRE of dataset.

By the way, $\boldsymbol{\theta}$ maximizing $rdepth(\boldsymbol{\theta}, Z_n)$ is obtained by a straight line passing two data points in simple regression. Then, an actual estimator is defined as follows.

Definition 5 *In a straight line passing two data points of* $\binom{n}{2}$, *the deepest regression estimator* $DR(Z_n)$ *is defined as the maximizing* $rdepth(\boldsymbol{\theta}, Z_n)$ *of the line* $\theta_{1i}x + \theta_{2i} (i = 1, \cdots, k)$, *where* k *is the number of straight lines having maximal regression depth.*

$$DR(Z_n) = (\overline{\theta}_{1\cdot}, \overline{\theta}_{2\cdot}).$$

Note that it is not necessarily for a straight line maximizing $rdepth(\boldsymbol{\theta}, Z_n)$ to pass two data points. For univariate data, regression depth is

$$rdepth(\theta, Z_n) = \min(\sharp\{y_i \leq \theta\}, \ \sharp\{y_i \geq \theta\}),$$

and so DRE is median. Therefore we can define regression depth in two dimensions as follows.

Definition 6 *In a straight line passing two data points of* $\binom{n}{2}$*, the deepest regression estimator* $DR(Z_n)$ *is defined as the maximizing* $rdepth(\boldsymbol{\theta}, Z_n)$ *of the line* $\theta_{1i}x + \theta_{2i}(i = 1, \cdots, k)$*, where* k *is the number of straight lines having maximal regression depth.*

$$DR(Z_n) = (\text{med } \theta_{1\text{i}},\ \text{med } \theta_{2\text{i}}).$$

We investigate the performance of the modified deepest regression estimator by simulation studies. A modified DRE is compared with the original DRE. The results are that mean squared error of our DRE is smaller than that of other estimators using regression depth in small and large sample size with outliers. (Fujiki and Shirahata, 2011)

4. Overlap

The Overlap can be defined by Difinition 1 and Difinition 2. Thus, the separation is defined as follows. (Christmann and Rousseeuw, 2001)

Definition 7 *The data set is completely separated completely if there exists* $\boldsymbol{\theta} \in \mathbb{R}^p$ *such that*

$$(\boldsymbol{x}_i, 1)\boldsymbol{\theta} > 0 \ if \ y_i = 1,$$
$$(\boldsymbol{x}_i, 1)\boldsymbol{\theta} < 0 \ if \ y_i = 0, \quad (i = 1, \cdots, n).$$

A data set is quasi-completely separated if there exists $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \{0\}$ *such that*

$$(\boldsymbol{x}_i, 1)\boldsymbol{\theta} \geq 0 \ if \ y_i = 1,$$
$$(\boldsymbol{x}_i, 1)\boldsymbol{\theta} \leq 0 \ if \ y_i = 0,$$

for all i *and if there exists* $j \in \{1, \cdots, n\}$ *such that* $(\boldsymbol{x}_j, 1)\boldsymbol{\theta} = 0$*.*

Definition 8 *The overlap is the number of observations that need to be removed to obtain complete or quasi-complete separation in binary regression model.*

In other words, Difinition 8 is the minimal number of missclassification in the training data for any linear discriminant function. From Difinition 7 and 8, a data set is said to have overlap if there is no complete separation and no quasi-complete separation. For logistic regression model, the maximum likelihood estimator of $\boldsymbol{\theta}$ does not exist if a data set has no overlap.

Figure 1 is the scatter plot of an artificial data set with

$$\begin{aligned} x_1 &= \{-1.5, -1, 0, 0, 1, 1, 2, 3, 3, 3.5\}, \\ x_2 &= \{0, 3, 1, 2, 2, 4, 2, 1, 3, 4\}, \\ y &= \{0, *, 0, 0, 0, 0, 1, 1, 1, 1\}. \end{aligned}$$

If $*$ is 1 ($y_2 = 1$), then the data sets $\{y_i = 0; i = 1, \cdots, n\}$ and $\{y_i = 1; i = 1, \cdots, n\}$ can not be separated by a hyperplane. In that case, the maximum likelihood estimator exists. However, if $*$ is 0 ($y_2 = 0$), then this data sets can be separated by an appropriate hyperplane. In that case, the maximum likelihood estimator does not exist, due to complete separation.

By the way, some statistical softwares such as SAS, S-PLUS or R execute an iteration method to obtain a maximum likelihood estimate. Furthermore, glm function implemented in R presents the result of the iteration with regard to the maximum likelihood estimate in spite of failing in convergence of the iteration. In this case, a standard error for regression parameter estimate is very large. However, if a data set has no overlap, the deepest regression estimator can be obtained. Figure 1 shows that the red line is the result of linear discriminant analysis, and the blue line is the result of the deepest regression in logistic regression model. Both of two lines yield similar results. Therefore,

we confirm that it is possible to use the deepest regression estimator in logistic regression model for the discriminant analysis, due to complete separation.

Figure1: Scatter plot of an artificial data set

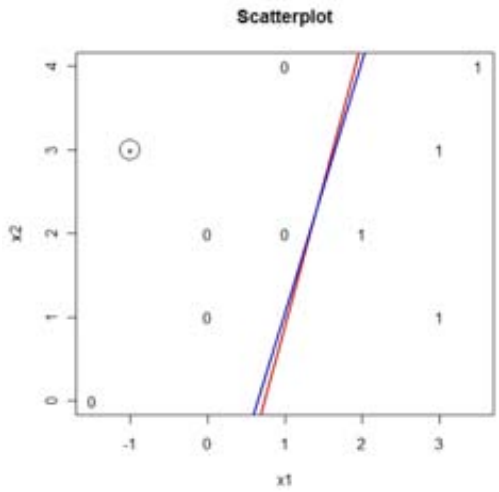

Figure 1: Red line is LDA, Blue line is DRE($*$ is $y_i = 0$ or $y_i = 1$).

Figure2: The result of generating data sets with complete separation or overlap

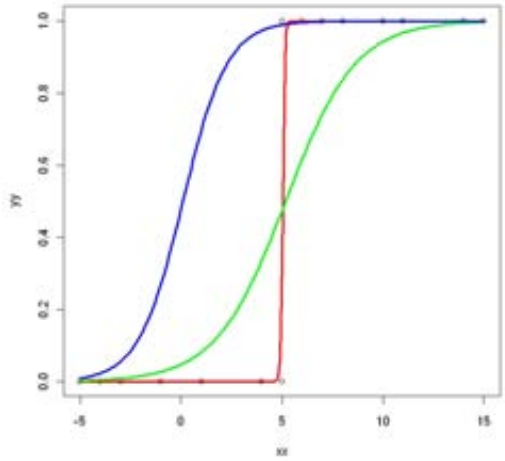

Figure 2: Overlap($k = 0$)

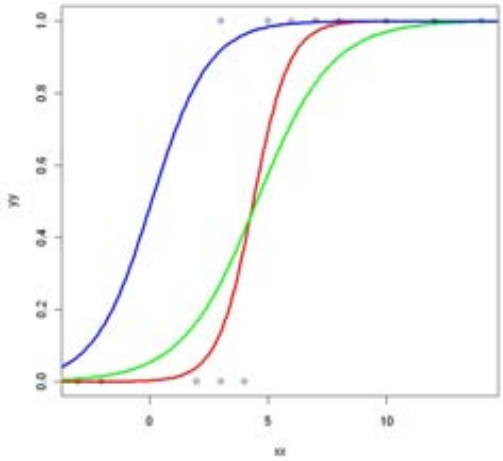

Figure 3: Overlap($k = 1$)

5. Simulation

Firth(1993) suggested a method to eliminate a bias of the maximum likelihood estimator. The regression parameter $\boldsymbol{\theta}$ is estimated by the modified log-likelihood function of Firth's method $\log FL(\boldsymbol{\theta})$

$$\log FL(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) + \frac{1}{2}\log|I(\boldsymbol{\theta})|$$

where $L(\boldsymbol{\theta})$ is the likelihood function of logistic regression model, $I(\boldsymbol{\theta})$ is Fisher's information matrix. As a result, Firth's method can estimate the regression parameter under the complete or the quasi-complete separation. (Heinze and Schemper, 2002)

King and Ryan (2002) examined the performance of the maximum likelihood estimator in the presence of near separation, and compared a maximum likelihood estimator with an estimator using the exact logistic regression. We compare three estimators, which are the maximum likelihood estimator(MLE), an estimator using Firth's method, and the deepest regression estimator(DRE) in logistc regression model. We refer the simulation method of King and Ryan(2002) and Ohkura and Kamakura(2007).

To measure the performance of three estimators(MLE, Firth's method, and DRE), we carried out the following simulation. We generate $n = 10,\ 20,\ 50,\ 100,\ 500$ from the uniform distribution, where k is the number of observations overlapping and $d(\geq k)$ is sample size in the range of overlap, as follows.

$$y_i = \begin{cases} 0 & (-5 \leq x_i \leq\ 5,\ i = 1, \cdots, n/2) \\ 1 & (\ 5 \leq x_i \leq 15,\ i = n/2-1, \cdots, n-d) \end{cases}$$

We decide n, k, and d. We generate $n/2$ from the uniform distribution in $[-5, 5]$. Let this data set for $y_i = 0$. Similarly, we generate $n/2 - d$ from an uniform distribution in $[5, 15]$. Let this data set be $y_i = 1$.

$$x_i = \begin{cases} [5, 15] & (i = n/2, \cdots, n-d) \\ \left[\max\limits_{1\leq j\leq n/2}\{x_j\} - k,\ \max\limits_{1\leq j\leq n/2}\{x_j\} - 1\right] & (i = n-d+1, \cdots, n) \end{cases}$$

$k = 0$ implies complete separation. Figure2 is the example of $k = 0$ and $k = 1$, and shows that red is MLE, green is Firth's method, and blue is DRE.

We want to fit a model $\log\{\pi_i/(1-\pi_i)\} = \theta_1 x_i + \theta_2$ for the generated data set, and we estimate θ_1 of MLE, Firth's method, and DRE.

Table 1 is the average of 100 θ_1 estimates in each estimator. When $k = 0$ is complete separation, as the sample size n increases, the estimate of MLE and Firth's method also increases. In that case, the estimate of DRE does not increase. As the overlap k increases, the estimate of MLE and Firth's method decreases. Similarly, as d increases, the estimate of MLE and Firth's method decreases. However, the estimate of DRE is approximately-constant in the case that complete separation or overlap exists. Therefore, DRE is not affected by complete separation or overlap.

REFERENCES (RÉFERENCES)

Aelst, S.V., Rousseeuw, P.J., Hubert.M., and Struyf.A. (2002). The Deepest Regression Method, *Journal of Multivariate Analysis*,81,138-166.

Christmann, A. and Rousseeuw, P.J. (2001). Measuring overlap in logistic regression, *Computational Statistics and Data Analysis*, 37, 65-75.

Table1: The result of simulation

	Overlap(k)	$k=0$			$k=1$			$k=2$		
	Estimator	MLE	Firth	DRE	MLE	Firth	DRE	MLE	Firth	DRE
$n=10$	$d=0$	19.2953	0.5165	1.0242						
	1				0.9092	0.3844	0.8869			
	2				0.7704	0.3229	0.8181	0.6366	0.3183	0.8394
	3				0.7871	0.3318	0.7305	0.5769	0.2773	0.7471
	4				0.7796	0.3414	0.6831	0.5605	0.2776	0.6791
	5				0.8296	0.4458	0.5178	0.5810	0.3654	0.4623
$n=20$	$d=0$	25.0378	0.8459	1.0169						
	1				1.2782	0.6572	0.9267			
	2				1.1224	0.6002	0.8931	0.8782	0.5460	0.8718
	3				1.0606	0.5751	0.8948	0.8154	0.5093	0.8560
	4				1.0866	0.5901	0.8556	0.7968	0.4980	0.8283
	5				1.0907	0.6216	0.8253	0.8143	0.5216	0.8291
$n=50$	$d=0$	27.3974	1.8543	0.9980						
	1				1.9400	1.2321	0.9591			
	2				1.6556	1.1483	0.9392	1.3071	0.9507	0.9257
	3				1.4695	1.0410	0.9232	1.1635	0.8838	0.9120
	4				1.4515	1.0268	0.9101	1.1328	0.8755	0.9040
	5				1.5227	1.0655	0.9148	1.0204	0.8014	0.8852
$n=100$	$d=0$	23.0449	2.7824	0.9642						
	1				2.6654	1.9557	0.9624			
	2				2.2231	1.7261	0.9570	1.8120	1.4686	0.9737
	3				1.9719	1.5708	0.9501	1.5589	1.2977	0.9492
	4				1.8633	1.4803	0.9269	1.4449	1.2197	0.9330
	5				1.7626	1.4278	0.9356	1.3132	1.1247	0.9176
$n=500$	$d=0$	20.7498	4.6174	0.9883						
	1				4.0108	3.5284	0.9849			
	2				3.3666	3.0748	0.9848	2.9581	2.7377	0.9857
	3				3.0127	2.7926	0.9840	2.6224	2.4584	0.9805
	4				2.7949	2.6174	0.9882	2.3545	2.2256	0.9846
	5				2.6377	2.4841	0.9864	2.1887	2.0778	0.9788

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.

Fujiki, M. and Shirahata, S. (2011). Robust estimator using the data depth in simple regression (In Japanese), *Bulletin of the Computational Statistics of Japan*, 23, 81-96.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression, *Statistics in Medicine*, 21, 2409-2419.

King, E.N. and Ryan, T.P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression, *The American Statistician*, 56, 163-170.

Ohkura, M. and Kamakura, T. (2007). Approximate Estimate for Exact Logistic Regression (In Japanese), *Japanese Journal of Applied Statistics*, 36, 87-98.

Rousseeuw, P.J. and Hubert, M.(1999). Regression Depth, *Journal of the American Statistical Association*, 94, 388-402.