Open Journal of Statistics, 2012, 2, 401-407
http://dx.doi.org/10.4236/ojs.2012.24048 Published Online October 2012 (http://www.SciRP.org/journal/ojs)

Flexible Model Selection Criterion for Multiple Regression

Kunio Takezawa
National Agriculture and Food Research Organization, Agricultural Research Center Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan
Email: nonpara@gmail.com

Received July 17, 2012; revised August 20, 2012; accepted August 31, 2012

ABSTRACT

Predictors of a multiple linear regression equation selected by *GCV* (Generalized Cross Validation) may contain undesirable predictors with no linear functional relationship with the target variable, but are chosen only by accident. This is because *GCV* estimates prediction error, but does not control the probability of selecting irrelevant predictors of the target variable. To take this possibility into account, a new statistics "GCV_f" ("*f*" stands for "flexible") is suggested. The rigidness in accepting predictors by GCV_f is adjustable; GCV_f is a natural generalization of *GCV*. For example, GCV_f is designed so that the possibility of erroneous identification of linear relationships is 5 percent when all predictors have no linear relationships with the target variable. Predictors of the multiple linear regression equation by this method are highly likely to have linear relationships with the target variable.

Keywords: *GCV*; GCV_f; Identification of Functional Relationship; Knowledge Discovery; Multiple Regression; Significance Level

1. Introduction

There are two categories of methods for selecting predictors of regression equations such as multiple linear regression. One includes methods using statistical tests such as the *F*-test. The other one includes methods of choosing predictors by optimizing statistics such as *GCV* or *AIC* (Akaike's Information Criterion). The former methods have a problem in that they examine only a part of multiple linear regression equations among many applicants of the predictors (e.g., p. 193 in Myers [1]). In this point, all possible regression procedures are desirable. It has spread the use of statistics such as *GCV* and *AIC* to produce multiple linear regression equations.

Studies of statistics such as *GCV* and *AIC* aim to construct multiple linear regression equations with a small prediction error in terms of residual sum of squares or log-likelihood. In addition, discussion on the practical use of multiple linear regression equations advances on the assumption of the existence of a linear relationship between the predictors adopted in a multiple linear regression equation and the target variables. However, we should consider the possibility that some predictors used in a multiple linear regression equation have no linear relationships with the target variable. If we cannot neglect the probability that some predictors with no linear relationships with the target variable reduce the prediction error by accident, there is some probability that one or more predictors with no linear relationships with the target variable may be selected among the many applicants of predictors. Hence, if our purpose is to select predictors with linear relationships with the target variable, we need a method different from those that choose a multiple linear regression equation yielding a small prediction error. We address this possibility in the following discussion.

We present an example that casts some doubt on the linear relationships between the predictors selected by *GCV* and the target variable in Section 2. In preparation to cope with this problem, in Section 3, we show the association between *GCV* (or *AIC*) and the *F*-test. In Section 4, on the basis of this insight, we suggest "GCV_f" ("*f*" stands for "flexible") to help solve this problem. Then, in Section 5, we propose a procedure for estimating the probability of the existence of linear relationships between the predictors and the target variable using GCV_f. Finally, we show the application of this method to the data which is used in Section 2.

2. Definition of the Problem Using Real Data

We use the first 50 sets of Boston house price data (named "Boston") retrieved from StatLib. These data consist of 14 variables. The applicants of the predictors ($\{x_1, x_2, x_3, x_4\}$) and the target variable (y) are selected among them:

x_1: per capita crime rate by town;
x_2: proportion of nonretail business acres per town;
x_3: average number of rooms per dwelling;

x_4: pupil-teacher ratio by town;

y: median value of owner-occupied homes in $1000's.

Figure 1 shows a matrix of scatter plots for showing the distributions of the above data. The correlations of the target variable with x_1 and x_3 appear to be high. The negative correlation between x_1 and y indicates that house prices in crime-ridden parts of the city tend to be low. The positive correlation between x_3 and y implies that house price is relatively high if the average number of rooms per household in an area is large. The result of the construction of a multiple linear regression equation using 50 datasets with all the predictors is shown below. The R command lm() installed by default was used for this purpose.

Coefficients:	Estimate	Std.Error	t value	$\Pr(>\|t\|)$
(Intercept)	−28.0049	10.6679	−2.625	0.01179*
x.1	−6.7668	1.6640	−4.067	0.00019***
x.2	−0.5825	0.3324	−1.752	0.08651
x.3	7.3779	1.1860	6.221	1.47e−07***
x.4	0.5784	0.3121	1.854	0.07037

Signif.codes: 0"***"0.001; "**"0.01; "*"0.05; "."0.1; " "1

Residual standard error: 2.841 on 45 degrees of freedom Multiple R-squared: 0.7957, Adjusted R-squared: 0.7776 F-statistic: 43.83 on 4 and 45 DF, p-value: 5.692e−15.

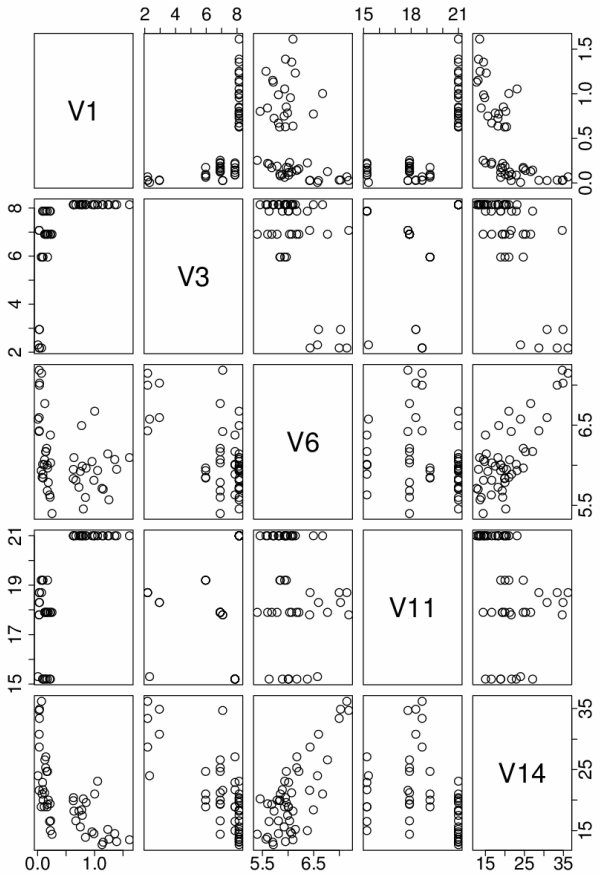

Figure 1. Matrix of scatter plots using four applicants of predictors and the target variable. The first 50 sets of Boston house price data (named "Boston") are used.

The above table shows that $\{x_1, x_3\}$ should be chosen as predictors if a 5 percent significant level is adopted in the t-test.

However, if predictors are not independent of each other, this result is not necessarily reliable. Then, all possible regression procedures using *GCV* were carried out to select predictors. *GCV* is defined as

$$GCV(q) = \frac{RSS(q)}{n\left(1-\frac{q+1}{n}\right)^2}, \tag{1}$$

where n is the number of data and q is the number of predictors. $RSS(q)$ is

$$RSS(q) = \sum_{i=1}^{n}(y_i - a_0)^2 \quad \text{if} \quad q = 0$$

$$RSS(q) = \sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{q} a_j x_{ij}\right)^2 \quad \text{if} \quad q \geq 1, \tag{2}$$

where $\{x_{ij}\}$ indicate the data of the selected predictors. $\{a_j\}$ are regression coefficients given by conducting the least squares using selected predictors. $\{y_i\}$ shows the data of the target variable. The above procedures were uses to select all predictors ($\{x_1, x_2, x_3, x_4\}$). Predictor selection by *GCV* results in a multiple linear regression equation that is expected to provide a small prediction error with the use of the regression equation for predictive purposes. Hence, since the multiple linear regression equation using $\{x_1, x_2, x_3, x_4\}$ is of great use for prediction, we are inclined to think that each of the predictors $\{x_1, x_2, x_3, x_4\}$ has a linear relationship with y.

To determine whether this is correct, the data ($\{y_i\}$) of the target variable (y) are randomly resampled with replacement to obtain n data $\left(\left\{y_i^B\right\}(1 \leq i \leq 50)\right)$. The data of the target variable remain unchanged. The procedure was repeated 500 times while varying the seed of the pseudo-random number generator. This procedure provided 500 sets of bootstrapped data. The values of the target variable of these bootstrapped data are provided by random sampling from a population with a distribution given by the data of the target variable; hence, they are not associated with the predictors. Therefore, if predictors are selected using these bootstrapped data, a constant seems to be almost always chosen as the best regression equation.

The result is shown in **Figure 2** where the frequencies of the number of selected predictors are illustrated. A constant is selected as the best regression equation in only 267 of the 500 data sets. This result shows that even

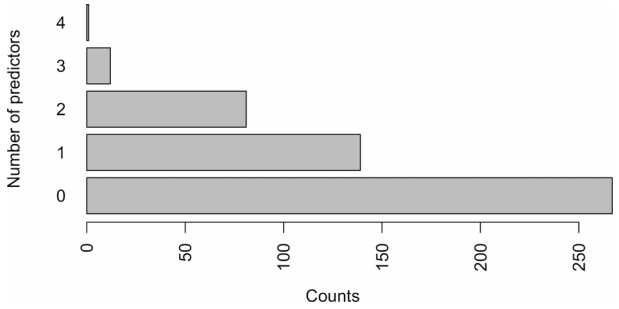

Figure 2. Frequencies of the number of predictors selected by *GCV*.

if the data do not have linear relationships between the predictors and the target variable, a functional relationship represented by a multiple linear regression equation or a simple regression equation is found at about 50 percent probability.

Therefore, if the predictors are chosen by all possible regression procedures using statistics such as *GCV*, we should not rule out the possibility that they contain one or more predictors with no linear relationships with the target variable. This implies that we need a new model selection criterion. This new criterion should choose predictors only if the predictors are highly likely to have linear relationships with the target variable.

3. Relationship between Model Selection Criterion and *F*-Test

$F(n, q)$ (F values) is defined as

$$F(n,q) = \frac{\dfrac{RSS(q-1) - RSS(q)}{1}}{\dfrac{RSS(q)}{n-q-1}} = (n-q-1)\left(\frac{RSS(q-1)}{RSS(q)} - 1\right). \tag{3}$$

Hence, we have

$$\frac{RSS(q-1)}{RSS(q)} = \frac{F(n,q)}{n-q-1} + 1. \tag{4}$$

Furthermore, Equation (1) leads to

$$\frac{GCV(q)}{GCV(q-1)} = \frac{RSS(q)\cdot\left(1-\dfrac{q}{n}\right)^2}{RSS(q-1)\cdot\left(1-\dfrac{q+1}{n}\right)^2} \tag{5}$$

The substitution of Equation (4) gives

$$\frac{GCV(q)}{GCV(q-1)} = \left(\frac{F(n,q)}{n-q-1} + 1\right)^{-1} \frac{(n-q)^2}{(n-q-1)^2}. \tag{6}$$

Therefore, when we have a multiple linear regression equation with $(q-1)$ predictors, the condition for accepting a q-th predictor is written as

$$\left(\frac{F(n,q)}{n-q-1} + 1\right)^{-1} \frac{(n-q)^2}{(n-q-1)^2} < 1. \tag{7}$$

That is,

$$F(n,q) > \frac{(n-q)^2}{n-q-1} - (n-q-1). \tag{8}$$

If the inequality sign in the above equation is replaced with an equality sign and $n = 25$, $F(n, q)$ is shown as in **Figure 3** (left panel). This shows that when we use *GCV*, $F(n, q)$ for determining whether the q-th predictor should be added to the multiple linear regression equation with $(q-1)$ predictors is nearly independent of q.

If the multiple linear regression equation with $(q-1)$ predictors is correct, $F(n, q)$ is written as

$$F(n,q) = \frac{\dfrac{R^2(q) - R^2(q-1)}{1-R^2(q)}}{n-q-1} \sim F_{1,n-q-1}. \tag{9}$$

$F_{1,n-q-1}$ stands for the F distribution; the first degree of freedom is 1 and the second degree of freedom is $(n-q-1)$. $R^2(q)$ is the coefficient of determination defined as

$$R^2(q) = \frac{\displaystyle\sum_{i=1}^{n}\left(\hat{y}^{(q)} - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2}{\displaystyle\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2}, \tag{10}$$

where $\left\{\hat{y}_i^{(q)}\right\}$ are estimates derived using a multiple linear regression equation with q predictors. Then, we calculate p that satisfies the equation

$$p = \int_{F(n,q)}^{\infty} den(1, n-q-1, x)\,dx, \tag{11}$$

where $den(1, n-q-1, x)$ is the probability density function of an F distribution; the first degree of freedom is 1 and the second degree of freedom is $(n-q-1)$. p is the value of the integral. The lower limit of the integration of the probability density function with respect to x is $F(n, q, p)$. This p represents the probability that F is larger than $F(n, q)$ when the multiple linear regression function with $(q-1)$ predictors is a true one. Hence, the values of $F(n, q)$ drawn in **Figure 3** (left panel) are substituted into Equation (11); the resultant values of p are shown in **Figure 4** (left panel). These values of p are the probability that the q-th predictor is wrongly accepted when the multiple linear regression equation with the $(q-1)$ predictors is correct. That is, this is the probability of a type one error. When the forward and backward selection

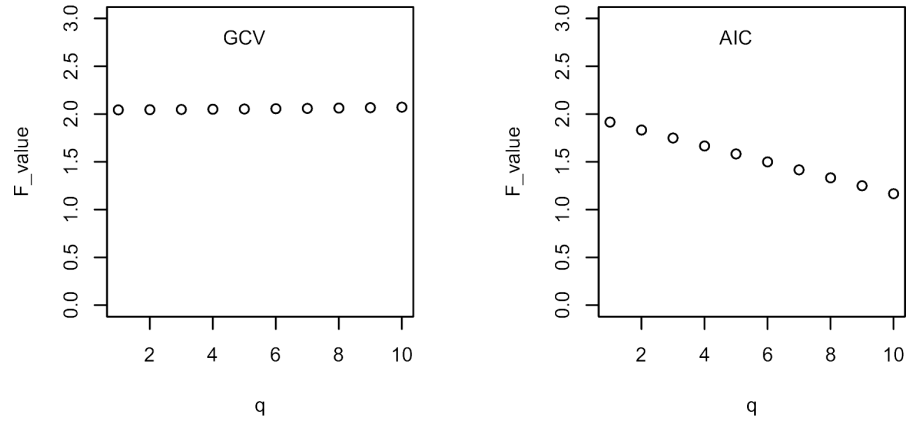

Figure 3. Relationship between *q* and *F*(25, *q*) corresponding to *GCV* and *AIC*.

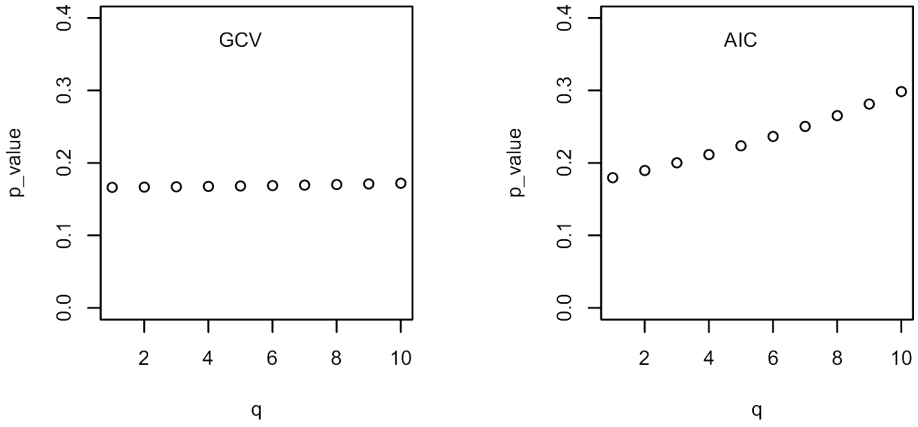

Figure 4. Relationship between *q* and *p* corresponding to *GCV* and *AIC* (*n* = 25).

method using *F* value is carried out, this probability is fixed at values ranging from 0.25 to 0.5 (e.g., p. 188 in Myers [1]), or 0.05 (e.g., p. 314 in Montgomery [2]). Therefore, the selection method for predictors by *GCV* has similar features with the forward and backward selection method with a fixed *p* because *p* in **Figure 4** (left panel) is nearly independent of *q*.

On the other hand, the forward and backward selection method does not compare the multiple linear regression equation with predictors of $\{x_1, x_2\}$ with that with predictors of $\{x_3, x_4\}$ for example. This type of comparison can be performed by *GCV*. All possible regression procedures using *GCV* entail such a comparison. Hence, the comparison of two multiple linear regression equations in the forward and backward selection method should be on par with that of the same multiple linear regression equations by all possible regression procedures.

On the other hand, *AIC* is defined as

$$AIC(q) = n\log(2\pi) + n\log\left(\frac{RSS(q)}{n}\right) + n + 2q + 4. \quad (12)$$

Hence, we have

$$AIC(q) - AIC(q-1) = n\log\left(\frac{RSS(q)}{n}\right) - n\log\left(\frac{RSS(q-1)}{n}\right) + 2 = n\log\left(\frac{RSS(q)}{RSS(q-1)}\right) + 2. \quad (13)$$

The substitution of Equation (4) leads to

$$AIC(q) - AIC(q-1) = n\log\left(\left(\frac{F(n,q)}{n-q-1}+1\right)^{-1}\right) + 2 < 0. \quad (14)$$

Therefore, if we have a multiple linear regression equation with $(q-1)$ predictors, the condition for accepting a *q*-th predictor is

$$\left(\frac{F(n,q)}{n-q-1}+1\right)^{-1} < \exp\left(\frac{-2}{n}\right). \quad (15)$$

That is,

$$F(n,q) > (n-q-1)\left(\exp\left(\frac{2}{n}\right)-1\right). \quad (16)$$

when the inequality sign in this equation is replaced with an equality sign, $F(n, q)$ (n = 25) is drawn in **Figure 3**

(left panel). The corresponding p is shown in **Figure 4** (right panel). It shows the characteristics of *AIC* which show that, when we have a multiple linear regression equation with $(q-1)$ predictors, p for determining whether to accept a q-th predictor augments with an increase in q. This is consistent with the tendency that *AIC* accepts a new predictor with comparative ease when the present multiple linear regression equation has many predictors.

4. Introduction of GCV_f

In the previous section, we associate *GCV* and *AIC* with the forward and backward selection method using F. This indicates that *GCV* is desirable as long as p is nearly independent of q. However, if p corresponding to a model selection criterion should be independent of q, we may well develop a new model selection criterion that meets the requirement. Then, if p is given, $F(n, q, p)$ is calculated using

$$p = \int_{F(n,q,p)}^{\infty} den(1, n-q-1, x)\,\mathrm{d}x. \tag{17}$$

The difference between Equations (17) and (11) is that Equation (11) is used to obtain p when $F(n, q)$ is given, whereas Equation (17) works as an equation for calculating $F(n, q, p)$ when p is in hand. Equation (4) indicates that the multiple linear regression equation with $(q-1)$ predictors accepts the q-th predictor when the following equation is satisfied:

$$\frac{RSS(q-1)}{RSS(q)} > \frac{F(n,q,p)}{n-q-1}+1. \tag{18}$$

Therefore, we suggest the following $GCV_f(q)$ as a new model selection criterion:

$$GCV_f(q) = \frac{RSS(q)}{n+2} \text{ if } q = 0$$

$$GCV_f(q) = \frac{RSS(q)}{n+2}\prod_{k=1}^{q}\left(\frac{F(n,k,p)}{n-k-1}+1\right) \text{ if } q \geq 1. \tag{19}$$

This criterion is justified because it is a criterion in which a multiple linear regression equation with $(q-1)$ predictors accepting a q-th predictor is depicted as Equation (18) ($q \geq 1$). Hence, $GCV_f(q)$, a new model selection criterion, is the same in function to the forward and backward selection method using the F-test with a p significant level in determining whether or not a q-th predictor is accepted. $GCV_f(q)$ stands for "flexible *GCV*". In discussion on model selection, the focus is on choosing between *GCV* and *AIC*, for example. However, $GCV_f(q)$ allows us to adjust the characteristics of the model selection criterion continuously by varying p. Therefore, if p is tuned according to the nature of the data or to the purpose of the regression, we have an appropriate model selection criterion. The background that made us think of $GCV_f(q)$ as a flexible version of the conventional $GCV(q)$ is as follows.

Let the coefficient of $\frac{RSS(q)}{n}$ in $GCV(q)$ (Equation (1)) be $CGCV(q)$. Then, we have

$$CGCV(q) = \frac{1}{\left(1-\frac{q+1}{n}\right)^2}. \tag{20}$$

Let the coefficient of $\frac{RSS(q)}{n}$ in $GCV_f(q)$ (Equation (19)) be $CGCV_f(q)$. Then, we have

$$CGCV_f(q) = \frac{n+2}{n}\prod_{k=1}^{q}\left(\frac{F(n,k,p)}{n-k-1}+1\right). \tag{21}$$

when n is large, the equation below holds:

$$F(n,q,0.1573) \approx F(\infty,q,0.1573) \approx 2.0. \tag{22}$$

Thus, if $q \geq 1$ and $p = 0.1573$ are satisfied, we obtain

$$\begin{aligned} CGCV(q) &= \frac{1}{\left(1-\frac{q+1}{n}\right)^2} \approx \frac{1}{\left(1-\frac{2q+2}{n}\right)} \\ &\approx 1+\frac{2q+2}{n}, \end{aligned} \tag{23}$$

$$\begin{aligned} CGCV_f(q) &= \frac{n+2}{n}\prod_{k=1}^{q}\left(\frac{F(n,k,0.1573)}{n-k-1}+1\right) \\ &\approx \frac{n+2}{n}\prod_{k=1}^{q}\left(\frac{2}{n-k-1}+1\right) \\ &\approx \left(1+\frac{2}{n}\right)\left(1+\frac{2q}{n}\right) \approx 1+\frac{2q+2}{n}. \end{aligned} \tag{24}$$

If $q = 1$, we have

$$CGCV(1) = \frac{1}{\left(1-\frac{1}{n}\right)^2} \approx 1+\frac{2}{n}. \tag{25}$$

$$CGCV_f = 1+\frac{2}{n}. \tag{26}$$

Because of Equations (23)-(26), $GCV_f(q)$ is approximately identical to $GCV(q)$ when $p = 0.1573$.

$CGCV(q)$ and $CGCV_f(q)$ when $n = 100$ are set are shown in **Figure 5**. It demonstrates that $GCV_f(q)$ when $p = 0.1573$ is set are approximately identical to those of $GCV(q)$. Furthermore, a small p is assumed in $CGCV_f(q)$; $CGCV_f(q)$ is large. This indicates that $GCV_f(q)$ with a small p selects a multiple linear regression equation with

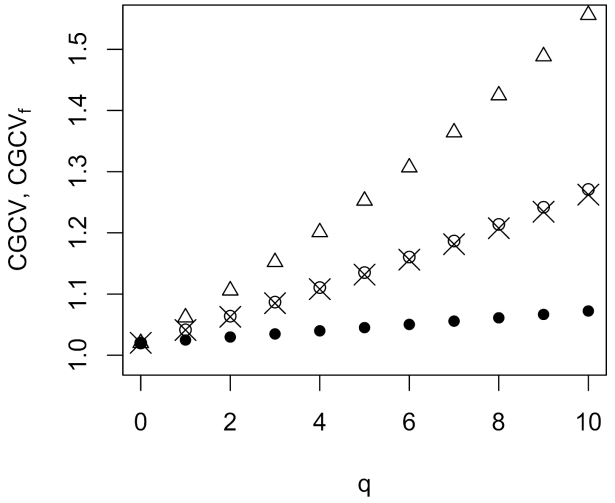

Figure 5. ***CGCV*(*q*) and *CGCV*$_f$*(q)*. " × " *CGCV*(*q*), "Δ" *CGCV*$_f$(*q*) (*p* = 0.05), "○" *CGCV*$_f$(*q*) (*p* = 0.1573), "•" *CGCV*$_f$(*q*) (*p* = 0.5).**

a small number of predictors.

5. Identification of Linear Functional Relationships Using GCV_f

The discussion in the previous section shows that $GCV_f(q)$ enables us to adjust the rigidness of accepting new predictors. By taking advantage of this feature, a method of including predictors with a clear causal connection is developed. This method is performed as follows:

1) Let the data of the target variable be $\{y_i\}(1 \le i \le n)$. n data are randomly resampled with replacement from $\{y_i\}$. Then, we have $\left(\{y_i^B\}(1 \le i \le n)\right)$. The procedure is repeated 500 times by varying the seed of the pseudo-random number generator. The data of predictors remain unchanged. Thus, we have 500 sets of bootstrapped data.

2) Using various values of p, predictors are selected by $GCV_f(q)$; this process is carried out for 500 sets of bootstrapped data. We choose p by which the probability of obtaining regression equations except a constant one is approximately 0.05.

3) Using p selected in (2), model selection by $GCV_f(q)$ is carried out for the original data.

This method generates sets of bootstrapped data of which the data of the target variable are resampled ones; hence, there is no causal connection at all between the data of the predictors and those of the target variable. This is because the values of the target variable are sampled from a population with a distribution given by the data of the target variable; the values of the target variable are not associated with those of the predictors. Although the predictor variables are selected using these bootstrapped data, regression equations except a constant may be produced with considerable probability. This indicates that the model selection criterion is very likely to accept predictors. Therefore, we should find p to make this probability approximately 5 percent. When the model selection is carried out using $GCV_f(q)$ given by the optimized p, a regression equations except a constant will be selected at a 5 percent probability when a constant should be chosen. This strategy quells our suspicion that a constant might be actually desirable even though regression equations except a constant were selected. This method is similar to Generalized Cross-validation Test (p. 87, in Wang [3]) in which the Monte Carlo method is carried out to test whether a regression equation should be parametric such as a simple regression equation.

Next, model selection was carried out for the data used in Section 2. GCV_f with various values of p was used for choosing predictors that are highly likely to have linear functional relationships with the target variable; the data of the target variable were bootstrapped. **Table 1** shows the results using the settings of p = 0.01, p = 0.06, p = 0.05 and p = 0.04. When GCV_f with p = 0.05 is employed, a constant was selected in 446 sets. Hence, if a model selection method adopts GCV_f with p = 0.05 and multiple linear regression equation except a constant are chosen, we reject the null hypothesis at a 5 percent significance level: there are no linear functional relationships between the predictors and the target variable. Then, a model selection by all possible regression procedures was carried out using GCV_f with p = 0.05. GCV_f is minimized when $\{x_1, x_3, x_4\}$ were chosen. On the other hand, a model selection by all possible regression procedures using GCV chose $\{x_1, x_2, x_3, x_4\}$ in Section 2. In view of **Figure 2**, this result follows our intuition that one or more predictors among the four selected ones may not have linear functional relationships with the target variable.

However, if the data are slightly altered, GCV_f with p = 0.05 may choose different predictors from $\{x_1, x_3, x_4\}$. If this possibility is correct, the selection of $\{x_1, x_3, x_4\}$ is not valid. To clarify this point, a bootstrap method in the usual sense is conducted for these data; 500 sets of bootstrapped data are generated. That is, the data set of $\{(x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_i)\}$ $(1 \le i \le 50)$ was randomly resampled with replacement whereas the set of values of the predictors

Table 1. Frequencies of the number of selected predictors.

Number of predictors	p = 0.1	p = 0.06	p = 0.05	p = 0.05
0	404	446	453	461
1	66	44	37	31
2	29	10	10	8
3	1	0	0	0
4	0	0	0	0

and the target variable was wrapped. When model selection by GCV_f with $p = 0.05$ was carried out for 500 datasets, we obtained **Table 2**. $\{x_1, x_3, x_4\}$ was chosen for 166 datasets. On the other hand, $\{x_1, x_2, x_3\}$ was selected for 157 datasets. Therefore, $\{x_1, x_3, x_4\}$ is not the only choice as a set of predictors with linear relationships with the target variable. $\{x_1, x_2, x_3\}$ is also a possible choice when we proceed with the discussion on this data.

6. Conclusions

We have assumed that when *GCV* or *AIC* yields a multiple linear regression equation with a small prediction error, there is a linear functional relationship between the predictors employed in the regression equation and the target variable. Not much attention has been paid to the probability that one or more selected predictors actually have no linear functional relationships with the target variable. However, we should not ignore the possibility that when several predictors with no linear functional relationships with the target variable are contained in the applicants of the predictors, one or more such predictors are adopted as appropriate predictors in a multiple linear regression equation. This is because when many applicants of the predictors have no linear relationships with the target variable, one or more such predictors will be selected at a high probability, since p in **Figure 4** does not depend on the number of applicants of the predictors.

Hence, another statistics for model selection based on an approach different from the use of prediction error is required for choosing predictors with linear relationships with the target variable. The new statistics should make the threshold high for accepting predictors when quite a few predictors have no linear functional relationship with the target variable. Although this strategy poses a relatively high risk of rejecting predictors that actually have linear relationships with the target variable, we have to accept this trade-off. This policy is quite similar to that of multiple comparison in which we accept the comparatively high risk of detecting no difference when there is actually a difference with the purpose of reducing the risk of mistakenly finding a difference when there is no difference.

Table 2. Frequencies of selected predictors.

Predictor	Frequency	Predictor	Frequency
$\{x_1, x_3, x_4\}$	166	$\{x_2, x_3\}$	14
$\{x_1, x_2, x_3\}$	157	$\{x_2, x_3, x_4\}$	4
$\{x_1, x_3\}$	87	$\{x_1, x_2\}$	1
$\{x_1, x_2, x_3, x_4\}$	71		

Using the statistics of GCV_f suggested here, we select one or more predictors at a 0.05 probability when no predictors have linear relationships with the target variable. If we select predictors using this new statistics, the chosen predictors are less likely to contain those that have no linear relationships with the target variable.

However, there is still room for further study of the detailed characteristics of GCV_f produced by the procedure presented here. In particular, we should know the behavior of GCV_f when there are high correlations between predictors.

The discussion so far indicates that the criteria for selecting predictors of a multiple linear regression equation are classified into two categories: one aims to minimize prediction error and the other is designed to select predictors with a high probability of having linear relationships with the target variable. *GCV* and GCV_f are examples of both categories, respectively. Interest has been focused on the derivation of multiple linear regression equations yielded using a criterion of prediction error. We expect that more attention will be paid to the probability of the existence of linear relationships. Furthermore, we should study whether a similar discussion is possible with respect to regression equations different from the multiple linear regression equation.

REFERENCES

[1] R. H. Myers, "Classical and Modern Regression with Applications (Duxbury Classic)," 2nd Edition, Duxbury Press, Pacific Grove, 2000.

[2] D. C. Montgomery, E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," 3rd Edition, Wiley, New York, 2001.

[3] Y. Wang, "Smoothing Splines: Methods and Applications," Chapman & Hall/CRC, Boca Raton, 2011. doi:10.1201/b10954

