Journal of Advances in Mathematics and Computer Science

32(5): 1-15, 2019; Article no.JAMCS.49483
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Parametric Bootstrapping Predictive Estimator for Logistic Regression

Kunio Takezawa[1*]

[1] *Division of Informatics and Inventory, Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Kannondai 3-1-3, Tsukuba, Ibaraki 305-8604, Japan.*

Author's contribution

The sole author designed, analyzed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v32i530154
Editor(s):
(1) Dr. Serkan Araci, Hasan Kalyoncu University, Turkey.
Reviewers:
(1) Olumide Adesina, Olabisi Onabanjo University, Nigeria.
(2) Siana Halim, Petra Christian University, Indonesia.
Complete Peer review History: http://www.sdiarticle3.com/review-history/49483

Received: 25 March 2019
Accepted: 30 May 2019
Published: 10 June 2019

Original Research Article

Abstract

This paper proposes a method for constructing a predictive estimator for logistic regression. We make a provisional assumption that the predictive estimator is given by multiplying the maximum likelihood estimators by constants, which are estimated using a parametric bootstrap method. The relative merits of the maximum likelihood estimator and the predictive estimator produced by this method are determined by cross-validation. The results show that the predictive estimators derived by this method lead to a smaller deviance than that obtained by the maximum likelihood estimator in many instances.

Keywords: Log-likelihood; future data; predictive estimator; logistic regression; maximum likelihood estimator; parametric bootstrap method.

2010 Mathematics Subject Classification: 60G25; 62F10; 62M20

1 Introduction

The maximum likelihood estimator is a commonly used tool for deriving regression coefficients in regression analysis. One of the reasons is that the maximum likelihood estimator has a number of

**Corresponding author: E-mail: nonpara@gmail.com*

desirable characteristics when the number of data is large (e.g. [1]). It has been found, however, that some estimators are better than the maximum likelihood estimator from the standpoint of fitting well to future data when the number of data is relatively small. Such an estimator is called a 'predictive estimator'. For example, when we use the 'third variance' for estimating the variance of normally distributed data, the resultant estimate tends to fit closely to future data ([2]; Chapter 5 of [3]). In contrast, a predictive estimator that gives better results than the maximum likelihood estimator for estimating the parameter of an exponential distribution has already been obtained ([4]), and a predictive estimator that works better than the maximum likelihood estimator for estimating the coefficients of correlation has also been found ([5]). The characteristics of such predictive estimators have been elucidated using asymptotic methods ([6]) and comparisons with cross-validation ([7]).

Hence, this paper considers a predictive estimator for logistic regression, which is a typical generalized linear regression (e.g. [8]; [9]; [10]; [11]). The maximum likelihood estimator is extensively used for estimating regression coefficients in generalized linear regression for practical purposes. One reason for this is that the estimates obtained by the maximum likelihood estimator are obtained by applying a weighted linear regression with relative ease; the obvious reason is that the maximum likelihood estimator is regarded as a desirable estimator. In simple regression, however, we have already found that a predictive estimator leads to better results than the maximum likelihood estimator with regard to prediction ([12]). Hence, if a predictive estimator for logistic regression were to be constructed, it could possibly be a better estimator than the maximum likelihood estimator from the perspective of prediction. Such considerations should have a substantial impact in the fields in which estimation using the maximum likelihood estimator are conventionally employed.

Therefore, an outline of the predictive estimator is given below. The predictive estimator for the variance of normally distributed data is first derived; this predictive estimator is called the third variance. Next, we suggest an application of the parametric bootstrap method (Section 6.5 of [13]; [14]) to estimate the constants in the predictive estimator for logistic regression. The usefulness of this method is explored using simulation data. "R 3.5.0" was used for these numerical simulations.

2 Basic Concept and Example of a Predictive Estimator using the Third Variance

Maximization of the log-likelihood leads to estimates of parameters using the maximum likelihood method. The log-likelihood is defined as

$$l = \sum_{i=1}^{n} \log\Big(f(\mathbf{x}_i|\boldsymbol{\theta})\Big), \tag{2.1}$$

where $f(\mathbf{x}|\boldsymbol{\theta})$ stands for the available data and $\boldsymbol{\theta}$ represents the parameters. The true values of $\boldsymbol{\theta}$ are denoted as $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}$ which maximizes l is called the maximum likelihood estimator.

In contrast, a maximum likelihood method that considers future data is also possible. This method estimates parameters by maximizing the log-likelihood in the light of future data. This method results in predictive estimators. While the maximum likelihood estimator gives estimates that fit well to the available data, the predictive estimator gives estimates that fit well to future data. Because the purpose of constructing a statistical model is to represent the behaviour of future data with great accuracy, the predictive estimator is a more desirable estimator than the maximum likelihood estimator in terms of the real purpose of statistical modelling.

The log-likelihood in the light of future data is defined as

$$l^* = \sum_{i=1}^{n} \log\Big(f(\mathbf{x}_i^*|\boldsymbol{\theta})\Big), \tag{2.2}$$

where $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*\}$ denotes future data. Future data are sampled from the same population as the one from which the available data were sampled. Hence, future data has the same statistical characteristics as that of available data, However, use of future data removes overfitting of parameters to available data. Because the number of future data is infinite, we consider the expectation of l^* with respect to future data. That is, the average is taken after an infinite number of future data are sampled. Hence, we obtain the equation

$$E_{\{\mathbf{x}_i^*\}}[l^*] = E_{\{\mathbf{x}_i^*\}}\left[\sum_{i=1}^{n} \log\Big(f(\mathbf{x}_i^*|\boldsymbol{\theta})\Big)\right] = n\int \log\Big(f(\mathbf{x}^*|\boldsymbol{\theta})\Big) f(\mathbf{x}^*|\boldsymbol{\theta}_0) d\mathbf{x}^*. \tag{2.3}$$

This is referred to as the expected log-likelihood. The term $\boldsymbol{\theta}$, which increases this value, is regarded as the estimate that fits closely to future data. However, we cannot estimate $\boldsymbol{\theta}$ by maximizing this value because we do not normally have access to future data in regression analysis. Instead, we consider the expectation of l^* with respect to both available data and future data. Then, the expectation given by sampling both available data and future data infinitely is represented as (Eq. (2.3) in [6])

$$\begin{aligned} E_{\{\mathbf{x}_i,\mathbf{x}_i^*\}}[l^*] &= E_{\{\mathbf{x}_i,\mathbf{x}_i^*\}}\left[\sum_{i=1}^{n} \log\Big(f\big(\mathbf{x}_i^*|\tilde{\boldsymbol{\theta}}(\{\mathbf{x}_i\})\big)\Big)\right] \\ &= n\int\cdots\int \log\Big(f\big(\mathbf{x}^*|\tilde{\boldsymbol{\theta}}(\{\mathbf{x}_i\})\big)\Big) f(\mathbf{x}_1|\boldsymbol{\theta}_0)\cdots f(\mathbf{x}_n|\boldsymbol{\theta}_0) f(\mathbf{x}^*|\boldsymbol{\theta}_0) d\mathbf{x}_1\cdots d\mathbf{x}_n d\mathbf{x}^*, \end{aligned} \tag{2.4}$$

where $\tilde{\boldsymbol{\theta}}(\{\mathbf{x}_i\})$ stands for $\boldsymbol{\theta}$, which is estimated by an estimator of some sort (not limited to the maximum likelihood method) using $\{\mathbf{x}_i\}$. The estimator that maximizes the value of Eq. (2.4) is defined as the predictive estimator. When least squares is used for estimation, Eq. (2.4) turns out to be the 'expected test mean squared error' (on page 34 in [15]). The idea of making the value of $E_{\{\mathbf{x}_i,\mathbf{x}_i^*\}}[l^*]$ large appears in the derivation of Akaike's information criterion (e.g. [16], [17], [18] and generalized cross-validation (e.g. [19], [20]). However, whereas Akaike's information criterion and generalized cross-validation are used for model selection, a predictive estimator is used for obtaining estimates.

Next, the predictive estimator for variance of normally distributed data is developed to provide an example of a predictive estimator. The available data is denoted as $\{x_1, x_2, \ldots, x_n\}$, and future data is represented as $\{x_1^*, x_2^*, \ldots, x_n^*\}$. Then, the expectation of the log-likelihood with respect to both the available data and future data is written as

$$E_{\{x_i,x_i^*\}}[l^*] = E_{\{x_i,x_i^*\}}\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{\sum_{i=1}^{n}(x_i^* - \bar{x})^2}{2\tilde{\sigma}^2}\right], \tag{2.5}$$

where $\tilde{\sigma}^2$ is the variance, which is derived using available data. In addition, $\tilde{\sigma}^2$ is set as

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - \alpha}, \tag{2.6}$$

where α is a constant and $\bar{x}$ is the average of $\{x_i\}$. Then, l^* is written as

$$l^* = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - \alpha}\right) - \frac{(n-\alpha)\sum_{i=1}^{n}(x_i^* - \bar{x})^2}{2\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.7}$$

By taking expectation with respect to both the available data and future data, we obtain the following equation.

$$E_{\{x_i,x_i^*\}}[l^*] = -\frac{n}{2}\log(2\pi) - \frac{n}{2}E_{\{x_i\}}\left[\log\left(\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-\alpha}\right)\right] - \frac{n-\alpha}{2}E_{\{x_i,x_i^*\}}\left[\frac{\sum_{i=1}^n (x_i^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] \tag{2.8}$$

To calculate the last term of the right-hand side of this equation, we use the equation

$$\frac{\sum_{i=1}^n (x_i^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\chi_{n+1}^2}{\chi_{n-1}^2} = \frac{n+1}{n-1}F_{(n+1,n-1)}, \tag{2.9}$$

where χ_{n+1}^2 is a random variable that obeys a chi-squared distribution with $(n+1)$ degrees of freedom, χ_{n-1}^2 is a random variable that obeys a chi-squared distribution with $(n-1)$ degrees of freedom, and χ_{n+1}^2 and χ_{n-1}^2 are independent each other. In addition, $F_{(n+1,n-1)}$ is a random variable that obeys an F-distribution with $(n+1, n-1)$ degrees of freedom. As a result, the equation below is obtained.

$$E_{\{x_i,x_i^*\}}\left[\frac{\sum_{i=1}^n (x_i^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] = \frac{n+1}{n-1}E[F_{(n+1,n-1)}] = \frac{n+1}{n-1}\cdot\frac{n-1}{n-3} = \frac{n+1}{n-3} \tag{2.10}$$

Here, $E_{\{x_i,x_i^*\}}[l^*]$ given by Eq. (2.8) is written as

$$E_{\{x_i,x_i^*\}}[l^*] = -\frac{n}{2}\log(2\pi) - \frac{n}{2}E_{\{x_i\}}\left[\log\left(\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-\alpha}\right)\right] - \frac{(n-\alpha)(n+1)}{2(n-3)}. \tag{2.11}$$

The essential part for deriving the optimal value of α is extracted from the equation above. Then, we have

$$g(\alpha) = \frac{n}{2}\log(n-\alpha) + \frac{\alpha(n+1)}{2(n-3)}. \tag{2.12}$$

We differentiate the equation above with respect to α and set it to zero. Then, the optimal value of α turns out to be

$$\hat{\alpha} = \frac{4n}{n+1} \approx 4. \tag{2.13}$$

Substitution of the above result into Eq. (2.6) provides the third variance.

Fortunately, $\hat{\alpha}$ does not depend upon the true value of the parameter in this instance. However, the most predictive estimators depend upon the true values of the parameters; such predictive estimators should be handled with ingenuity.

3 Maximum Likelihood Estimator and Predictive Estimator for Logistic Regression

We assume that data $(\{(x_i, m_i, y_i)\}(1 \le i \le n))$ are available for carrying out logistic regression. Each $\{x_i\}(1 \le i \le n)$ is the value of a predictive variable at the i-th data point, each $\{m_i\}(1 \le i \le n)$ stands for the number of trials at the i-th data point, and each $\{y_i\}$ is the number of success at the i-th data point. x_i and m_i are non-random variables, while y_i is a realization of a random variable Y_i. $\{x_i\}$ and $\{m_i\}$ are fixed before an experiment and $\{y_i\}$ are results of the experiment. Hence, $(m_i - y_i)$ represents the number of failures at the i-th data point.

The regression equation of the logistic equation is written as (page 103 in [9])

$$E[Y_i] = m_i P(x_i) = \frac{m_i}{1+\exp(-a_0 - a_1 x_i)}, \tag{3.1}$$

where Y_i shows the number of successes at the i-th data point; this random variable obeys a binomial distribution. Further, $E[Y_i]$ is the expectation of the number of successes for a probability process obeying the binomial distribution, $P(x_i)$ represents the probability of success when the value of the predictive variable is set to x_i, and a_0 and a_1 are regression coefficients.

The deviance of logistic regression is defined as follows (page 174 in [9]). This is the deviance of the regression equation in the light of available data because this deviance is obtained by considering the effect of available data on the regression equation.

$$D^0 = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right), \tag{3.2}$$

where $\{\hat{\mu}_i\}(1 \leq i \leq n)$ are estimates corresponding to $\{y_i\}(1 \leq i \leq n)$. They are defined as

$$\hat{\mu}_i = \frac{m_i}{1 + \exp(-\hat{a}_0 - \hat{a}_1 x_i)}. \tag{3.3}$$

The maximum likelihood method minimizes D^0 to obtain the values of a_0 and a_1. These regression coefficients given by this method are maximum likelihood estimators ($\hat{a}_0$ and $\hat{a}_1$).

Next, let us consider the deviance of the maximum likelihood estimators considering future data. It is defined as

$$D = 2\sum_{i=1}^{n}\left(y_i^* \log\left(\frac{y_i^*}{\hat{\mu}_i}\right) + (m_i - y_i^*)\log\left(\frac{m_i - y_i^*}{m_i - \hat{\mu}_i}\right)\right), \tag{3.4}$$

where $\{y_i^*\}(1 \leq i \leq n)$ are future data and y_i^* denotes the number of successes on the i-th data point. In contrast, the deviance of predictive estimator considering future data is

$$D^* = 2\sum_{i=1}^{n}\left(y_i^* \log\left(\frac{y_i^*}{\mu_i^+}\right) + (m_i - y_i^*)\log\left(\frac{m_i - y_i^*}{m_i - \mu_i^+}\right)\right). \tag{3.5}$$

The values of x_i and m_i corresponding to y_i^* are the same as those corresponding to y_i. Here, μ_i^+ is written as

$$\mu_i^+ = \frac{m_i}{1 + \exp(-a_0^+ - a_1^+ x_i)}. \tag{3.6}$$

We assume here that a_0^+ and a_1^+ are represented as

$$a_0^+ = \alpha_0 \hat{a}_0, \qquad a_1^+ = \alpha_1 \hat{a}_1, \tag{3.7}$$

where α_0 and α_1 are constants that create the difference between the maximum likelihood estimator and the predictive estimator. The definition of the predictive estimator as given in Eq. (3.7) is based on the two limited observations: (1) these equations are simple and (2) both the third variance and the predictive estimators for exponential distribution are represented in this form. Therefore, we should keep in mind that the different definition of the predictive estimator for logistic regression from Eq. (3.7) may lead to better results.

Equation (2.4) indicates that regression coefficients (a_0^+ and a_1^+) that minimize the value below give the predictive estimator; a_0^+ and a_1^+ lead to $\{\mu_i^+\}$.

$$E_{\{\mathbf{y}_i, \mathbf{y}_i^*\}}[D^*] \tag{3.8}$$

However, an infinite number of future data ($\{y_i^*\}(1 \leq i \leq n)$) cannot be used in most situations. Moreover, only one set of data $\{y_i\}(1 \leq i \leq n)$ can be used as available data. Hence, we suggest using the parametric bootstrap method to calculate an approximate value of Eq. (3.8) in the following section.

4 Predictive Estimator from the Parametric Bootstrap Method

Using bootstrap data provided by the parametric bootstrap method, Eq. (3.8) is approximated to be

$$E_{\{\mathbf{y}_i,\mathbf{y}_i^*\}}[D^*] \approx \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} D_{j,k}^b, \tag{4.1}$$

where $D_{j,k}^b$ is defined as

$$D_{j,k}^b = 2\sum_{i=1}^{n}\left(y_{i,j,k}^f \log\left(\frac{y_{i,j,k}^f}{\hat{\mu}_{i,j}^b}\right) + (m_i - y_{i,j,k}^f)\log\left(\frac{m_i - y_{i,j,k}^f}{m_i - \hat{\mu}_{i,j}^b}\right)\right), \tag{4.2}$$

where Eq. (3.5) is used. $D_{j,k}^b$ $(1 \le i \le n, 1 \le j \le q, 1 \le k \le r)$ stands for deviance between $\{\hat{\mu}_{i,j}^b\}$ and $\{y_{i,j,k}^f\}$. To calculate the value of Eq. (4.2), $\{y_{i,j}^b\}(1 \le i \le n, 1 \le j \le q)$ are first constructed. Here, $\{y_{i,j}^b\}$ are l sets of bootstrap data. These bootstrap data are obtained using the estimates ($\hat{a}_0$ and $\hat{a}_1$), which are given by the maximum likelihood method using available data ($\{y_i\}$). Then, the following equation holds.

$$E[y_{i,j}^b] = \frac{m_i}{1 + \exp(-\hat{a}_0 - \hat{a}_1 x_i)} \tag{4.3}$$

These bootstrap data ($\{y_{i,j}^b\}$) are used as approximations of the available data.

Next, using $\{y_{i,j}^b\}(1 \le i \le n)$, a_0 and a_1 are estimated using the maximum likelihood method. The results are depicted as $\hat{a}_{0,j}^b$ and $\hat{a}_{1,j}^b$. The use of Eq. (3.6) provides the estimates corresponding to y_i as below.

$$\hat{\mu}_{i,j}^b = \frac{m_i}{1 + \exp(-\alpha_0 \hat{a}_{0,j}^b - \alpha_1 \hat{a}_{1,j}^b x_i)} \tag{4.4}$$

Moreover, $\{y_{i,j,k}^f\}(1 \le i \le n)$ are $l \times m$ sets of bootstrap data. These bootstrap data are obtained by the parametric bootstrap method with $\hat{a}_0$ and $\hat{a}_1$. Then, the equation below holds.

$$E[y_{i,j,k}^f] = \frac{m_i}{1 + \exp(-\hat{a}_0 - \hat{a}_1 x_i)} \tag{4.5}$$

The initial values of pseudorandom numbers for calculating $\{y_{i,j,k}^f\}$ are different from those for calculating $\{y_{i,j}^b\}$, which are used as approximations of future data. These values are substituted into Eq. (4.2) to obtain α_0 and α_1 by minimizing the right-hand side of Eq. (4.1). Then, Eq. (3.7) gives the approximated predictive estimator.

5 Execution Examples

Simulation data were constructed using Eq. (3.1). We assumed that $n = 50$ and $\{x_i\} = \{1, 2, 3, \ldots, 50\}$. Each $\{m_i\}(1 \le i \le n)$ were identically 1. Additionally, $a_0 = -3$ and $a_1 = 0.1$ were set. On the basis of these settings, 500 sets of available data ($\{y_i\}$) and 500×500 sets of future data ($\{y_i^*\}$) were created; 500 sets of future data were used for one set of available data. Using these simulation data, the values of D^* (Eq. (3.5)) were calculated and the average of 500 values of D^* were derived and represented as $\bar{D}^*$. The average of 500 values of D given by Eq. (3.4) was also derived and denoted as $\bar{D}$. One of 11 values of $\{0.7, 0.73, 0.76, \ldots, 1\}$ was used as the value of α_0, and one of 11 values of $\{0.7, 0.73, 0.76, \ldots, 1\}$ was also used as the value of α_1. If $\bar{D}^*$ is smaller than $\bar{D}$, it indicates that the predictive estimator is superior to the maximum likelihood estimator from the perspective of prediction.

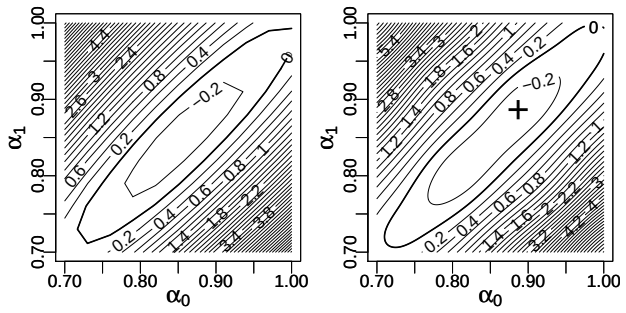

Fig. 1. Relationship between the values of α_0, α_1 and those of $(\bar{D}^* - \bar{D})$ (left). The right graph illustrates the result of smoothing the values of the left graph using smoothing splines. The thick line shows that the value of $(\bar{D}^* - \bar{D})$ is 0. The symbol '+' indicates that $(\bar{D}^* - \bar{D})$ takes the minimum value at that point.

The relationship between the values of α_0, α_1, and $(\bar{D}^* - \bar{D})$ are illustrated in Fig. 1(left). The values in Fig. 1(left) were smoothed using smoothing splines to illustrate Fig. 1(right). For this purpose, the R mgcv package (version 1.8-23, [10]) was used. In Fig. 1(right), the point where $(\bar{D}^* - \bar{D})$ takes the minimum value is indicated by '+'. This optimal point is depicted as $\alpha_0 = 0.882$ and $\alpha_1 = 0.887$. These optimal values of α_0 and α_1 are used in Eq. (3.7) to obtain the optimized predictive estimator.

In most situations, however, future data cannot be utilized. Hence, the value of D^* in Eq. (3.5) cannot be derived. Hence, α_0 and α_1 were derived using Eq. (4.2) and the right-hand side of Eq. (4.1). The same simulation data as those used for Fig. 1 were generated to implement 50 simulations, and $l = 100$ and $m = 50$ were used. Again, one of 11 values of $\{0.7, 0.73, 0.76, \ldots, 1\}$ were used as the values of α_0 and α_1.

The values of α_0 and α_1 given by minimizing the right-hand side of Eq. (4.1) are superimposed on Fig. 1(right) to illustrate Fig. 2. By alternating the initial values of the pseudorandom numbers, 50 numerical simulations were carried out. Each of the four graphs illustrates the results obtained by conducting this numerical simulation four times with other initial values of the pseudorandom numbers. In the first of the four graphs, '×' symbols indicate that 49 out of 50 points are located in the area where the value of $(\bar{D}^* - \bar{D})$ is negative. In the second (upper right), third (lower left), and fourth (lower right) graphs, 48, 47, and 49 points out of 50 points, respectively, are positioned in the area where the value of $(\bar{D}^* - \bar{D})$ is negative. Therefore, we conclude that when the parametric bootstrap method is used, the predictive estimator (Eq. (3.7)) performs better than the maximum likelihood estimator in terms of prediction with a probability of more than 95%. Figure 3 compares a logistic curve given by the maximum likelihood method with one given by the predictive estimator; the data employed here are from one of the simulation datasets used for Fig. 1. The values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ used for Fig. 3(left) were obtained by averaging the values of the '×' points in the first graph of Fig. 2, that is, the averages of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ yielded by the parametric bootstrap method were adopted. The values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ used for Fig. 3(right) represent the position where

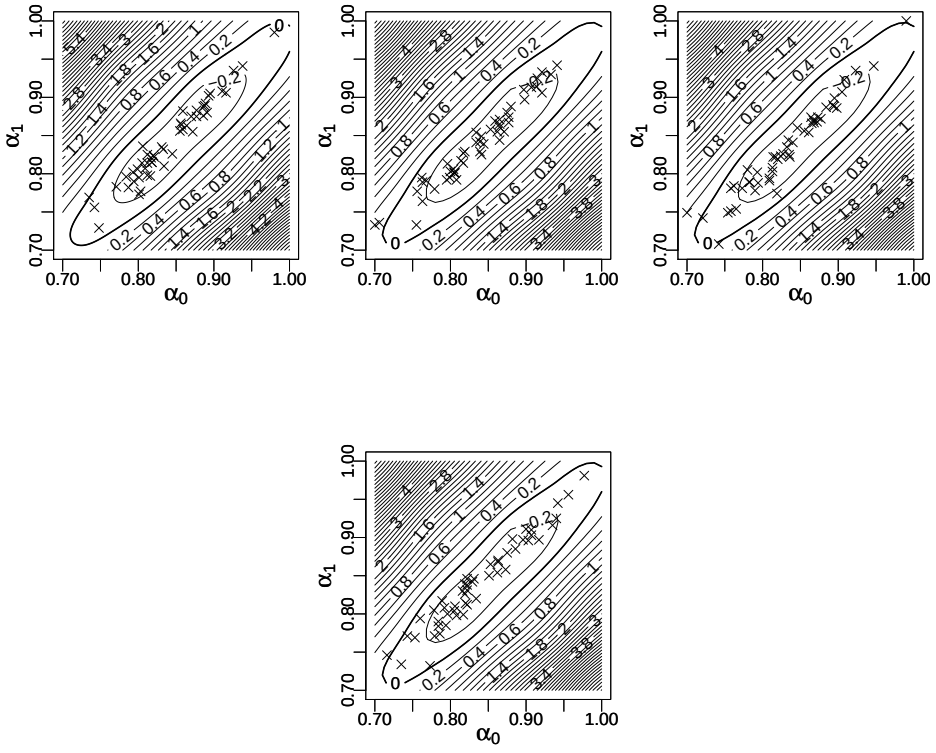

Fig. 2. Values of α_0 and α_1 optimized by the parametric bootstrap method superimposed on Fig. 1(right) (upper left). The upper right, lower left, and lower right graphs use the results given by altering the initial value of pseudorandom numbers in the simulation, yielding the upper left graph.

$(\bar{D}^* - \bar{D})$ takes the minimum value in Fig. 1(right). That is, the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ derived by an infinite number of future data are adopted. Both graphs in Fig. 3 show that the estimates given by the predictive estimator are non-negligibly different from those given by the maximum likelihood estimator. Because we cannot obtain future data in the usual regression analysis, we cannot use future data to confirm the superiority of the predictive estimator (Eq. (3.7)), which contains the optimized values of α_0 and α_1 using the parametric bootstrap method, over the maximum likelihood estimator from the standpoint of prediction. Furthermore, we have to take into account the fact that the bootstrap data provided by parametric bootstrap method are slightly different from real future data. Hence, we use cross-validation to investigate whether the predictive estimator using the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ given by the parametric bootstrap method surpasses the maximum likelihood method from the aspect of prediction. For this purpose, the cross-validated deviance defined below

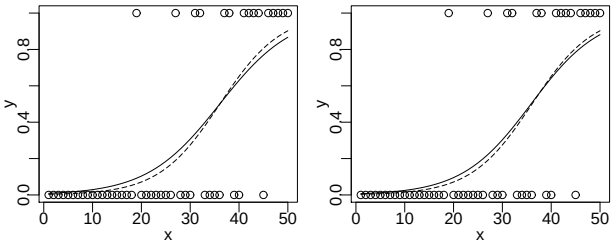

Fig. 3. Logistic curve given by the maximum likelihood estimator (dashed line) and that given by the averaged predictive estimator when one of the simulation datasets used for Fig. 1(left) are adopted (thick line). Here, '◯' symbols depict 50 data points. In the left graph, $\hat{\alpha}_0 = 0.84024$ and $\hat{\alpha}_1 = 0.84066$ (the average of the '×' in the upper left graph of Fig. 2) are set. In the right graph, $\hat{\alpha}_0 = 0.882$ and $\hat{\alpha}_1 = 0.887$, where $(\bar{D}^* - \bar{D})$ takes the minimum value in Fig. 1(right), are set.

is calculated.

$$D^{c*} = \sum_{i=1}^{n} 2\left(y_i \log\left(\frac{y_i}{\mu^+_{i(-i)}} \right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \mu^+_{i(-i)}} \right) \right), \tag{5.1}$$

where $\mu^+_{i(-i)}$ is

$$\mu^+_{i(-i)} = \frac{m_i}{1 + \exp(-\hat{\alpha}_0 \hat{a}_{0(-i)} - \hat{\alpha} \hat{a}_{0(-i)} x_i)}, \tag{5.2}$$

where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ stand for the optimized values of α_0 and α_1 using the parametric bootstrap method, $\hat{a}_{0(-i)}$ and $\hat{a}_{1(-i)}$ represent the regression coefficients given by carrying out the maximum likelihood method after the i-th data has been deleted. Moreover, when the maximum likelihood method is employed, that is, $\hat{\alpha}_0 = 1$ and $\hat{\alpha}_1 = 1$ are set, the resultant D^{c*} is termed D^c.

Cross-validation in the exact sense should use the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ given by deleting the i-th data. However, the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ used here are calculated using all the data because this reduces the computational cost. The resultant values of $(D^{c*} - D^c)$ are illustrated in Fig. 5. In these four graphs, the numbers of datasets that make the value of $(D^{c*} - D^c)$ positive are 0 sets, 1 set, 2 sets, and 0 sets, respectively. These results show that $(D^{c*} - D^c)$ takes a positive value in rare settings. We note that although the value of $(D^{c*} - D^c)$ is the approximated value of $(\bar{D}^* - \bar{D})$, the use of a dataset that makes the value of $(D^{c*} - D^c)$ positive does not always result in the maximum likelihood method outperforming the predictive estimator. However, when the predictive error provided by cross-validation does not indicate the superiority of the predictive estimator, we should conclude that the validity of the results given by the parametric bootstrap method is not confirmed and hence the maximum likelihood method is a safe choice.

An alternative idea is to use cross-validation for estimating the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ for constructing the predictive estimator. This is because we prefer to use cross-validation for constructing a desirable predictive estimator when the performance of the predictive estimator and that of the maximum likelihood estimator are compared using cross-validation. However, it is well known that the results obtained by cross-validation are highly variable (e.g. [21], [22], [23]). We found a similar tendency with the predictive estimator proposed here, which is that the estimation of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ gives highly variable results. Therefore, our tentative suggestion is the procedure noted above: (1) $\hat{\alpha}_0$ and $\hat{\alpha}_1$ should be estimated by the parametric bootstrap method and (2) the results should be examined

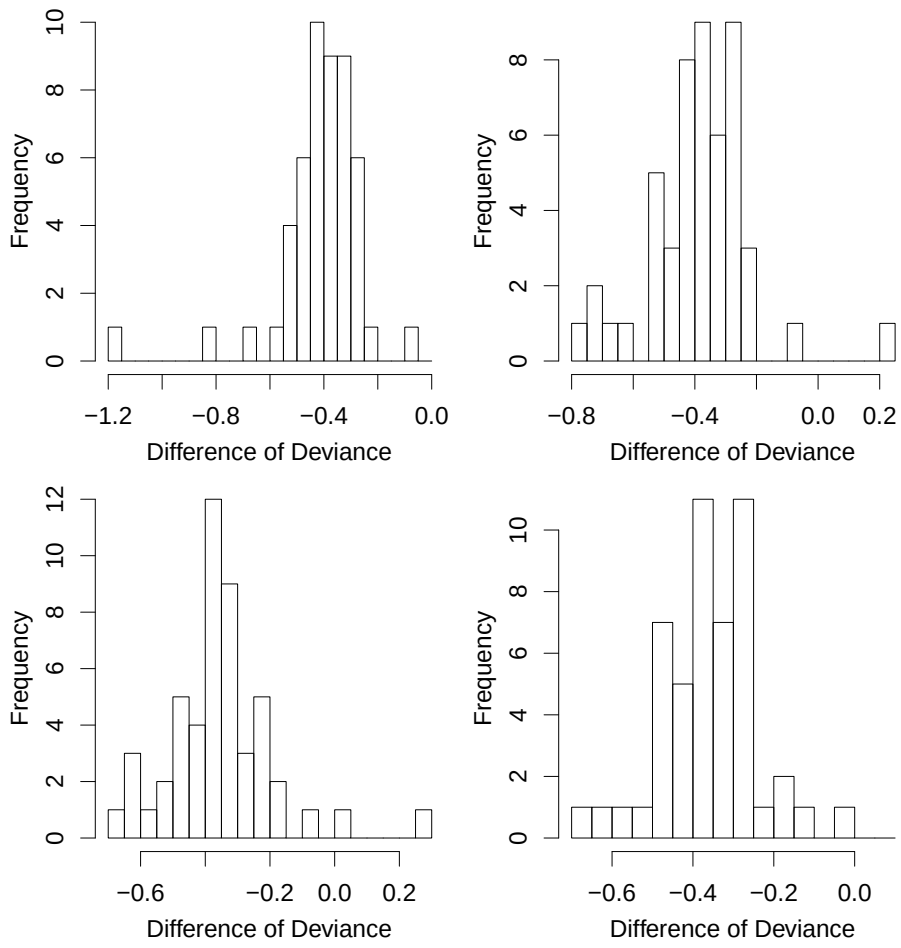

Fig. 4. Values of $(D^{c*} - D^c)$ derived by 50 sets of simulation data used in each of the four graphs in Fig. 2.

by cross-validation. Next, we set $n = 100$ and $\{x_i\} = \{1, 2, 3, \ldots, 100\}$. Each $\{m_i\}(1 \leq i \leq n)$ are all set to 1 again. We also set $a_0 = -3$ and $a_1 = 0.05$. The value of α_0 is one of 11 values of $\{0.7, 0.73, 0.76, \ldots, 1\}$ again. The value of α_1 is also one of 11 values of $\{0.7, 0.73, 0.76, \ldots, 1\}$. These settings lead to the results in Fig. 5, which corresponds to Figs. 2 and 4. A total of 48 sets of values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ out of the 50 datasets led by the parametric bootstrap method are located in the region where the value of $(\bar{D}^* - \bar{D})$ is negative. Hence, it is very probable that the use of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ derived by the bootstrap method creates a predictive estimator that outperforms the maximum likelihood estimator. The results in Fig. 5(right), given by cross-validation, show that 49 out of 50 datasets led to predictive estimators that yielded better results than the maximum likelihood estimator in terms of prediction. Consequently, the predictive estimator is recommended for 49 data sets and the maximum likelihood estimator is recommended for the remaining data set. Next, we newly set $n = 30$ and $\{x_i\} = \{1, 2, 3, \ldots, 30\}$. Each $\{m_i\}(1 \leq i \leq n)$ are all 1 again. We

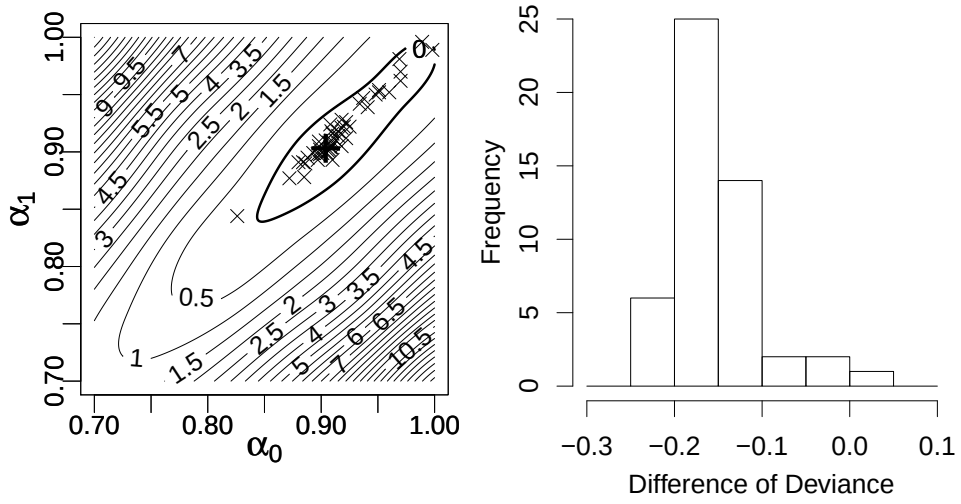

Fig. 5. Results corresponding to Fig. 2 and Fig. 4 when $n = 100$, $\{x_i\} = \{1, 2, 3, \ldots, 100\}$, $a_0 = -3$, and $a_1 = 0.05$ are assumed. In the left graph, $\hat{\alpha}_0 = 0.899$ and $\hat{\alpha}_1 = 0.904$(indicated by '+') minimizes the value of $(\bar{D}^* - \bar{D})$. The right graph shows the values of $(D^{c*} - D^c)$ given by 50 sets of simulation data.

also newly set $a_0 = -1$ and $a_1 = 0.08$. The value of α_0 is one of 11 values of $\{0.2, 0.28, 0.36, \ldots, 1\}$. The value of α_1 is also one of 11 values of $\{0.2, 0.28, 0.36, \ldots, 1\}$. These settings lead to the results in Fig. 6 which corresponds to Figs. 2 and 4.

The values of $(\bar{D}^* - \bar{D})$ are shown in Fig. 6(left). All 50 sets of values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ out of the 50 datasets obtained by the parametric bootstrap method are located in the region where the value of $(\bar{D}^* - \bar{D})$ is negative. Hence, it is very probable that use of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ derived by the bootstrap method creates a predictive estimator that outperforms the maximum likelihood estimator.

Figure 6(right) given by cross-validation shows that all 50 datasets yielded predictive estimators that lead to better results than the maximum likelihood estimator in terms of prediction. Thus, the predictive estimator is recommended for all 50 data sets.

Figure 7(left) compares a regression curve given by the averages of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ obtained by the parametric bootstrap method with that given by the maximum likelihood method. This graph shows that the regression curve derived by the predictive estimator using the bootstrap method is somewhat different from that derived by the maximum likelihood estimator. Figure 7(right) compares the regression curve given by the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ that minimize the value of $(\bar{D}^* - \bar{D})$ (Fig. 6(left)) with that given by the maximum likelihood method. We find that when the values of $(\bar{D}^* - \bar{D})$ obtained by a large number of future data are employed, the resultant regression curve is appreciably different from those given by the maximum likelihood estimator. This example

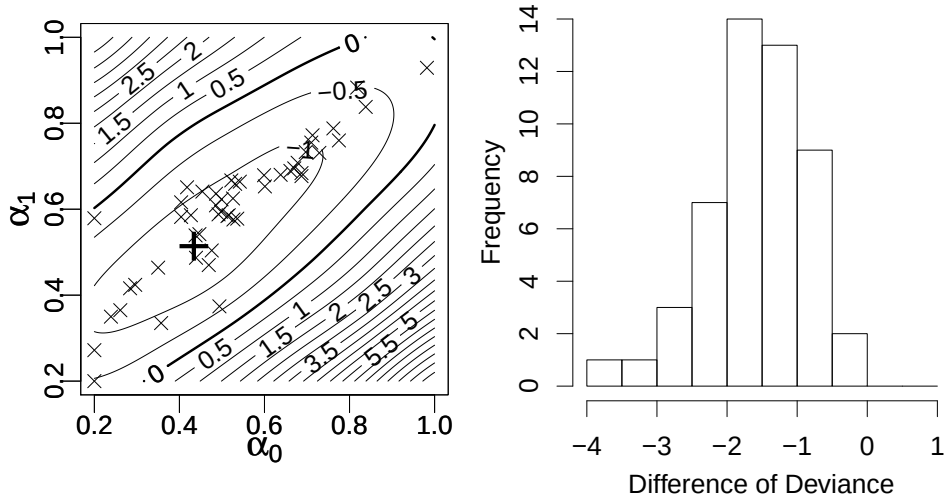

Fig. 6. Results corresponding to Figs. 2 and 4 when $n = 30$, $\{x_i\} = \{1, 2, 3, \ldots, 30\}$, $a_0 = -1$, and $a_1 = 0.08$ are assumed. In the left graph, $\hat{\alpha}_0 = 0.503$ and $\hat{\alpha}_1 = 0.434$ minimizes the value of $(\bar{D}^* - \bar{D})$. The right graph presents the values of $(D^{c*} - D^c)$ given by 50 sets of simulation data.

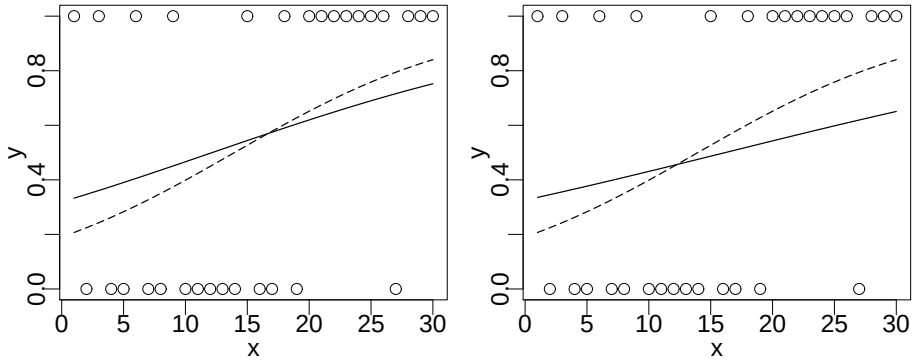

Fig. 7. Results corresponding to Fig. 3 when $n = 30$, $\{x_i\} = \{1, 2, 3, \ldots, 30\}$, $a_0 = -1$, and $a_1 = 0.08$ are assumed.

emphasizes the fact that the difference between fitting to available data and fitting to future data is sometimes too large to be neglected. Although the regression curve given by the predictive

estimator using the parametric bootstrap method is a little bit different from that obtained by $(\bar{D}^* - \bar{D})$, we conclude that the parametric bootstrap method led to favourable changes in the maximum likelihood method with respect to prediction because Figs. 6(left) and 6(right) show that many of the predictive estimators derived by the parametric bootstrap method are better from the perspective of prediction than the maximum likelihood estimators.

It is true that the conditions of these numerical simulation are very limited. The number of independent variables in all numerical simulations is one because implementation of parametric bootstrap method for multiple independent variable data is too computer intensive for us. Further, convergence of bootstrap data are not investigated here; this will be a topic of future research. Nevertheless, these simulation studies show that further consideration will be needed for the maximum likelihood method and principle of maximum likelihood.

6 Conclusions

The maximum likelihood estimator is widely used for estimating parameters in logistic regression. The background of this tradition is the naive belief that a regression model that fits well to data in the past will fit well to data in the future. However, the results of numerical simulations in the last section indicate that the predictive estimator (Eq. (3.7)) given by estimating the values of α_0 and α_1 using the parametric bootstrap method provides more favourable results than the maximum likelihood estimator from the aspect of prediction. Moreover, the high practical utility of the predictive estimator obtained using the parametric bootstrap method can be verified in terms of prediction by applying cross-validation to the results. The findings given by these simulation studies imply that in many estimations such as generalized linear regression in which the maximum likelihood estimator is universally used, the construction of predictive estimators by making straightforward prediction can provide more desirable results than the maximum likelihood estimator.

For this purpose, diverse forms of predictive estimator should be investigated; the form should not be limited to multiplication of the maximum likelihood estimator and constants (Eq. (3.7)). Another promising choice is the construction of an estimator for the purpose of prediction using regularization such as ridge regression, LASSO, or both; 'glmnet' ([24]; [25]) is a typical example. Moreover, if the analytical maximization of Eq. (2.3) is realized, that is, the constants in the predictive estimator such as α_0 and α_1 in Eq. (3.7) are derived analytically, numerical methods such as the parametric bootstrap method will no longer be needed. Even if the constants in the predictive estimator are not obtained analytically, further analytical studies on the constants should provide more efficient predictive estimators than those given by the parametric bootstrap method.

Estimators for beneficial prediction play an important role in prediction, control, and decision-making. The advent of a predictive estimator has freed us from the constraints of the maximum likelihood estimator. That is, we have reached a new stage at which we have obtained a clear understanding that the best prediction using the available data is different from constructing regression equations that fit closely to the available data. This notion is commonly known in the field of statistical learning. On page 30 in [15], it says:

Rather, *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data* (emphasis in original).

The concept of a predictive estimator is created by generalizing this philosophy to the whole of statistical estimation methods. We firmly hope that various features of the predictive estimator are elucidated by analytical methods and numerical ones to create predictive estimators of high practical value.

Acknowledgement

The author is very grateful to the referees for carefully reading the paper and for their comments and suggestions which have improved the paper.

Competing Interests

The author declare that no competing interests exist.

References

[1] Millar RB. Maximum likelihood estimation and inference: with examples in R, SAS and ADMB. Wiley. NJ, U.S.A.; 2011.

[2] Takezawa K. A revision of AIC for normal error models. Open Journal of Statistics. 2012;2(3):309-312.

[3] Takezawa K. Learning regression analysis by simulation. Springer, Tokyo, Japan; 2014.

[4] Takezawa K. Estimation of the exponential distribution in the light of future data. British Journal of Mathematics & Computer Science. 2015;5(1):128-132.

[5] Ogasawara H. Predictive estimation of a covariance matrix and its structural parameters. Journal of the Japanese Society of Computational Statistics. 2017;30:45-63.

[6] Ogasawara H. A family of the adjusted estimators maximizing the asymptotic predictive expected log-likelihood. Behaviormetrika. 2017;44:57-95.

[7] Ogasawara H. An asymptotic equivalence of the cross-data and predictive estimators. Communications in Statistics - Theory and Methods. 2019;1-14.

[8] Mc Cullagh P, Nelder JA. Generalized linear models, second edition. Chapman & Hall/CRC. Boca Raton, FL, U.S.A.; 1989.

[9] Myers RH, Montgomery DC. Vining GG, Robinson TJ. Generalized Linear Models: With Applications in Engineering and the Sciences (first edition). Wiley. NJ, U.S.A.; 2002.

[10] Wood SN. Generalized additive models: An introduction with R, second edition. Chapman & Hall/CRC. Boca Raton, FL, U.S.A.; 2017.

[11] Dobson AJ, Barnett AG. An introduction to generalized linear models, fourth edition. Chapman & Hall/CRC. Boca Raton, FL, U.S.A.; 2018.

[12] Takezawa K. Predictive estimator for simple regression. Journal of Advances in Mathematics and Computer Science. 2017;24(4):1-14.

[13] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall/CRC. Boca Raton, FL, U.S.A.; 1993.

[14] Takezawa K. Optimal estimator with respect to expected log-likelihood. International Journal of Innovation in Science and Mathematics. 2014;2(6):494-508.

[15] James G, Witten G, Hastie D, Tibshirani T, James R. An introduction to statistical learning: With applications in R. New York: Springer; 2013.

[16] Akaike H. Information theory and an extension of the maximum likelihood principle. Proceedings of 2nd International Symposium on Information Theory (Petrov BN. Csaki F. (Eds.)),. Budapest: Akademiai Kiado. 1973;267-281.

[17] Akaike H. A new look at the statistical model identification. IEEE Transaction on Automatic Control. 1974;19(6):716-723.

[18] Konishi S, Kitagawa G. Information criteria and statistical modelling. New York: Springer; 2008.

[19] Wahba G. Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial and Applied Mathematics; 1990.

[20] Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 1979;21(2):215-223.

[21] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association. 1983;78;316-331.

[22] Efron B, Tibshirani RJ. Improvements on cross-validation: The .632+ bootstrap method. Journal of the American Statistical Association. 1997;92(438):548-560.

[23] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010;4;40-79.

[24] Friedman JH, Hastie T, Tibshirani R. Regularized paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33(1);1-22.

[25] Hastie T, Qian J. Glmnet Vignette. Stanford; 2016.
Available:https://web.stanford.edu/ hastie/Papers/Glmnet$_V$*ignette.pdf*

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle3.com/review-history/49483