Journal of Advances in Mathematics and Computer Science

24(4): 1-14, 2017; Article no.JAMCS.35869

Previously known as British Journal of Mathematics & Computer Science
ISSN: 2231-0851

Predictive Estimator for Simple 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 and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2017/35869
Editor(s):
(1) Metin Basarir, Department of Mathematics, Sakarya University, Turkey.
Reviewers:
(1) Marcos Grilo, Universidade Estadual de Feira de Santana, Brasil.
(2) R. Shenbagavalli, SRM University, India.
(3) Jagdish Prakash, University of Botswana, Botswana.
(4) Ajay Auddy, The University of Burdwan, India.
(5) Suk-Bok Kang, Yeungnam University, South Korea.
Complete Peer review History: http://www.sciencedomain.org/review-history/21049

Received: 31st July 2017
Accepted: 6th September 2017
Published: 19th September 2017

Original Research Articles

Abstract

The predictive estimator of the gradient in simple regression is assumed to be the product of the gradient given by least-squares fitting and a constant (ρ). The results of numerical simulations show that when generalized cross-validation is used to obtain the optimal ρ, the resultant predictive estimator is not of great use. However, when the parametric bootstrap method is applied for this purpose, the resulting predictive estimator is often superior to the maximum likelihood estimator in terms of prediction accuracy. Therefore, statistics reflecting the characteristics of data should be used to determine which estimator should be adopted.

Keywords: Expected log-likelihood; future data; maximum likelihood estimator; predictive estimator; simple regression.

2010 Mathematics Subject Classification: 60G25, 62F10, 62M20.

**Corresponding author: E-mail:nonpara@gmail.com, takezawa@affrc.go.jp;*

1 Introduction

The maximum likelihood estimator does not always lead to the best results in terms of prediction accuracy. This is because the estimates that gave the best fit to the data in the past may not fit well with data in the future. Therefore, while the estimator that gives the best fit to past data is called the maximum likelihood estimator, the estimator which yields the best predictions is called the "predictive estimator". For example, the "third variance" ($(n-1)$ (n is the number of data) in unbiased variance is replaced with $(n-4)$ for reducing predictive error.) [1, 2], which refers to the variance for the purpose of prediction, has been derived; a method based on series expansion gives an estimator which is asymptotically identical to the third variance [3], and predictive estimators have been constructed [4] using the bootstrap method (section 6.7 of [5]).

In simple regression, when the absolute value of the gradient is adjusted to be slightly less than the value obtained by the least-squares method, the estimates tend to fit well with future data [4]. Hence, specific procedures for obtaining predictive estimators for simple regression when data are available are presented with the help of numerical simulations. The second section presents an outline of the predictive estimator for simple regression. The third section introduces a method of deriving the predictive estimator using Generalized Cross-Validation (GCV) ([6]; Section 4.3 of [7]; [8]), and the characteristics of this estimator are investigated. The fourth section derives a method of producing the predictive estimator based on the parametric bootstrap method, and its features are examined using numerical simulations. The fifth section provides a technique for using the predictive estimator and maximum likelihood estimator differently.

2 Predictive Estimator for Simple Regression

The variates x and y are described as follows:

$$y = \alpha + \beta x + \epsilon,$$

where x is a predictive variable and y is an objective variable. ϵ is a random error obeying $N(0, \sigma^2)$ (normal distribution with mean 0 and variance σ^2). α and β are constants called the intercept and the gradient, respectively. It is assumed that n sets of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are available, and these data are given by the regression equation above. We may assume $\sum_{i=1}^{n} x_i = 0$ without loss of generality. Hence, we use this assumption hereafter.

Now, the following equation holds for the available data:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where ϵ_i is a realization of ϵ. Eliminating errors from $\{y_i\}$ yields $\{\tilde{y}_i\}$, which are the true values corresponding to $\{y_i\}$. Then, we have

$$\tilde{y}_i = \alpha + \beta x_i, \qquad i = 1, 2, \ldots, n.$$

Using the n sets of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the estimates of α and β given by the least-squares method are denoted by $\hat{\alpha}$ and $\hat{\beta}$. The least-squares estimates ($\hat{y}$) are written as

$$\hat{y} = \hat{\alpha} + \hat{\beta} x.$$

The relationship $\sum_{i=1}^{n} x_i = 0$ leads to the following equation (e.g. page 13 in [9]):

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i}{n}, \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2.2}$$

The estimate corresponding to each $\{y_i\}$ is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i.$$

Hence, the third variance gives the estimate of σ^2 (i.e. $\hat{\sigma}^2$):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n - \gamma}, \tag{2.3}$$

where γ is defined as

$$\gamma = n\left(1 - \frac{n - q - 3}{n + q + 1}\right),$$

for $q = 1$ (because simple regression is considered here). Derivation of γ above is described in [1, 2].The hat matrix for Eq. (2.2) is (e.g. page 134 in [9]):

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t.$$

$\mathbf{X}$ and $\mathbf{y}$ are written as

$$\begin{aligned} \mathbf{X} &= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \\ \mathbf{y} &= (y_1, y_2, \ldots, y_n)^t. \end{aligned}$$

Now, let us assume that $\{\epsilon_i^*\}$ $(1 \leq i \leq n)$ are errors which are independent of $\{\epsilon_i\}$ and obey $N(0, \sigma^2)$. Then, $\{\tilde{y}_i + \epsilon_i^*\}$ $(1 \leq i \leq n)$ are treated as future data. The estimate (y^+) which fits this future data well is represented as

$$y^+ = \hat{\alpha} + \rho\hat{\beta}x. \tag{2.4}$$

Although, ρ is ofted used as a correlation coefficient, this ρ is not related to the correlation coefficient. As the estimates $\{\hat{\alpha} + \hat{\beta}x\}$ fit the available data well, we assume that slightly different estimates could fit the future data well. In [10], it is assumed that the following gives a good fit to the true value:

$$y^+ = \rho\hat{\alpha} + \rho\hat{\beta}x. \tag{2.5}$$

That is, the constant term is multiplied by ρ. Furthermore, [11] discusses the use of cross-validation for this sort of estimation in simple regression and multiple regression, although the value of the constant term is adjusted somewhat differently than in Eq. (2.5). However, Eq. (2.4) is adopted here to ensure addition invariance for the value of the objective variable and consistency with [4].

Then, $\hat{\rho}$ is obtained to minimize the following value by adjusting ρ:

$$er_\rho = \sum_{i=1}^{n}(\tilde{y}_i + \epsilon_i^* - \hat{\alpha} - \rho\hat{\beta}x_i)^2, \tag{2.6}$$

where $\{\tilde{y}_i + \epsilon_i^*\}$ are future data. The expectation of Eq. (2.6) with respect to $\{\epsilon_i^*\}$ gives

$$\begin{aligned} E^*[er_\rho] &= E^*\left[\sum_{i=1}^{n}(\tilde{y}_i + \epsilon_i^* - \hat{\alpha} - \rho\hat{\beta}x_i)^2\right] \\ &= \sigma^2 + \sum_{i=1}^{n}(\tilde{y}_i - \hat{\alpha} - \rho\hat{\beta}x_i)^2. \end{aligned} \tag{2.7}$$

The value of ρ which minimizes the above equation also minimizes the expectation of the sum of squares between the estimates given by the regression equation and the future data. The resultant ρ gives the predictive estimator.

The minimization of $E^*[er_\rho]$ defined by Eq. (2.7) is equivalent to that of $E^*[er_\rho]-\sigma^2$ defined below:

$$\begin{aligned} E^*[er_\rho]-\sigma^2 &= E^*\left[\sum_{i=1}^{n}(\tilde{y}_i+\epsilon_i^*-\hat{\alpha}-\hat{\rho}\hat{\beta}x_i)^2\right]-\sigma^2 \\ &= \sum_{i=1}^{n}(\tilde{y}_i-\hat{\alpha}-\hat{\rho}\hat{\beta}x_i)^2. \end{aligned}$$

The estimates which minimize the above value also minimize the Mean Squared Error (MSE): the expectation of the sum of squares of the difference between estimates given by the regression equation and the true values. Hence, in this problem, the predictive estimator is identical to the estimator that minimizes MSE. The James–Stein estimator ([12, 13]) is a well-known example that minimizes MSE, although there are no reports of the James–Stein estimator being applied to simple regression.

3 Predictive Estimator Given by GCV

To minimize $E^*[er_\rho]$ (Eq.(2.7)), the minimization of the following GCV is a possible strategy:

$$GCV = \frac{\sum_{i=1}^{n}(y_i-y_i^+)^2}{n\cdot\left(1-\frac{\sum_{i=1}^{n}[\mathbf{H}^*]_{ii}}{n}\right)^2}, \tag{3.1}$$

where $\mathbf{H}^*$ is written as

$$\mathbf{H}^* = \mathbf{DH}, \tag{3.2}$$

and $\mathbf{D}$ is a diagonal matrix in which the diagonal elements are $\left\{\frac{\hat{\alpha}+\rho\hat{\beta}x_i}{\hat{y}_i}\right\}(1\le i\le n)$; this setting leads to Eq. (2.4).

In optimizing ρ so as to minimize GCV defined in Eq. (3.1), the resultant simple regression equation is considered to be optimal in terms of prediction accuracy. The value of ρ given by this method is denoted as $\hat{\rho}_{gcv}$.

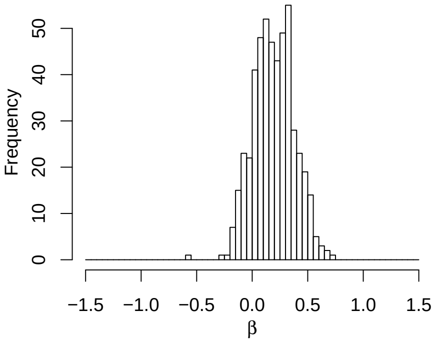

Fig. 1. Histogram of $\hat{\beta}$ obtained using 500 simulation data

To examine how this method performs, the following numerical simulation was conducted. Firstly, the number of data (n) in Eq. (2.1) was set to 21, and the following values were assigned: $\{x_i\} = \{-10, -9, -8, \ldots, 10\}$. Furthermore, $\alpha = 0$ and $\beta = 0.2$ were set, and $\{\epsilon_i\}$ were assumed to be realizations of $N(0.0, 5^2)$ (normal distribution with mean 0.0 and variance 5^2). A total of 500 sets of data were generated by altering the initial pseudo-random values. For each dataset, $\hat{\beta}$ was estimated using Eq. (2.2). A histogram of the distribution of the resulting $\hat{\beta}$ is shown in Fig. 1.

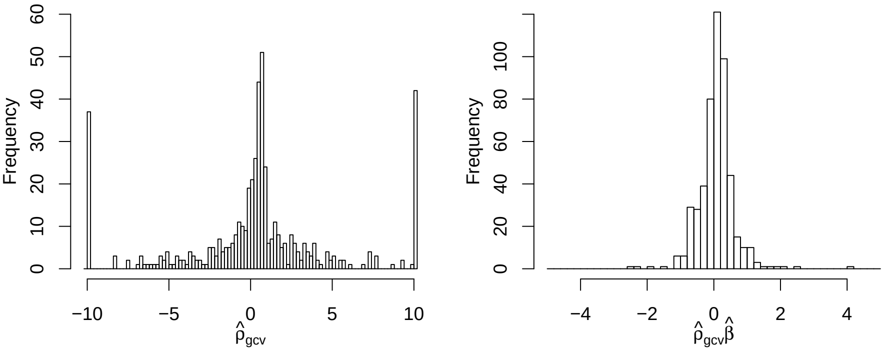

Fig. 2. Histogram of the distribution of $\{\hat{\rho}_{gcv}\}$ (left). Histogram of the distribution of $\{\hat{\rho}_{gcv}\hat{\beta}\}$ (right)

Next, ρ was set to one of $\{-10, -9.95, 9.9, \ldots, 10\}$(401 values), and ρ was derived using GCV defined in Eq. (3.1). A histogram of the resulting $\{\hat{\rho}_{gcv}\}$ is shown in Fig. 2 (left). These values are multiplied by $\{\hat{\beta}\}$ to obtain $\{\hat{\rho}_{gcv}\hat{\beta}\}$. A histogram of $\{\hat{\rho}_{gcv}\hat{\beta}\}$ is shown in Fig. 3(right).

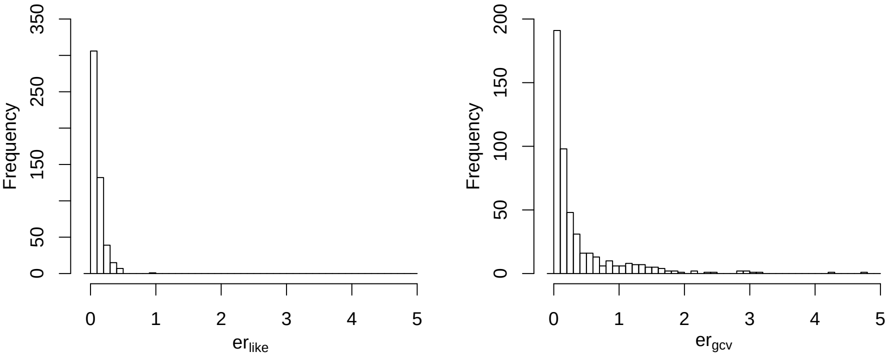

Fig. 3. Histogram of values of er_{like} (left). Distribution of values of er_{gcv} (right)

The prediction error given by the regression coefficients derived using the maximum likelihood method is defined as

$$er_{like} = \sum_{i=1}^{n}(\tilde{y}_i - \hat{\alpha} + \hat{\beta}x_i)^2. \tag{3.3}$$

The distribution of the values of er_{like} given by 500 numerical simulations is illustrated in Fig. 3

(left). The mean of these er_{like} values is 0.0998.

Moreover, the prediction error yielded by the optimal ρ in terms of GCV (i.e. $\hat{\rho}_{gcv}$) is defined as

$$er_{gcv} = \sum_{i=1}^{n}(\tilde{y}_i - \hat{\alpha} + \hat{\rho}_{gcv}\hat{\beta}x_i)^2.$$

The distribution of er_{gcv} given by 500 numerical simulations is shown in Fig. 3(right). In Fig. 3, six values greater than 5 have been eliminated (6.2243, 7.3938, 7.6779, 9.1003, 11.4130, 22.9210). The mean of the er_{gcv} values is 0.5072.

Therefore, the prediction error given by the simple regression equation derived using GCV is far higher than that derived using the maximum likelihood method. This result seems to reflect the fact that optimization using cross-validation gives estimates with a large variance (e.g. [14]). Thus, even if ρ is optimized in terms of prediction accuracy, the resulting estimates are less reliable for making predictions because of the high variance of the optimized ρ.

4 Predictive Estimator Given by Parametric Bootstrap Method

The numerical simulation in the previous section indicates that, even if a predictive estimator is intended to reduce the prediction error and the parameters in the estimator are optimized in terms of prediction accuracy, the resulting estimator does not always outperform the maximum likelihood estimator. However, GCV is not the only tool for optimizing the parameters contained in predictive estimators, and other approaches can be used. Thus, the application of the parametric bootstrap method is described below.

The value of ρ that minimizes Eq. (2.7) (i.e. $\hat{\rho}$) is written as (page 33 in [9]):

$$\hat{\rho} = \frac{\sum_{i=1}^{n} \hat{y}_i \tilde{y}_i}{\sum_{i=1}^{n} \hat{y}_i^2}, \tag{4.1}$$

where the $\{\tilde{y}_i\}$ do not contain errors. That is, these are the unknown true values. Then, the $\{\hat{y}_i\}$ are used as approximations of $\{\tilde{y}_i\}$. Next, $\{\hat{y}_i\}$ in Eq. (4.1) are replaced with $\{\hat{y}_i^*\}$ defined as

$$\hat{y}_i^* = \mathbf{H}^* y_i^*,$$

where $\mathbf{H}^*$ is defined in Eq. (3.2). The $\{y_i^*\}$ are written as

$$y_i^* = \hat{y}_i + \sqrt{\hat{\sigma}^2} u_i,$$

where $\{u_i\}$ $(1 \le i \le n)$ are the errors, which obey $N(0, 1)$.

Using these equations, Eq. (4.1) is approximated as

$$\hat{\rho} \approx \frac{\sum_{i=1}^{n} \mathbf{H}^* y_i^* \hat{y}_i}{\sum_{i=1}^{n} (\mathbf{H}^* y_i^*)^2}.$$

As the $\{u_i\}$ are random, the value of ρ obtained by taking the expectation with respect to $\{u_i\}$ is denoted as $\hat{\rho}_{boot}$:

$$\begin{aligned} \hat{\rho}_{boot} = E_{\{u_i\}}\big[\hat{\rho}\big] &= E_{\{u_i\}}\left[\frac{\sum_{i=1}^{n} \mathbf{H}^* y_i^* \hat{y}_i}{\sum_{i=1}^{n} (\mathbf{H}^* y_i^*)^2}\right] \\ &= E_{\{u_i\}}\left[\frac{\sum_{i=1}^{n} \mathbf{H}^* (\hat{y}_i + \sqrt{\hat{\sigma}^2} u_i) \hat{y}_i}{\sum_{i=1}^{n} \big(\mathbf{H}^* (\hat{y}_i + \sqrt{\hat{\sigma}^2} u_i)\big)^2}\right]. \end{aligned} \tag{4.2}$$

In the actual estimation, Eq. (4.2) is approximated as

$$\hat{\rho}_{boot} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\sum_{i=1}^{n}\mathbf{H}^{*}(\hat{y}_i + \sqrt{\hat{\sigma}^2}u_{ik})\hat{y}_i}{\sum_{i=1}^{n}\left(\mathbf{H}^{*}(\hat{y}_i + \sqrt{\hat{\sigma}^2}u_{ik})\right)^2}\right), \tag{4.3}$$

where $\{u_{ik}\}$ $(1 \leq i \leq n)$ are the errors, which obey $N(0,1)$. K is set so that $\hat{\rho}_{boot}$ is almost independent of the initial value of the pseudo-random numbers which generate $\{u_{ik}\}$.

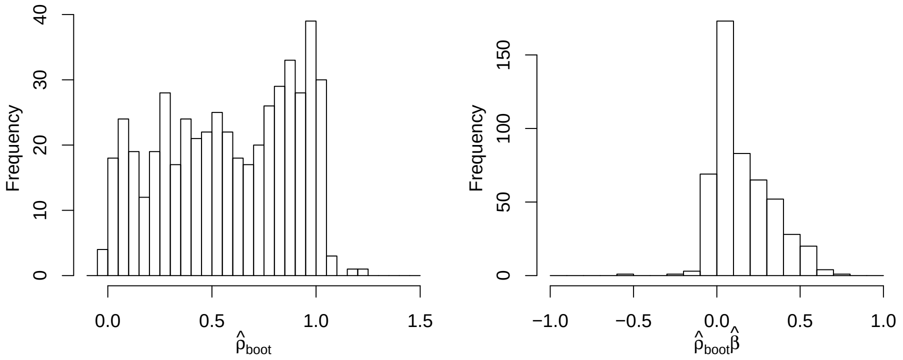

Fig. 4. Histogram of $\{\hat{\rho}_{boot}\}$ (left). Histogram of $\{\hat{\rho}_{boot}\hat{\beta}\}$ (right)

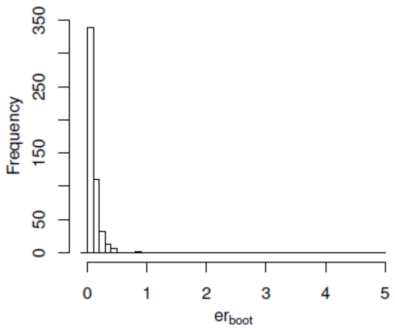

Fig. 5. Histogram of values of er_{boot}

Using the results of the numerical simulation in the previous section, $\hat{\rho}_{boot}$ was calculated using Eq. (4.3) for $K = 200$. The distribution of the resulting $\hat{\rho}_{boot}$ is shown in Fig. 4(left), and the distribution of $\hat{\rho}_{boot}\hat{\beta}$ is shown in Fig. 4(right).

The prediction error given by this result is defined as

$$er_{boot} = \sum_{i=1}^{n}(\tilde{y}_i - \hat{\alpha} + \hat{\rho}_{boot}\hat{\beta}x_i)^2. \tag{4.4}$$

The distribution of er_{boot} is illustrated in Fig. 5. The mean of er_{boot} is 0.0967. This is less than that given by the maximum likelihood method (= 0.0998). Therefore, the predictive estimator given by this method has the potential to outperform the maximum likelihood estimator in terms of prediction accuracy.

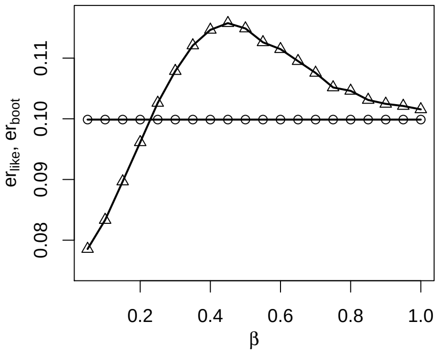

Fig. 6. er_{like} **(○) and** er_{boot} **(△) when** β **is set to be one of** $\{0.05, 0.1, 0.15, \ldots, 1\}$

However, this method does not always give better results than those of the maximum likelihood estimator from the aspect of prediction accuracy. For example, Fig. 6 illustrates the values of er_{like} and er_{boot} when β was set to one of 20 values: $\{0.05, 0.1, 0.15, \ldots, 1\}$; the other settings were the same as in the above numerical simulations. These results indicate that when β is less than 0.225, the prediction error provided by the predictive estimator is less than that given by the maximum likelihood estimator. Additionally, the difference between the two prediction errors peaks at $\beta = 0.45$ and then gradually becomes smaller, although the relationship between the two prediction errors is not reversed (this is not fully guaranteed by the limited number of simulations though). Hence, a threshold (such as $\beta = 0.225$ in this example) should be derived to determine whether to use the maximum likelihood estimator or a predictive estimator.

5 Distinct Usage of Predictive Estimator and Maximum Likelihood Estimator

Fig. 6 indicates that the predictive ability of the predictive estimator based on the parametric bootstrap method can be inferior to that of the maximum likelihood estimator; it is data dependent. Estimations with high predictive ability are to be expected if the data is represented using some statistics which prefer a predictive estimator above a certain threshold; otherwise, the maximum likelihood estimator is adopted. These statistics should choose the maximum likelihood estimator when the value of $\hat{\beta}$ is large and choose a predictive estimator if the value of $\hat{\beta}$ is small. The Δ statistic [4] has been defined for this purpose:

$$\Delta = \frac{\sqrt{var(\hat{\beta})}}{|\beta|}, \tag{5.1}$$

where $var(\hat{\beta})$ denotes the variance of $\hat{\beta}$, $|\beta|$ denotes the absolute value of β, and $var(\hat{\beta})$ is defined as (e.g. page 14 in [9]):

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\bar{x}$ is the mean of $\{x_i\}$. Although Eq. (5.1) is derived intuitively, numerical simulations prove its validity [4]. Because σ^2 is the true value (in the population) of the error variance, σ^2 should be replaced with $\hat{\sigma}^2$ (Eq. (2.3)). Then, $var(\hat{\beta})$ should be replaced with $\widehat{var}(\hat{\beta})$. When the result is represented as $\hat{\Delta}$, we have

$$\hat{\Delta} = \frac{\sqrt{\widehat{var}(\hat{\beta})}}{|\hat{\beta}|}, \tag{5.2}$$

where $\widehat{var}(\hat{\beta})$ is written as

$$\widehat{var}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

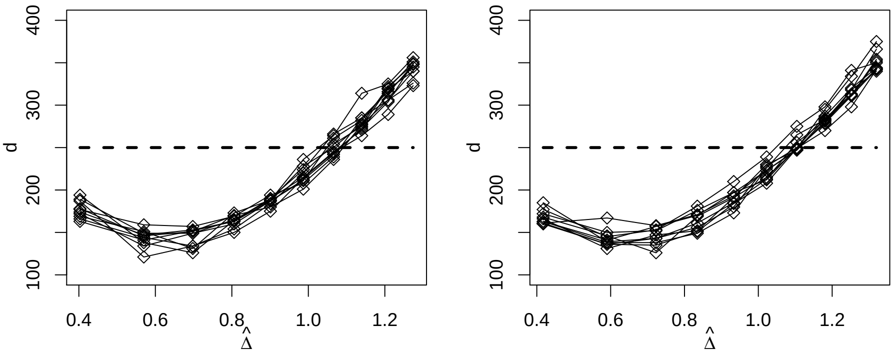

Fig. 7. Relationship between $\hat{\Delta}$ and d in the first numerical simulation (the dashed line indicates $d = 250$) (left). Relationship between $\hat{\Delta}$ and d in the second numerical simulation (right)

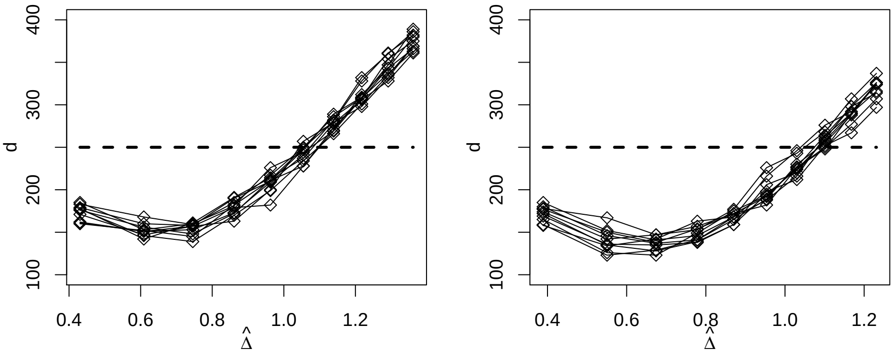

Fig. 8. Relationship between $\hat{\Delta}$ and d in the third numerical simulation (the dashed line indicates $d = 250$) (left). Relationship between $\hat{\Delta}$ and d in the fourth numerical simulation (right)

To examine whether $\hat{\Delta}$ (Eq.(5.2)) works as an index for selecting when to use the maximum likelihood estimator and a predictive estimator, four numerical simulations were conducted.

The first numerical simulation assumed that the number of data (n) was 21 and $\{x_i\} = \{-10, -9, -8, \ldots, 10\}$. Furthermore, $\alpha = 0$ and $\beta = 0.2$ were set, and $\{\epsilon_i\}$ were taken as realizations of $N(0.0, \sigma^2)$ (normal distribution with mean 0.0 and variance σ^2). σ^2 was set to one of $\{5, 10, 15, \ldots, 50\}$. By altering the initial value of the pseudo-random numbers, 500 sets of data were generated.

$\hat{\rho}_{boot}$ was calculated using Eq. (4.3) with $K = 200$. Using the results of these calculations, er_{boot} (Eq. (4.4)) was derived. er_{like} (Eq. (3.3)) was calculated using the maximum likelihood method. The results in Fig. 5(left) show the relationship between $\hat{\Delta}$ (Eq. (5.2)) and d. The value of d indicates the number of simulation data satisfying $er_{like} > er_{boot}$; the total number of simulation data was 500. For example, $d = 200$ means that 200 out of 500 sets of data indicate that the predictive estimator outperforms the maximum likelihood estimator.

The second numerical simulation used $n = 41$ and $\{x_i\} = \{-20, -19, -18, \ldots, 20\}$. Again, $\alpha = 0$ and $\beta = 0.2$ were set, and $\{\epsilon_i\}$ were taken as realizations of $N(0.0, \sigma^2)$. σ^2 was set to one of $\{40, 80, 120, \ldots, 400\}$. By altering the initial value of the pseudo-random numbers, 500 sets of data were generated. $\hat{\rho}_{boot}$ was calculated with $K = 200$. Using these results, er_{boot} (Eq. (4.4)) was calculated. er_{like} (Eq. (3.3)) was derived using the maximum likelihood method. The results in Fig. 5(right) show the relationship between $\hat{\Delta}$ (Eq. (5.2)) and d.

The third numerical simulation used $n = 31$ and $\{x_i\} = \{1^2 - \bar{\xi}, 2^2 - \bar{\xi}, 3^2 - \bar{\xi}, \ldots, 31^2 - \bar{\xi}\}$; $\bar{\xi} = \frac{1}{31}\sum_{i=1}^{31} i^2)$. We again set $\alpha = 0$ and $\beta = 0.2$ and took $\{\epsilon_i\}$ as realizations of $N(0.0, \sigma^2)$. σ^2 was set to one of $\{20,000, 40,000, 60,000, \ldots, 200,000\}$. By altering the initial value of the pseudo-random numbers, 500 sets of data were generated. $\hat{\rho}_{boot}$ was calculated with $K = 200$. Using these results, er_{boot} (Eq. (4.4)) was calculated. er_{like} (Eq. (3.3)) was derived using the maximum likelihood method. The results in Fig. 5(left) show the relationship between $\hat{\Delta}$ (Eq. (5.2)) and d.

The fourth numerical simulation used $n = 31$ and $\{x_i\} = \{\sqrt{1} - \bar{\xi}, \sqrt{2} - \bar{\xi}, \sqrt{3} - \bar{\xi}, \ldots, \sqrt{31} - \bar{\xi}\}$; $\bar{\xi} = \frac{1}{31}\sum_{i=1}^{31}\sqrt{i}$. Once again, $\alpha = 0$ and $\beta = 0.2$ were set, and $\{\epsilon_i\}$ were taken as realizations of $N(0.0, \sigma^2)$. σ^2 was set to one of $\{0.3, 0.6, 0.9, \ldots, 3\}$. By altering the initial value of the pseudo-random numbers, 500 sets of data were generated. $\hat{\rho}_{boot}$ was calculated with $K = 200$. Using these results, er_{boot} (Eq. (4.4)) was calculated. er_{like} (Eq. (3.3)) was derived using the maximum likelihood method. The results in Fig. 5(right) show the relationship between $\hat{\Delta}$ (Eq. (5.2)) and d.

Figs. 7 (left)(right) and 8(left)(right) show that $\hat{\Delta} = 1.05$ can be used as a rough threshold to determine the relative merit of the predictive estimator and the maximum likelihood estimator.

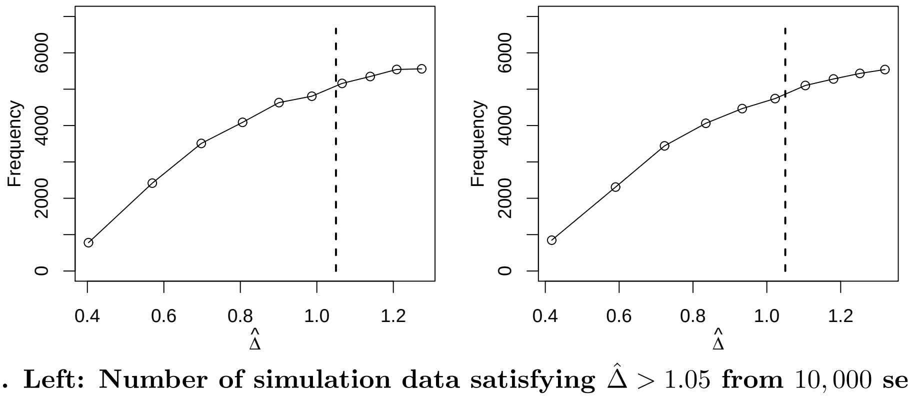

Fig. 9. Left: Number of simulation data satisfying $\hat{\Delta} > 1.05$ from $10,000$ sets of simulation data generated using the same conditions as Fig. 5(left). Right: Number of simulation data satisfying $\hat{\Delta} > 1.05$ from $10,000$ sets of simulation data generated using the same conditions as Fig. 5(right)

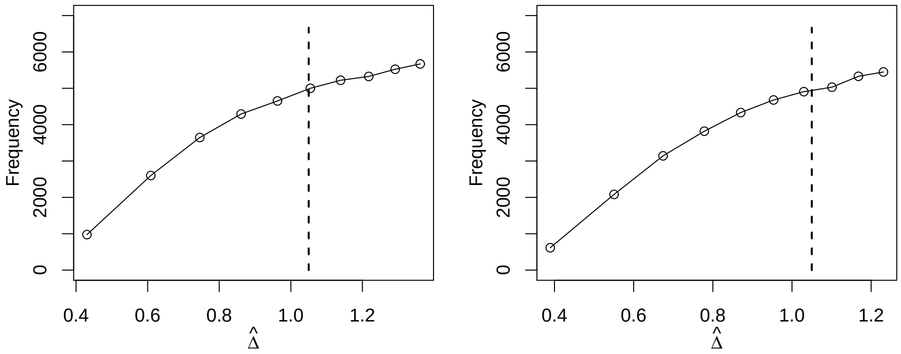

Fig. 10. Left: Number of simulation data satisfying $\hat{\Delta} > 1.05$ from $10,000$ sets of simulation data generated using the same conditions as Fig. 5(left). Right: Number of simulation data satisfying $\hat{\Delta} > 1.05$ from $10,000$ sets of simulation data generated using the same conditions as Fig. 5(right)

We then generated $10,000$ sets of simulation data under the same conditions as in the first numerical simulation. $\hat{\Delta}$ (Eq. (5.2)) was calculated for each dataset, and the number of datasets for which $\hat{\Delta} > 1.05$ was counted. The relationship between these counts and $\hat{\Delta}$ is shown in Fig. 9(left). When the simulation data have a population parameter satisfying $\hat{\Delta} < 1.05$, $\hat{\Delta} > 1.05$ holds in less than half of the datasets. Moreover, when the simulation data have a population parameter satisfying $\hat{\Delta} > 1.05$, $\hat{\Delta} > 1.05$ holds in more than half of the datasets.

Thus, when we use the predictive estimator and the maximum likelihood estimator on the basis of the magnitude relationship between $\hat{\Delta}$ and 1.05, appropriate use of the estimator is realized with a probability of greater than 0.5.

With $10,000$ sets of simulation data generated under the same conditions as in the second numerical simulation, we obtain Fig. 9(right).

Generating $10,000$ sets of simulation data under the same conditions as in the third and fourth numerical simulations gives Fig. 10 (left), (right), respectively.

These simulations show that we can use a predictive estimator or the maximum likelihood estimator according to the magnitude relationship between $\hat{\Delta}$ and 1.05.

6 Conclusions

To construct a predictive estimator, the parameter values of a function with data as its arguments are optimized in terms of prediction accuracy. This function may be, for example, the product of a constant and the maximum likelihood estimator. However, the resultant estimator is not always suitable for practical use because, if the parameters in the predictive estimator depend on the data, a large variance of the parameters would augment the variance of the predictive estimator. In the example used in this paper, ρ is the only parameter in the predictive estimator and the variance of ρ is not 0 because it depends on the data. Hence, the prediction error given by the predictive estimator could surpass that of the maximum likelihood estimator.

Note that parameter estimation of the exponential distribution is a particular case, and should not be generalized. That is, a possible predictive estimator for this setting is obtained by multiplying $\left(1 - \dfrac{1}{n}\right)$ (n is the number of data) by the maximum likelihood estimator [15]. The variance of this predictive estimator is obviously less than that of the maximum likelihood estimator. Therefore, this predictive estimator is superior to the maximum likelihood estimator in terms of prediction accuracy.

The example of the simple regression using GCV in section 3 shows that the variance given by the GCV-based method is large. Hence, in most cases, the prediction error produced by the predictive estimator with the resulting ρ is inferior to that given by the maximum likelihood estimator. This indicates that the specification of the form of a predictive estimator and the optimization of parameters in the estimator are not good enough for our purpose. Therefore, the prediction error of each estimator should be quantified by taking account of the variance of the resulting parameters. In light of this consideration, we should choose the best predictive estimator. Furthermore, if the predictive ability of the selected predictive estimator outperforms that of the maximum likelihood estimator, the use of the predictive estimator should be recommended.

Note that the relative merits of a predictive estimator and the maximum likelihood can depend upon the appearance of the data. Therefore, statistics such as $\hat{\Delta}$ (Eq.(5.2)) are needed to quantify the appearance of the data in order to choose between the predictive estimator and the maximum likelihood estimator. If we have more than one predictive estimator in mind, such statistics can be used to choose among these predictive estimators on the basis of the appearances of the data.

Moreover, we still do not have a theorem that gives the lower bound of variance for predictive estimators and the James–Stein estimator, whereas we have the Cramér–Rao inequality for unbiased estimators (e.g. page 181 in [16]). Hence, we cannot conclude that a specific predictive estimator is the best in terms of the variance or prediction error for the given data.

Therefore, even if a good predictive estimator for some specific data is known, we cannot deny the possibility that another predictive estimator would give superior predictions. Hence, a new predictive estimator could be better than the predictive estimators obtained so far in terms of prediction accuracy.

The numerical simulations reported in this paper show how this methodology works in simple regression. We have shown that, although the use of GCV is typical for simple regression from the perspective of prediction, this method does not always give a practical predictive estimator. Moreover, it has also been revealed that when the parametric bootstrap method is used, the relative merits of the maximum likelihood estimator and the predictive estimator depend upon the characteristics of the data. To determine when to use the maximum likelihood estimator or the predictive estimator, $\hat{\Delta}$ (Eq.(5.2)) is a promising tool. Better statistics, however, could lead to more appropriate use of the two estimators. Thus, we conclude that estimations which take account of predictive estimators will differ appreciably from the conventional estimations given by the maximum likelihood estimator and unbiased estimators.

Predictive estimators will be built for various statistical estimation scenarios. For this purpose, the characteristics of predictive estimators that reflect the features of the data should be investigated from all perspectives. This would allow practical predictive estimators to be selected and compared with the maximum likelihood estimator in terms of prediction accuracy. Estimations based on this strategy would clarify the overall quality of the available data, which traditional methods have not been able to do.

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

Author has declared that no competing interests exist..

References

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

[2] Takezawa K. Learning regression analysis by simulation. Springer; 2014.

[3] Ogasawara H. A family of the adjusted estimators maximizing the asymptotic predictive expected log-likelihood. Behaviormetrika. 2016;1-39.
DOI: 10.1007/s41237-016-0004-6

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

[5] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994.

[6] Craven P, Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik. 1979;31(4):377-403.

[7] Wahba G. Spline models for observational data (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial and Applied Mathematics; 1990.

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

[9] Myers RH. Classical and modern regression with applications (Duxbury Classic) 2 edition. Duxbury Press; 1990.

[10] Copas JB. Regression, prediction and shrinkage. Journal of the Royal Statistical Society. Series B. 1983;45(3):311-354.

[11] Copas JB. Cross-validation shrinkage of regression predictors. Journal of the Royal Statistical Society. Series B. 1987;49(2):175-183.

[12] Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematics, Statistics and Probability. University of California Press, Berkeley. 1956;1:197-206.

[13] James W, Stein C. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability. University of California Press, Berkeley. 1961;1:361-380.

[14] Chiu ST. Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. Biometrika. 1990;77:222-226.

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

[16] Bickel PJ, Doksum KA. Mathematical statistics: Basic ideas and selected topics, Vol I (2nd Edition). Prentice Hall; 2000.

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://sciencedomain.org/review-history/21049