TR-No. 22-03, Hiroshima Statistical Research Group, 1–19

Confidence Intervals in Multiple Linear Regression Conditioned on the Selected Model via the Kick-One-Out Method

Kyohei Mochizuki[1†] and Hirokazu Yanagihara[2*]

[1]Department of Mathematics, Graduate School of Science, Hiroshima University,
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan
[2]Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University,
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan

(Last Modified: June 5, 2022)

Abstract

This paper deals with the construction of a post-selection confidence interval for a regression coefficient via the kick-one-out (KOO) method in multiple linear regression. We derive a confidence interval for a regression coefficient with $1 - \alpha$ coverage conditioned on the selection event whereby the specific model is selected by the KOO method. In the KOO method, it is necessary to establish a discriminant function that is a difference of variable selection criteria when deciding whether to select a particular variable. In this paper, by deriving a general expression for the discriminant function, we systematically construct confidence intervals conditioned on the selection event via the KOO method when more than one variable selection criteria are used. Our results consider the case of the KOO method when various well-known variable selection criteria such as the AIC, BIC, and C_p criterion are employed.

Key words: GIC, GC_p criterion, Linear regression model, Post-selection inference, Variable selection.

*Corresponding author
E-mail address: yanagi-hiro@hiroshima-u.ac.jp (Hirokazu Yanagihara)

1. Introduction

Consider the following situation: Let $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ be an n-dimensional vector of response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top$ be an $n \times k$ matrix of non-stochastic k explanatory variables, where n is the sample size, and $^\top$ denotes the transpose of a matrix or vector. The multiple linear regression model with an assumption of normality can be expressed as

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

[†]Current Address: Innovation Technology Laboratories, AGC Inc, 1-1 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ is a k-dimensional vector of unknown regression coefficients, σ^2 is an unknown variance, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Here, in order to ensure a possibility of estimating the model, we assume that $\mathrm{rank}(\boldsymbol{X}) = k \ (< n)$.

With regard to the interpretability of the model and its prediction accuracy, it is not known whether the model that includes all of the potential explanatory variables is best; thus, it is necessary to search for the optimal subset of explanatory variables. Given k candidate variables, 2^{k-1} subsets are possible. The variable selection task, then, is to choose the optimal subset of explanatory variables among the 2^{k-1} candidates. One variable selection method, commonly referred to as the "full search method", would have us search the subset of explanatory variables minimizing the variable selection criterion among all candidate subsets of explanatory variables. However, since the computation of 2^{k-1} variable selection criterion values is required with this approach, the full search method is not feasible in terms of computational complexity when the number of explanatory variables is large. The kick-one-out (KOO) method proposed by Zhao, Krishnaiah and Bai (1986) and named by Bai, Fujikoshi and Hu (2018) is a simpler method than the forward-backward stepwise method and offers a variable selection approach that dramatically reduces computational complexity.

In the general flow of data analysis, after the best subset of variable is selected, confidence intervals of the regression coefficients are constructed for each selected variable. We derive a confidence interval for the regression coefficient of the jth explanatory variable in the model with $1-\alpha$ coverage when the model is the product of a specific selection method. Lee *et al.* (2016) constructed such confidence intervals when the specific model is selected by the lasso (Tibshirani, 1996). Charkhi and Claeskens (2017) constructed confidence intervals when the specific model is selected by the full search method using Akaike's information criterion (AIC) proposed by Akaike (1973; 1974). The purpose of this paper is to derive conditional confidence intervals when the model is selected by the KOO method, following the idea of the derivation in Charkhi and Claeskens (2017). In particular, we derive a general expression for the discriminant function to decide whether or not to select each variable in the KOO method when more than one variable selection criteria are being applied, including the AIC, BIC (Bayesian information criterion) (Schwarz, 1978), and C_p criterion (Mallows, 1973; 1995).

The paper is organized as follows: In Section 2, we describe the discriminant function for the KOO method when various variable selection criteria are used and show a general expression for the discriminant function in the KOO method. Importantly, we show that the statistic in each discriminant function is exactly the same even when a different variable selection criterion is used and establish that only the threshold determining whether a variable is selected changes as a function of the variable selection criterion being applied. In Section 3, we construct the target confidence interval by following the approach of Charkhi and Claeskens (2017). The proposed confidence interval consists of a truncated normal distribution, where the truncated region of the truncated normal distribution is expressed systematically using a general form of the discriminant function. We also show that the truncated region depends on the selected variables but not on any of the unselected variables. In Section 4, we report the results of numerical experiments comparing the coverage probabilities of the proposed confidence interval constructions with those obtained by a naive method that does not consider post-selection inference. In addition, we investigate how the proposed confidence intervals behave when the threshold in the discriminant function is increased. Technical details are provided in the Appendix.

2. General Expression of the Discriminant Function in the KOO Method

Let M be a subset of $\Omega = \{1, \ldots, k\}$, i.e., $M \subseteq \Omega$, which indicates the variables used in the model. Accordingly, in this paper, the model will be identified by the set M. Let $\boldsymbol{x}_{(j)}$ be an n-dimensional vector of the jth column of $\boldsymbol{X}$, i.e., $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$, and let $\boldsymbol{X}_M$ be an $n \times k_M$ matrix consisting of columns of $\boldsymbol{X}$ corresponding to the elements of M, where k_M denotes the number of elements of M, i.e., $k_M = \#(M)$. As an example of $\boldsymbol{X}_M$, if $M = \{2, 4, 6\}$, $\boldsymbol{X}_M = (\boldsymbol{x}_{(2)}, \boldsymbol{x}_{(4)}, \boldsymbol{x}_{(6)})$ and $k_M = 3$. In particular, we note that $\boldsymbol{X}_\Omega = \boldsymbol{X}$ and $k_\Omega = k$. Using $\boldsymbol{X}_M$, the candidate model can be expressed as follows:

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}_M \boldsymbol{\beta}_M, \sigma^2 \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{\beta}_M$ is the k_M-dimensional vector of regression coefficients given by minimizing the mean square error, that is,

$$\boldsymbol{\beta}_M = \arg\min_{\boldsymbol{b} \in \mathbb{R}^{k_M}} \mathbb{E}\left[||\boldsymbol{y} - \boldsymbol{X}_M \boldsymbol{b}||^2 \right] = \boldsymbol{X}_M^+ \boldsymbol{\mu},$$

where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}]$ and $\boldsymbol{X}_M^+$ is the Moore-Penrose inverse matrix of $\boldsymbol{X}_M$, i.e., $\boldsymbol{X}_M^+ = (\boldsymbol{X}_M^\top \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^\top$ (for details of the Moore-Penrose inverse matrix, see, e.g., Harville, 1997, chap. 20). Here, we call (1) model M. It is well known that the estimator of $\boldsymbol{\beta}_M$ can be obtained by minimizing the sum of squared residuals

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{b} \in \mathbb{R}^{k_M}} \frac{1}{n} ||\boldsymbol{y} - \boldsymbol{X}_M \boldsymbol{b}||^2 = \boldsymbol{X}_M^+ \boldsymbol{y}. \tag{2}$$

Let $\mathrm{SC}(M)$ denote a variable selection criterion for the model M. In the KOO method with $\mathrm{SC}(M)$, the ith variable is selected if $\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) > 0$ holds, where Ω_i is a model in which only the ith variable is removed from Ω, i.e., $\Omega_i = \Omega \backslash \{i\}$. Therefore, in this paper, $\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) > 0$ is called the discriminant function of the ith variable, or the ith discriminant function. The model selected via the KOO method using $\mathrm{SC}(M)$ is formulated by the discriminant functions as

$$\hat{M} = \{i \in \Omega \mid \mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) > 0\}. \tag{3}$$

Let $\boldsymbol{P}_M$ be the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_M$, i.e., $\boldsymbol{P}_M = \boldsymbol{X}_M \boldsymbol{X}_M^+$, and let s_M^2 be an unbiased estimator of the variance σ^2 under the model M defined by

$$s_M^2 = \frac{1}{n - k_M} \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_M) \boldsymbol{y}.$$

In this paper, the following six variable selection criteria are used in the KOO method:

$$GC_p(M|\gamma) = nc_M \frac{s_M^2}{s_\Omega^2} + \gamma k_M, \qquad \mathrm{GIC}(M|\gamma) = n \log\left(c_M s_M^2\right) + \gamma k_M,$$

$$\mathrm{GBNP}(M|\gamma) = -\frac{c_\Omega^2 s_\Omega^2}{c_M s_M^2} + \gamma k_M, \qquad \mathrm{GCV}_C(M|\gamma) = \log\left\{nc_M\left(c_M - \frac{\gamma}{n}\right)^{-2} s_M^2\right\},$$

$$\mathrm{AIC}_C(M) = n\log\left(c_M s_M^2\right) + \frac{2n(k_M+1)}{n-k_M-2}, \quad \mathrm{EGCV}(M|\gamma) = \log\left(nc_M^{1-\gamma} s_M^2\right),$$

where c_M is a constant depending on k_M and n, i.e., $c_M = 1 - k_M/n$, and γ is a positive parameter to adjust the strength of the penalty for increasing the number of variables. The GC_p criterion was first proposed by Atkinson (1980), GIC was proposed by Nishii (1984), the GCV_C criterion was proposed by Boonstra, Mukherjee and Taylor (2015), AIC_C was proposed by Sugiura (1978), Hurvich & Tsai (1989), and the EGCV criterion was proposed by Ohishi, Yanagihara and Fujikoshi (2020). For computational convenience, the GCV_C and EGCV criteria are expressed as logarithmic transformations of the original criteria. The GBNP criterion (in particular, the GBNP criterion with $\gamma = 2$ is called the BNP criterion) is a variable selection criterion measuring the goodness of fit of the model using a Bartlett-Nanda-Pillai-type test statistic, and is included in the class of variable selection criteria proposed by Ohishi (2021). The GC_p criterion can be regarded as a variable selection criterion measuring the goodness of fit of the model using the Lawley-Hoteling-type test statistic. GIC is essentially a variable selection criterion measuring the goodness of fit of a model using a likelihood ratio-type test statistic. In GIC and the GC_p, GCV_C and EGCV criteria, the various existing variable selection criteria can be expressed by changing γ. The relationship between these criteria and existing variable selection criteria is as follows:

$$GC_p:\ \gamma = \begin{cases} 2 & (C_p;\ \text{Mallows, 1973, 1995}) \\ 2\left(1 - \dfrac{2}{nc_M}\right)^{-1} & (MC_p;\ \text{Fujikoshi \& Satoh, 1997}) \end{cases},$$

$$\mathrm{GIC}:\ \gamma = \begin{cases} 2 & (\text{AIC; Akaike, 1973, 1974}) \\ \log n & (\text{BIC; Schwarz, 1978}) \\ 2\log\log n & (\text{HQC; Hannan \& Quinn, 1979}) \\ 1 + \log n & (\text{CAIC; Bozdogan, 1987}) \end{cases},$$

$$\mathrm{GCV}_C:\ \gamma = 0, \quad \mathrm{EGCV}:\ \gamma = 2, \quad (\text{GCV; Craven \& Wahba, 1979}).$$

It follows from equation (3.4) in Oda and Yanagihara (2020) that

$$\boldsymbol{P}_\Omega - \boldsymbol{P}_{\Omega_i} = \boldsymbol{r}_i \boldsymbol{r}_i^\top, \quad \boldsymbol{r}_i = \frac{1}{\|(\boldsymbol{X}^+)^\top \boldsymbol{e}_i\|}(\boldsymbol{X}^+)^\top \boldsymbol{e}_i \quad (i = 1, \ldots, k), \tag{4}$$

where $\boldsymbol{e}_i$ is a k-dimensional vector in which only the ith element is 1 and the other elements are 0. Equation (4) implies that

$$\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\Omega_i})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y} + \left(\boldsymbol{y}^\top \boldsymbol{r}_i\right)^2. \tag{5}$$

Let T_i be a statistic defined by

$$T_i = \frac{(\boldsymbol{y}^\top \boldsymbol{r}_i)^2}{s_\Omega^2}. \tag{6}$$

Using equations (5) and (6), the ith discriminant function $\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega)$ is calculated as follows (the proof is given in Appendix A.1):

$$\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) = \begin{cases} T_i - \gamma & (GC_p), \\ n \log\left(1 + \dfrac{T_i}{n-k}\right) - \gamma, & (\mathrm{GIC}), \\ T_i \left(1 + \dfrac{T_i}{n-k}\right)^{-1} - \gamma & (\mathrm{GBNP}), \\ \log\left(1 + \dfrac{T_i}{n-k}\right) - 2\log\left(\dfrac{n-k-\gamma+1}{n-k-\gamma}\right) & (\mathrm{GCV}_C), \\ n \log\left(1 + \dfrac{T_i}{n-k}\right) - \dfrac{2n(n-1)}{(n-k-1)(n-k-2)} & (\mathrm{AIC}_C), \\ \log\left(1 + \dfrac{T_i}{n-k}\right) - \gamma \log\left(\dfrac{n-k+1}{n-k}\right) & (\mathrm{EGCV}). \end{cases} \tag{7}$$

Equation (7) indicates that for any variable selection criterion, the first term of the discriminant function of the ith variable is a function of $(\boldsymbol{y}^\top \boldsymbol{r}_i)^2 / s_\Omega^2$ and the second term is a function of γ. Thus, using strictly monotonically increasing functions whose domain of definition is non-negative f and g (for GCV_C, the domain of definition of g is $[0, n-k)$), the discriminant function can be expressed, in general, as

$$\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) = f(T_i) - g(\gamma).$$

Notice that

$$\mathrm{SC}(\Omega_i) - \mathrm{SC}(\Omega) > 0 \iff f(T_i) - g(\gamma) > 0 \iff T_i - f^{-1}(g(\gamma)) > 0.$$

Here, $f^{-1}(g(\gamma))$ for each variable selection criterion is as follows (the proof is given in Appendix A.2):

$$f^{-1}(g(\gamma)) = \begin{cases} \gamma & (GC_p), \\ (n-k)\left\{\exp\left(\dfrac{\gamma}{n}\right) - 1\right\} & (\mathrm{GIC}), \\ \dfrac{\gamma(n-k)}{\gamma+n-k} & (\mathrm{GBNP}), \\ (n-k)\left\{\left(\dfrac{n-k-\gamma+1}{n-k-\gamma}\right)^2 - 1\right\} & (\mathrm{GCV}_C), \\ (n-k)\left\{\exp\left(\dfrac{2(n-1)}{(n-k-1)(n-k-2)}\right) - 1\right\} & (\mathrm{AIC}_C), \\ (n-k)\left\{\left(\dfrac{n-k+1}{n-k}\right)^\gamma - 1\right\} & (\mathrm{EGCV}). \end{cases} \tag{8}$$

Let $\delta = f^{-1}(g(\gamma))$. Using these equations, the selected model $\hat{M}$ in (3) can be rewritten as follows, with $T_i - \delta$ as the discriminant function of the ith variable:

$$\hat{M} = \{i \in \Omega \mid T_i - \delta > 0\}, \quad T_i = \frac{(\boldsymbol{y}^\top \boldsymbol{r}_i)^2}{s_\Omega^2}, \quad \delta = f^{-1}(g(\gamma)) \quad (i = 1, \ldots, k). \tag{9}$$

3. Confidence Interval Conditioned on the Selection Event

In this section, under the condition that the model M is selected via the KOO method, we construct a confidence interval for the regression coefficient of the jth explanatory variable in the model M that satisfies the following equation:

$$\mathbb{P}\left(\beta_M^{(j)} \in C_M^{(j)}(\delta) \,\middle|\, \mathcal{S}(M|\delta)\right) = 1 - \alpha \quad (j \in M), \tag{10}$$

where $\mathcal{S}(M|\delta)$ denotes the selection event that the model M is selected via the KOO method, i.e., $\{\hat{M} = M\}$. Let $t_M^{(j)}$ be the jth element of $\hat{\boldsymbol{\beta}}_M$ given by (2), which is given by

$$t_M^{(j)} = \boldsymbol{y}^\top \boldsymbol{\eta}_M^{(j)}, \quad \boldsymbol{\eta}_M^{(j)} = (\boldsymbol{X}_M^+)^\top \boldsymbol{e}_j, \tag{11}$$

where $\boldsymbol{e}_j$ is a k-dimensional vector in which only the jth element is 1 and the other elements are 0, as defined in Section 2. Then, we consider the following conditional distribution of the jth element of $\hat{\boldsymbol{\beta}}_M$ to obtain the target confidence interval $C_M^{(j)}(j)$ satisfying (10):

$$t_M^{(j)} \,\middle|\, \mathcal{S}(M|\delta), \quad t_M^{(j)} \sim N\left(\boldsymbol{\mu}^\top \boldsymbol{\eta}_M^{(j)}, \sigma^2 \left\|\boldsymbol{\eta}_M^{(j)}\right\|^2\right). \tag{12}$$

In the following, we consider rewriting the selection event $\mathcal{S}(M|\delta)$ as the region of $t_M^{(j)}$ to obtain the conditional distribution in (12). It follows from equation (9) that the selection event can be expressed as

$$\mathcal{S}(M|\delta) = \bigcap_{i \in M} \{T_i - \delta > 0\} \cap \bigcap_{i \notin M} \{T_i - \delta \leq 0\}. \tag{13}$$

Let $\boldsymbol{z}_M^{(j)}$ and $\boldsymbol{w}_M^{(j)}$ be n-dimensional vectors of random variables and constants, respectively, defined by

$$\boldsymbol{z}_M^{(j)} = \boldsymbol{y} - t_M^{(j)} \boldsymbol{w}_M^{(j)}, \quad \boldsymbol{w}_M^{(j)} = \frac{\boldsymbol{\eta}_M^{(j)}}{||\boldsymbol{\eta}_M^{(j)}||^2}. \tag{14}$$

We note that $t_M^{(j)}$ and $\boldsymbol{z}_M^{(j)}$, $t_M^{(j)}$ and s_Ω^2, s_Ω^2 and $\boldsymbol{z}_M^{(j)}$ are independent (the proof is given in A.3). The definition of $\boldsymbol{z}_M^{(j)}$ in (14) implies that $\boldsymbol{y} = \boldsymbol{z}_M^{(j)} + t_M^{(j)} \boldsymbol{w}_M^{(j)}$. Expanding T_i after substituting this equation into (6) yields

$$\begin{aligned} T_i &= \frac{1}{s_\Omega^2} \left\{ \left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right)^2 \left(t_M^{(j)}\right)^2 + 2\left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)\left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right) t_M^{(j)} + \left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)^2 \right\} \\ &= \frac{1}{s_\Omega^2} q\left(t_M^{(j)} \,\middle|\, \left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right)^2, 2\left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)\left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right), \left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)^2\right), \end{aligned}$$

where $\boldsymbol{r}_i$ is an n-dimensional vector given by (4), and $q(x|a, b, c)$ denotes a quadratic function with coefficients a, b, and c, i.e., $q(x|a, b, c) = ax^2 + bx + c$. Here, since $M \subset \Omega_i$ is satisfied if $i \notin M$, $\boldsymbol{P}_{\Omega_i} \boldsymbol{X}_M = \boldsymbol{X}_M$ for $i \notin M$ holds. Hence, we have

$$\boldsymbol{r}_i \boldsymbol{r}_i^\top \boldsymbol{\eta}_M^{(j)} = (\boldsymbol{P}_\Omega - \boldsymbol{P}_{\Omega_i}) \boldsymbol{X}_M \left(\boldsymbol{X}_M^\top \boldsymbol{X}_M\right)^{-1} \boldsymbol{e}_j = \boldsymbol{0}_n \iff \boldsymbol{r}_i^\top \boldsymbol{\eta}_M^{(j)} = 0.$$

Since the above result also shows $\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)} = 0$, T_i for unused variables in the model M can be rewritten as

$$T_i = \frac{1}{s_\Omega^2}\left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right).$$

This indicates that $t_M^{(j)}$ does not appear in the region $T_i - \delta < 0$ for the unselected variables. Since $t_M^{(j)}$ is independent of $\boldsymbol{z}_M^{(j)}$ and s_Ω^2, respectively, we can see that the events on unselected variables via the KOO method are independent of the target conditional distribution in (12). From this result, the selection event in (13) is rewritten by using only events on the selected variables as

$$\mathcal{S}(M|\delta) = \bigcap_{i\in M}\left\{q\left(t_M^{(j)} \;\middle|\; \left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right)^2, 2\left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)\left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right), \left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)^2 - s_\Omega^2\delta\right) > 0\right\}. \tag{15}$$

Applying the quadratic formula to q, we have

$$\begin{aligned} &q\left(t_M^{(j)} \;\middle|\; \left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right)^2, 2\left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)\left(\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}\right), \left(\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}\right)^2 - s_\Omega^2\delta\right) > 0 \\ \iff\; & t_M^{(j)} \in \left(\infty, A_i^{(j)}(\delta)\right] \cup \left[B_i^{(j)}(\delta), \infty\right), \end{aligned}$$

where boundaries $A_i^{(j)}(\delta)$ and $B_i^{(j)}(\delta)$ are given by

$$A_i^{(j)}(\delta) = -\frac{\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}}{\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}} - \frac{\sqrt{s_\Omega^2\delta}}{|\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}|}, \quad B_i^{(j)}(\delta) = -\frac{\boldsymbol{r}_i^\top \boldsymbol{z}_M^{(j)}}{\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}} + \frac{\sqrt{s_\Omega^2\delta}}{|\boldsymbol{r}_i^\top \boldsymbol{w}_M^{(j)}|}.$$

Let $I_M^{(j)}(\delta)$ be a region defined by

$$I_M^{(j)}(\delta) = \bigcap_{i\in M}\left\{\left(\infty, A_i^{(j)}(\delta)\right] \cup \left[B_i^{(j)}(\delta), \infty\right)\right\}. \tag{16}$$

It should be emphasized that from a simple calculation, the region $I_M^{(j)}(\delta)$ can be represented as the union of several disjoint intervals. Using the region in (16), the selection event in (15) can be given by

$$\mathcal{S}(M|\delta) = \left\{t_M^{(j)} \in I_M^{(j)}(\delta)\right\}.$$

This implies that the conditional distribution in (12) is the truncated normal distribution with mean $\boldsymbol{\mu}^\top \boldsymbol{\eta}_M^{(j)}$ and variance σ^2 constrained by the truncated region $I_M^{(j)}(\delta)$. Here, although the truncated region $I_M^{(j)}(\delta)$ includes closed intervals, the cumulative probability remains the same whether or not it includes boundary values, since the truncated normal distribution is a continuous distribution. Consequently, region $I_M^{(j)}(\delta)$ will be treated hereafter as the union of disjoint open intervals.

Let $\Phi(x)$ be the distribution function of $N(0, 1)$, and $F(x; \mu, \sigma^2, D)$ denote the distribution function of the truncated normal distribution with mean μ and variance σ^2 constrained by the truncated region $D = \cup_{i=1}^m (a_i, b_i)$ which is written as, for $x \in (a_r, b_r)$,

$$F(x; \mu, \sigma^2, D) = \frac{\sum_{i=1}^{r-1} p_i + \Phi\left((x-\mu)/\sigma\right) - \Phi\left((a_r-\mu)/\sigma\right)}{\sum_{i=1}^{m} p_i}, \tag{17}$$

where $p_i = \Phi((b_i - \mu)/\sigma) - \Phi((a_i - \mu)/\sigma)$. From Lemma A.1 in Lee *et al.* (2016), we can see that $F(x; \mu, \sigma^2, D)$ is monotone decreasing in μ. Let $L_M^{(j)}(\delta)$ and $U_M^{(j)}(\delta)$ be the solutions of the following equations for L and U, respectively:

$$F\left(\boldsymbol{y}^\top \boldsymbol{\eta}_M^{(j)}; L, \sigma^2, I_M^{(j)}(\delta)\right) = 1 - \frac{\alpha}{2}, \quad F\left(\boldsymbol{y}^\top \boldsymbol{\eta}_M^{(j)}; U, \sigma^2, I_M^{(j)}(\delta)\right) = \alpha.$$

Using the solutions, the $1 - \alpha$ confidence interval $C_M^{(j)}(\delta)$ in (10), the focus of this section, can be given as

$$C_M^{(j)}(\delta) = \left[L_M^{(j)}(\delta), U_M^{(j)}(\delta)\right]. \tag{18}$$

That $C_M^{(j)}(\delta)$ in (18) satisfies (10) is clear from the fact that the distribution function in (17) is monotonic decreasing with respect to μ. In addition, since $L_M^{(j)}(\delta)$ and $U_M^{(j)}(\delta)$ in $C_M^{(j)}(\delta)$ include an unknown variance σ^2, in practice, these can be derived using s_Ω^2 instead of σ^2.

4. Numerical Experiments

4.1. Examining Confidence Interval Coverage Probabilities

We performed a series of numerical experiments to determine whether the confidence intervals for the regression coefficients produced by the proposed method actually contained the true regression coefficients at a rate of $100 \times (1 - \alpha)\%$. The motivation for this relates to the fact that the actual coverage probability in the case of naive interval estimation without conditioning, i.e., an ordinary confidence interval based on the t-distribution, differs from the nominal confidence level of $1 - \alpha$. Thus, we compared the coverage probabilities of ordinary confidence intervals in the naive case with the coverage probabilities of the proposed confidence interval estimates.

In our numerical experiments, vectors of the explanatory variables were generated as

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim i.i.d.\ N_k(\mathbf{0}_k, \boldsymbol{\Psi}_k),$$

where $\mathbf{0}_k$ denotes a k-dimensional vector of zeros, and $\boldsymbol{\Psi}_k$ denotes a $k \times k$ autocorrelation matrix in which the (a, b)th element is $0.6^{|a-b|}$. Using these vectors, simulation data were generated as

$$\boldsymbol{y} \sim N_n\left(\boldsymbol{X}\boldsymbol{\theta}_a^*, \boldsymbol{I}_n\right), \quad \boldsymbol{\theta}_a^* = (1.04, 1.71, 2.26, 0.85, 2.20/a, \mathbf{0}_{k-5}^\top)^\top. \tag{19}$$

It should be noted that the first five coefficients of $\boldsymbol{\theta}_a^*$ in (19) are non-zero and the remaining $k - 5$ coefficients are all zero. In the experiments, one of three models was selected, with a sample size $n = 100$, 500, and 1,000. and with $k = 10$ and 30 explanatory variables. The three models considered were the true model $M_* = \{1, 2, 3, 4, 5\}$, an overspecified model $M_*^+ = \{1, 2, 3, 4, 5, 6\}$ with a sixth explanatory variable not included in the true model added to M_*, and an underspecified model $M_*^- = \{1, 2, 3, 4\}$, where the fifth variable in the true model was removed. The calculations were repeated until each of the M_*, M_*^+, and M_*^- models was selected 10,000 times using the six variable selection criteria: the C_p criterion, AIC, BIC, HQC, CAIC, and the BNP criterion. When conducting the experiment in the case of $\hat{M} = M_*$ or $\hat{M} = M_*^+$, we used 1 as a in (19). On the other hand, in the case of $\hat{M} = M_*^-$, a in (19) was set to 30 in order to facilitate the selection of model M_*^-. For each specified model selected, a 95% confidence interval was constructed using the method we propose

Table 1. Coverage probabilities when $\hat{M} = M_*$ and $n = 100$

		C_p		AIC		BIC		HQC		CAIC		BNP	
k	j	P	N	P	N	P	N	P	N	P	N	P	N
10	1	93.85	95.69	93.63	95.46	93.68	95.48	94.00	95.65	93.53	95.42	93.80	95.65
	2	93.81	94.29	93.73	94.22	93.28	93.81	93.74	94.23	93.33	93.84	93.82	94.27
	3	93.61	94.20	93.45	94.06	93.23	93.83	93.61	94.14	93.42	94.00	93.66	94.25
	4	93.94	94.64	94.02	94.64	93.51	94.29	93.99	94.67	93.60	94.42	93.97	94.67
	5	93.46	96.48	93.69	96.57	93.52	96.61	93.72	96.82	93.52	96.60	93.43	96.47
	Prob.	47.42		42.99		82.47		64.01		89.21		48.26	
30	1	91.49	96.80	92.33	97.15	90.58	96.03	91.09	96.66	90.05	95.83	91.39	96.72
	2	91.35	94.66	92.00	94.99	90.52	93.62	91.57	94.62	90.44	93.66	91.51	94.75
	3	91.91	95.32	92.06	95.38	90.73	94.63	91.41	95.01	90.75	94.84	91.83	95.22
	4	91.50	96.67	92.11	96.61	90.19	95.90	91.20	96.37	90.54	96.22	91.22	96.45
	5	91.48	97.96	92.07	97.94	90.75	97.79	91.21	97.69	90.69	97.89	91.44	97.91
	Prob.	3.54		0.61		21.97		5.06		35.96		3.88	

Table 2. Coverage probabilities when $\hat{M} = M_*$ and $n = 500$

		C_p		AIC		BIC		HQC		CAIC		BNP	
k	j	P	N	P	N	P	N	P	N	P	N	P	N
10	1	94.48	94.92	94.52	94.97	94.18	94.65	94.35	94.79	94.15	94.60	94.46	94.90
	2	94.83	95.13	94.80	95.11	94.84	95.11	94.65	94.95	94.81	95.07	94.78	95.08
	3	94.84	94.92	94.85	94.94	94.91	94.99	94.96	95.04	94.96	95.04	94.84	94.92
	4	93.92	94.40	93.92	94.41	93.97	94.58	94.06	94.59	94.06	94.56	93.93	94.41
	5	94.78	98.08	94.74	98.07	95.00	98.32	94.85	98.21	94.85	98.35	94.98	98.09
	Prob.	46.90		45.99		93.83		75.75		96.39		47.10	
30	1	94.06	94.99	93.92	94.85	93.81	94.71	94.23	95.05	93.65	94.56	94.01	94.96
	2	94.44	95.09	94.31	95.00	94.09	94.77	94.35	95.05	94.05	94.73	94.41	95.04
	3	95.10	95.50	95.29	95.68	94.18	94.60	94.86	95.23	94.30	94.76	95.07	95.48
	4	94.05	94.91	94.22	95.06	93.94	94.79	93.81	94.63	93.74	94.61	94.01	94.87
	5	94.44	97.79	94.50	97.89	94.68	97.86	94.60	97.91	94.67	97.90	94.58	97.79
	Prob.	2.69		1.99		69.86		23.82		80.66		2.76	

and a naive method using the t-distribution; the proportion of true regression coefficients included in the confidence interval was then calculated as the coverage probability.

Tables 1 through 9 show the coverage probabilities, i.e., the probability that the regression coefficient of the jth variable is within the corresponding confidence interval. The "P" column gives the coverage probabilities for the confidence intervals produced by the proposed method; the "N" column gives the coverage probabilities for the confidence intervals produced by the naive method. "Prob." indicates the probability that the specific model was selected. A dash ("–") indicates that no values were entered. (Depending on the values of n and k, and the applied variable selection criterion, the specified model may not be selected at all, meaning that a probability and confidence interval could not be calculated.) As shown, for the confidence intervals produced by the proposed method, the coverage probability differed slightly from 95% when n was small but was very close to 95% when n was large. On the other hand, for the confidence intervals constructed with the

Table 3. Coverage probabilities when $\hat{M} = M_*$ and $n = 1{,}000$

k	j	C_p P	C_p N	AIC P	AIC N	BIC P	BIC N	HQC P	HQC N	CAIC P	CAIC N	BNP P	BNP N
10	1	94.37	94.82	94.40	94.84	94.42	94.89	94.54	95.02	94.42	94.91	94.38	94.83
	2	94.25	94.65	94.31	94.71	94.48	94.95	94.49	94.96	94.49	94.95	94.26	94.65
	3	94.19	94.86	94.17	94.84	94.84	95.39	94.59	95.17	94.80	95.35	94.20	94.85
	4	95.07	95.60	95.03	95.55	94.70	95.26	94.97	95.51	94.64	95.19	95.08	95.62
	5	94.80	98.18	94.80	98.16	94.82	98.11	94.88	98.19	94.83	98.11	94.80	98.18
	Prob.	46.75		46.32		95.92		79.74		97.70		46.87	
30	1	94.21	94.89	94.13	94.78	93.97	94.79	94.05	94.79	93.97	94.75	94.23	94.91
	2	94.47	95.04	94.35	95.00	94.19	94.85	94.18	94.80	94.22	94.82	94.44	95.01
	3	93.97	94.99	94.01	95.03	94.01	94.82	94.23	94.98	94.02	94.84	94.02	95.02
	4	93.82	94.78	93.71	94.66	94.13	95.03	94.24	95.01	94.27	95.17	93.85	94.81
	5	94.87	97.95	94.07	97.95	94.21	97.98	94.29	97.91	94.16	97.94	94.09	97.95
	Prob.	2.59		2.21		79.83		30.61		87.57		2.62	

Table 4. Coverage probabilities when $\hat{M} = M_*^+$ and $n = 100$

k	j	C_p P	C_p N	AIC P	AIC N	BIC P	BIC N	HQC P	HQC N	CAIC P	CAIC N	BNP P	BNP N
10	1	93.04	94.81	93.05	94.94	92.54	94.18	93.06	94.80	92.16	93.91	92.88	94.79
	2	93.68	93.58	93.51	93.47	92.77	92.57	93.43	93.34	92.68	92.37	93.56	93.53
	3	93.50	93.65	93.64	93.80	92.64	92.71	93.29	93.58	92.64	92.70	93.35	93.67
	4	93.70	94.45	93.73	94.45	92.66	93.31	92.96	93.65	92.35	93.16	93.73	94.41
	5	94.27	88.77	94.22	89.02	93.14	84.49	93.96	87.34	92.88	82.60	94.34	88.79
	6	91.39	77.19	91.25	79.08	90.72	46.00	91.28	67.09	90.96	33.40	92.22	76.74
	Prob.	8.09		8.43		3.41		6.26		2.19		8.04	
30	1	91.60	95.46	92.33	95.75	90.39	94.66	91.57	95.77	90.35	94.31	91.55	95.54
	2	90.82	94.00	91.70	95.73	89.90	93.35	91.25	94.27	90.14	93.47	90.82	94.00
	3	91.26	94.59	91.77	95.17	94.43	93.72	91.14	94.78	89.31	92.94	91.41	94.69
	4	91.62	95.77	92.01	96.18	90.65	95.33	90.98	95.50	90.62	94.88	91.57	95.67
	5	92.88	88.34	93.37	90.19	90.84	84.32	92.20	87.21	90.77	81.86	92.61	87.97
	6	91.28	87.15	91.42	90.75	90.97	77.97	91.25	85.74	91.21	72.12	91.63	86.62
	Prob.	0.40		0.11		1.11		0.51		1.23		0.43	

naive method, there were cases where the coverage probability did not approach 95% even when n increased, especially for M_*^+.

Figure 1 shows the confidence intervals and estimated regression coefficients when each model is selected by BIC for $n = 1{,}000$ and $k = 30$. The red intervals correspond to the proposed method; the blue intervals correspond to the naive method. The black dots show the value of the estimated regression coefficient. From the figure, it is apparent that the intervals for the proposed method tended to be slightly narrower than the those based on the naive method for M_* and M_*^-, and wider for M_*^+. It can also be seen that for a variable whose true value was zero (the sixth variable in M_*^+), the confidence interval produced by the proposed method includes zero, while the interval produced by the naive method does not.

Table 5. Coverage probabilities when $\hat{M} = M_*^+$ and $n = 500$

k	j	C_p P	C_p N	AIC P	AIC N	BIC P	BIC N	HQC P	HQC N	CAIC P	CAIC N	BNP P	BNP N
10	1	94.03	94.63	94.04	94.68	94.07	94.72	94.01	94.67	93.80	94.45	94.03	94.66
	2	94.66	94.93	94.74	94.91	94.34	94.53	94.91	95.09	94.02	94.34	94.75	94.94
	3	94.90	94.90	94.88	94.91	94.18	94.22	94.70	94.75	94.53	94.61	94.96	94.92
	4	94.18	94.57	94.26	94.52	94.21	94.33	94.65	95.07	94.57	94.74	94.20	94.57
	5	96.14	86.31	96.11	86.35	94.71	74.65	95.55	81.35	94.57	71.12	96.10	86.29
	6	93.93	84.29	94.02	84.57	94.41	39.76	93.34	68.38	94.58	30.79	93.92	84.20
	Prob.	7.55		7.53		1.21		4.18		0.73		7.53	
30	1	94.83	95.03	94.20	95.24	93.95	94.77	94.30	95.10	93.60	94.45	94.17	95.05
	2	94.10	94.65	93.96	94.67	93.99	94.40	94.30	94.65	93.52	94.19	94.06	94.61
	3	93.80	94.66	93.67	94.60	94.07	94.74	93.96	94.68	93.61	94.37	93.77	94.61
	4	93.96	94.57	93.70	94.34	93.82	94.37	94.30	94.73	93.85	94.41	93.88	94.49
	5	94.84	86.85	95.08	87.49	93.21	77.45	93.67	83.51	92.91	74.51	94.90	86.91
	6	93.24	83.61	93.10	84.54	93.93	42.46	92.86	67.80	93.64	33.03	93.29	83.57
	Prob.	0.41		0.33		1.00		1.35		0.69		0.41	

Table 6. Coverage probabilities when $\hat{M} = M_*^+$ and $n = 1,000$

k	j	C_p P	C_p N	AIC P	AIC N	BIC P	BIC N	HQC P	HQC N	CAIC P	CAIC N	BNP P	BNP N
10	1	94.69	95.13	94.82	95.21	–	–	94.73	95.08	–	–	94.71	95.15
	2	94.96	95.19	94.96	95.25	–	–	94.60	94.80	–	–	94.91	95.18
	3	94.97	95.09	94.98	95.03	–	–	94.60	94.53	–	–	94.97	95.04
	4	94.97	95.14	95.00	95.16	–	–	94.28	94.47	–	–	95.04	95.13
	5	95.13	86.83	95.02	86.83	–	–	94.55	81.74	–	–	95.16	86.80
	6	93.79	84.00	93.79	84.15	–	–	93.72	65.18	–	–	93.84	83.97
	Prob.	7.38		7.42		–		3.64		–		7.37	
30	1	94.42	95.13	94.25	94.91	94.35	94.96	94.21	94.63	94.32	94.77	94.50	95.11
	2	94.34	95.01	94.43	95.03	94.27	94.85	94.24	94.66	94.39	94.82	94.45	95.01
	3	94.54	95.15	94.69	95.24	93.97	94.44	94.15	94.64	94.44	94.80	94.57	95.13
	4	94.12	94.81	94.08	94.74	94.21	94.93	94.24	95.01	94.06	94.78	94.20	94.79
	5	95.30	88.09	95.19	88.11	93.65	75.01	94.26	81.94	93.70	71.95	95.12	88.13
	6	93.65	82.28	93.69	82.91	94.04	29.11	93.94	62.57	94.60	20.37	93.63	82.24
	Prob.	0.40		0.36		0.72		1.45		0.46		0.40	

4.2. Behavior of Confidence Intervals with Increasing δ

In the previous subsection, we conducted numerical experiments regarding the coverage probability of the proposed confidence interval with a fixed δ in (18). However, as the value of δ is changed, the truncated region in (16) is also changed, and the confidence interval in (18) is changed as well. Since the selected model will transition if δ exceeds a certain value that exists for the number of variables, it is easy to see that the confidence interval shifts discretely once δ exceeds one of these points. In this subsection, we use a numerical experiment to examine the behavior of the truncated

Table 7. Coverage probabilities when $\hat{M} = M_*^-$ and $n = 100$

		C_p		AIC		BIC		HQC		CAIC		BNP	
k	j	P	N	P	N	P	N	P	N	P	N	P	N
10	1	93.79	95.51	93.81	95.58	93.77	95.49	93.97	95.56	93.73	95.48	93.77	95.51
	2	93.80	94.37	93.76	94.36	93.36	93.89	93.62	94.10	93.45	93.98	93.76	94.33
	3	93.90	94.66	93.47	94.65	93.40	94.21	93.85	94.60	93.22	94.07	93.91	94.65
	4	93.77	96.45	94.04	96.56	93.51	96.28	93.68	96.48	93.58	96.38	93.69	96.42
	Prob.	37.85		33.32		76.32		54.79		84.59		38.66	
30	1	91.80	96.91	92.34	97.22	90.60	96.24	91.15	96.63	90.20	96.03	91.55	96.78
	2	91.71	94.87	92.14	95.37	90.64	93.86	91.69	94.75	90.65	93.70	91.67	94.83
	3	92.20	95.49	91.92	95.59	91.06	94.95	91.85	95.21	90.72	94.91	92.23	95.42
	4	91.61	99.00	91.81	98.99	90.73	98.72	91.39	98.94	90.68	98.82	91.44	99.00
	Prob.	3.09		0.49		20.43		4.44		34.26		3.38	

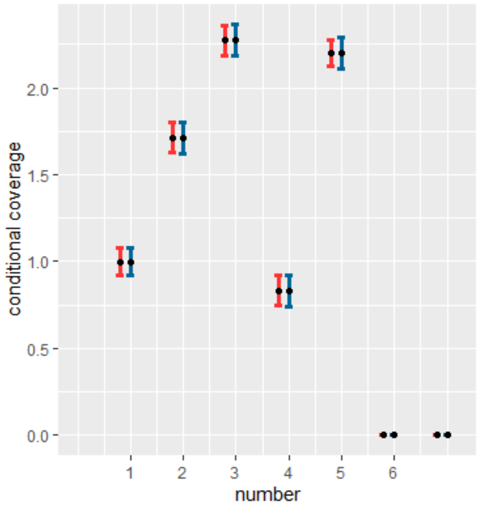
Table 8. Coverage probabilities when $\hat{M} = M_*^-$ and $n = 500$

		C_p		AIC		BIC		HQC		CAIC		BNP	
k	j	P	N	P	N	P	N	P	N	P	N	P	N
10	1	94.22	94.83	94.21	94.83	94.12	94.72	94.47	95.02	94.13	94.75	94.23	94.84
	2	94.90	95.17	94.88	95.15	94.88	95.16	94.76	95.05	94.84	95.10	94.90	95.15
	3	94.62	94.70	94.67	94.75	94.83	94.90	94.95	95.03	94.87	94.94	94.64	94.72
	4	93.92	97.30	94.02	97.38	93.82	97.24	93.93	97.27	93.88	97.30	93.93	97.32
	Prob.	29.11		28.29		85.91		59.84		90.97		29.25	
30	1	93.83	94.84	94.01	95.01	94.02	94.98	94.14	95.10	93.76	94.73	93.76	94.78
	2	94.22	95.05	94.04	94.86	94.12	94.83	94.42	95.01	94.01	94.68	94.29	95.09
	3	94.89	95.37	94.74	95.21	94.04	94.51	95.18	95.53	94.23	94.73	94.97	95.46
	4	94.13	97.67	94.12	97.67	93.53	97.52	93.84	97.56	93.25	97.48	94.11	97.68
	Prob.	1.71		1.23		63.99		18.76		75.96		1.75	

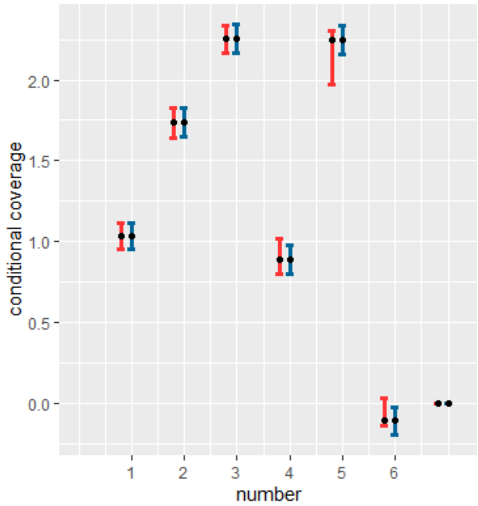
Table 9. Coverage probabilities when $\hat{M} = M_*^-$ and $n = 1,000$

		C_p		AIC		BIC		HQC		CAIC		BNP	
k	j	P	N	P	N	P	N	P	N	P	N	P	N
10	1	94.27	94.83	94.28	94.83	94.35	94.84	94.25	94.72	94.38	94.87	94.28	94.84
	2	94.15	94.63	94.18	94.66	94.46	95.04	94.18	94.74	94.48	95.10	94.15	94.63
	3	93.90	94.89	93.92	94.92	94.76	95.36	94.58	95.36	94.96	95.52	93.92	94.89
	4	94.83	98.09	94.85	98.09	94.81	97.93	94.65	97.98	94.83	97.95	94.81	98.08
	Prob.	7.51		7.37		59.03		27.81		66.55		7.54	
30	1	–	–	–	–	94.51	95.33	94.81	95.54	94.36	95.19	–	–
	2	–	–	–	–	94.30	95.12	94.48	95.25	94.20	94.97	–	–
	3	–	–	–	–	93.96	94.65	94.12	95.15	94.06	94.71	–	–
	4	–	–	–	–	93.69	97.34	93.83	97.57	93.70	97.29	–	–
	Prob.	–		–		48.79		10.54		59.53		–	

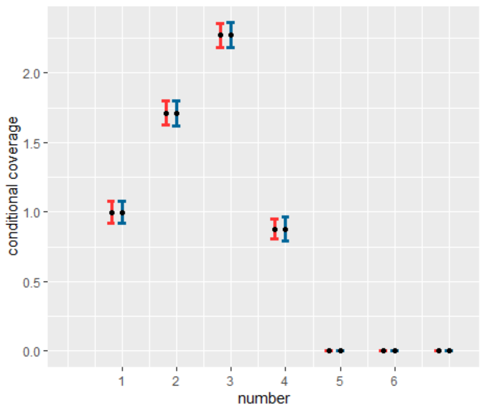
region and confidence interval under an increasing δ when $(n, k) = (500, 10)$. The simulation data were generated by the model in (19) used in subsection 4.1, with $\boldsymbol{\theta}_a^*$ replaced by



Case of $\hat{M} = M_*$



Case of $\hat{M} = M_*^+$



Case of $\hat{M} = M_*^-$

Figure 1. Confidence intervals when BIC is used and (n, k) = (1,000, 30) (1st of 10,000 iterations)

$$(-0.08, 0.26, -0.14, 0.06, 0.31, 0, 0, 0, 0, 0)^\top.$$

Figure 2 shows the confidence intervals and truncated regions of the fifth variable for increasing δ. They were created using the slider object in the R package "plotly" (see Sievert, 2019). In the upper portion of each figure, the horizontal axis represents δ, the red dot denotes the estimate of $\beta_M^{(j)}$, and the black lines denote the upper and lower boundaries of the confidence interval $C_M^{(j)}(\delta)$. The purple or green vertical bars indicate the model change points at which each variable is no longer selected, with the purple vertical bars representing the model change point at which variables with true regression coefficients of zero are no longer selected, and the green vertical bars representing the model change point at which variables with true non-zero regression coefficients are no longer selected. We note that the model change points are the ordered statistics of $T_1, ..., T_k$. The lower part of each figure shows the truncated region of the distribution function for

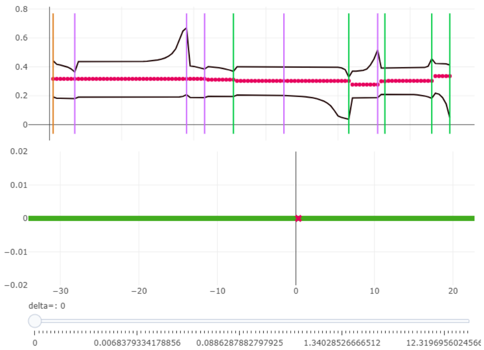
$\delta = 0.0, 1.4\times10^{-3}, 1.96\times10^{-2}, 3.39\times10^{-2}, 7.13\times10^{-2}$ and 19.579, for which an orange vertical bar is drawn in the upper part of the figure. The red $\times$ is an estimate of $\beta_M^{(j)}$. Figure 3 shows the change in estimates and confidence intervals for the first to the tenth variables when increasing δ, i.e., the upper part of each figure in Figure 2 for the first to the tenth variables. We found the following tendencies in the behavior of the confidence intervals and the truncated regions for increases in δ by moving the slider object:

- The closer δ is to the model change point, the more extreme the transition of one side of the confidence interval.
- The closer δ is to the model change point, the narrower the truncated region becomes.
- When the truncated region is narrow, the closer the boundary of an open interval containing the estimated regression coefficient is to the estimated value, the more extreme the transition of the confidence interval.
- The confidence interval shifts in the positive or negative direction as the right or left boundary of the truncated interval approaches the estimate of the regression coefficient.

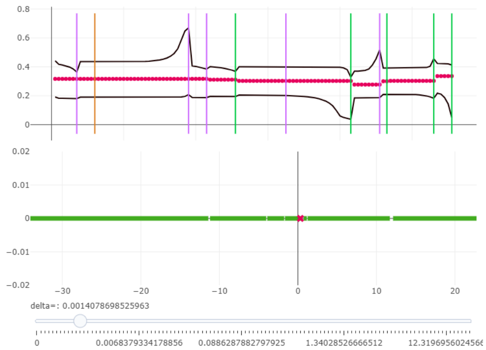
5. Conclusion

In this paper, we derive conditional confidence intervals for $\beta_M^{(j)}$ with $1-\alpha$ coverage, conditioned on the model M selected by the KOO method. To systematically construct such confidence intervals when applying various variable selection criteria, we give a general expression of the discriminant function used in the KOO method. For the variable selection criteria considered in this paper, the ith discriminant function $\mathrm{SC}(\Omega) - \mathrm{SC}(\Omega_i)$ is represented as the difference between T_i in (6), which is a statistic common to all the variable selection criteria, and the threshold δ in (9), which varies with the variable selection criterion. The confidence interval is derived from the conditional distribution of $t_M^{(j)}$ in (11) conditioned on the selection event consisting of the general expression of the discriminant function, that is, the truncated normal distribution with the truncated region $I_M^{(j)}(\delta)$ in (16). We showed that the truncated region depends on the selected variables, but not on the unselected variables. Numerical experiments confirmed that the coverage probability of the proposed confidence intervals tends to be closer to the nominal confidence level than that of confidence intervals obtained from the naive method. Furthermore, various transitions of the confidence interval were confirmed by increasing the threshold δ in numerical experiments.

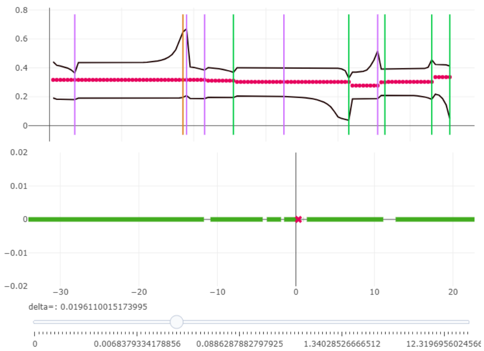
In addition to the variable selection criteria considered in this paper, there are other widely known variable selection criteria, including the cross-validation (CV) criterion proposed by Stone (1974) and the extended information criterion (EIC) proposed by Ishiguro, Sakamoto and Kitagawa (1997), which uses bootstrapping. It should be noted that the confidence interval expression in this paper cannot be used when the model is selected by the KOO method with the CV criterion, EIC, or the bias-corrected CV criteria proposed by Yanagihara, Tonda and Matsumoto (2006), and Yanagihara



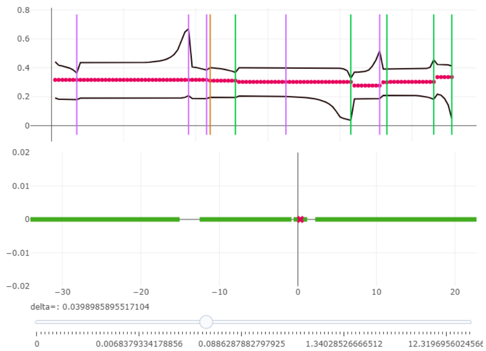
$\delta = 0.0$



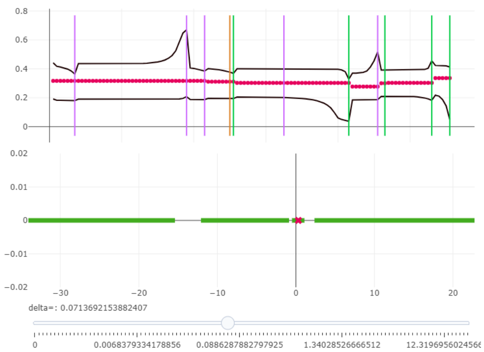
$\delta = 0.0014$



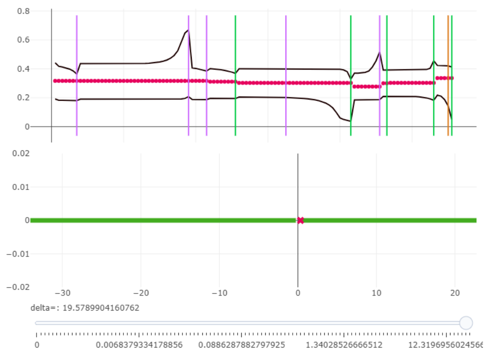
$\delta = 0.0196$



$\delta = 0.0399$



$\delta = 0.0713$



$\delta = 19.579$

Figure 2. **Truncated region when increasing δ for the 5th variable**

and Fujisawa (2012). It is left to future work to obtain a general expression of the confidence interval that includes the case in which variables are selected by the KOO method using these additional criteria. For more information on variable selection criteria in linear regression, see, for example, Yanagihara *et al*. (2017).

Finally, although the behavior of the proposed confidence interval in response to increases in the threshold δ was studied, we did not produce any theoretical results regarding this behavior. We hope that further study will lead to a solution to the problem of how to optimize δ.

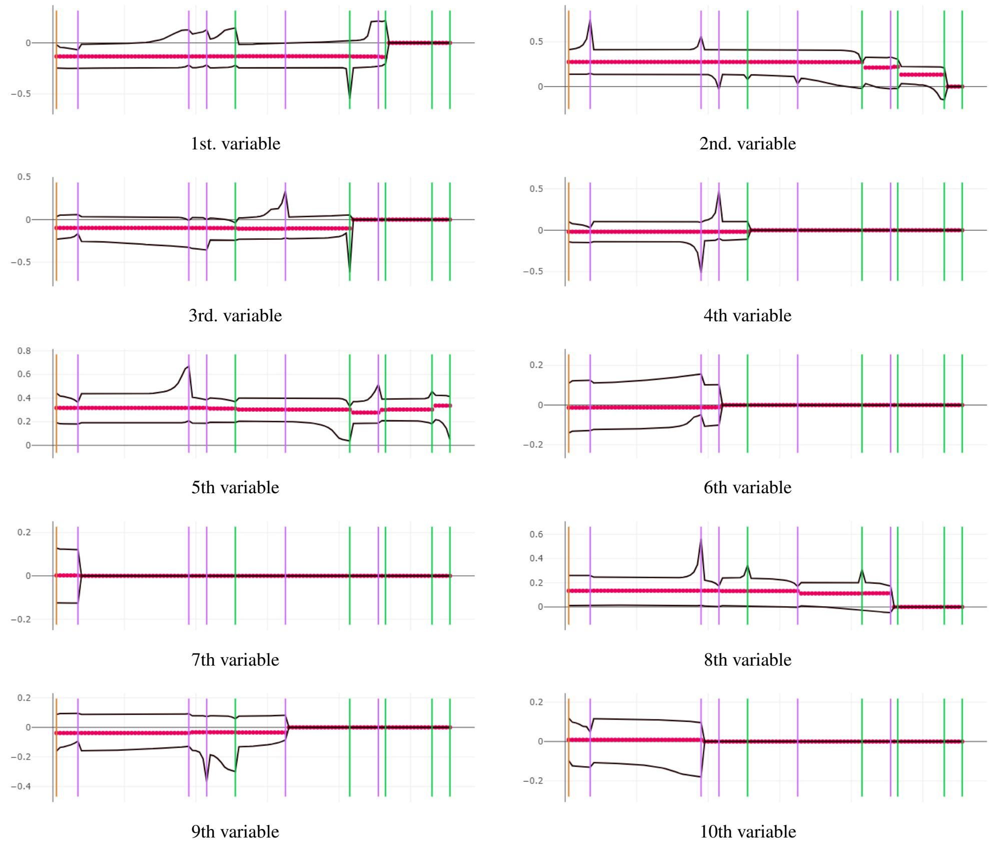


Figure 3. **Behavior of confidence intervals when increasing δ for each variable**

Acknowledgment The authors would like to thank Professor Yoshiyuki Ninomiya of the Institute of Statistical Mathematics, and Dr. Ryoya Oda and Dr. Mineaki Ohishi of Hiroshima University for their helpful comments. The second author's research was partially supported by JSPS KAKENHI Grant Number 20H04151 and JSPS Bilateral Program Grant Number JPJSBP120219927.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceeding of 2nd. International Symposium on information Theory* (eds. B. N. Petrov & F. Csáki), pp. 267–281, Akadèmiai Kiadò, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.

Bai, Z. D., Fujikoshi, Y. & Hu, J. (2018). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. TR No. 18-09, *Statistical Research Group*, Hiroshima University.

Boonstra, P. S., Mukherjee, B. & Taylor, J. M. G. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statist. Sinica*, **25**, 1185–1206.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Charkhi, A. & Claeskens, G. (2017). AIC post-selection inference in linear regression. In *Proceedings of 20th The European Young Statisticians Meetings*, 14–18 August 2017, Uppsala University. https://indico.uu.se/event/317/contributions/284/

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method og generalized cross-validation. *Number. Math.*, **31**, 377–403.
Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.
Ishiguro, M., Sakamoto, Y. & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.*, **49**, 411–434.
Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser.* **B**, **41**, 190–195.
Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York.
Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.
Mallows C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.
Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.
Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.
Oda, R. & Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron. J. Stat.*, **14**, 1386–1412.
Ohishi, M. (2021). Ridge parameters optimization based on minimizing model selection criterion in multivariate generalized ridge regression. *Hiroshima Math. J.*, **51**, 177–226.
Ohishi, M., Yanagihara, H. & Fujikoshi, Y. (2020). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *J. Statist. Plann. Infer.*, **204**, 187–205.
Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Sievert, C. (2019). Interactive web-based data visualization with R, plotly, and shiny. https://plotly-r.com.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**, **36**, 111–147.
Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods*, **A7**, 13–26.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser.* **B**, **58**, 267–288.
Yanagihara, H. & Fujisawa, H. (2012). Iterative bias correction of the cross-validation criterion. *Scand. J. Stat.*, **39**, 116–130.
Yanagihara, H., Tonda, T. & Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *J. Multivariate Anal.*, **97**, 1965–1975.
Yanagihara, H., Kamo, K., Imori, S. & Yamamura, M. (2017). A study on the bias-correction effect of the AIC for selecting variables in normal multivariate linear regression models under model misspecification. *REVSTAT-Stat. J.*, **15**, 299–332.
Zhao, L. C., Krishnaiah, P. R. & Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.

Appendix

A. Mathematical Details

A.1. Proof of (7)

It follows from (5) and $= \boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y} = (n-k)s_\Omega^2$ that

$$\frac{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\Omega_i})\boldsymbol{y}}{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y}} = 1 + \frac{T_i}{n-k}, \tag{A.1}$$

where T_i is given by (6). Equation (A.1) and the result $k_{\Omega_i} = k - 1$ imply that

$$\begin{aligned}
nc_{\Omega_i}\frac{s_{\Omega_i}^2}{s_\Omega^2} - nc_\Omega\frac{s_\Omega^2}{s_\Omega^2} &= (n-k)\frac{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\Omega_i})\boldsymbol{y}}{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y}} - (n-k) = T_i, \\
\log\left(c_{\Omega_i}s_{\Omega_i}^2\right) - \log\left(c_\Omega s_\Omega^2\right) &= \log\left\{\frac{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\Omega_i})\boldsymbol{y}/n}{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y}/n}\right\} = \log\left(1 + \frac{T_i}{n-k}\right), \\
-\frac{c_\Omega^2 s_\Omega^2}{c_{\Omega_i}s_{\Omega_i}^2} + \frac{c_\Omega^2 s_\Omega^2}{c_\Omega s_\Omega^2} &= -(n-k)\frac{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y}}{\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_{\Omega_i})\boldsymbol{y}} + (n-k) \\
&= (n-k)\left\{1 - \left(1 + \frac{T_i}{n-k}\right)^{-1}\right\} = T_i\left(1 + \frac{T_i}{n-k}\right)^{-1}.
\end{aligned} \tag{A.2}$$

Moreover, we have

$$\begin{aligned}
\log\left\{\left(c_{\Omega_i} - \frac{\gamma}{n}\right)^{-2}\right\} - \log\left\{\left(c_\Omega - \frac{\gamma}{n}\right)^{-2}\right\} &= -2\log\left(\frac{n-k-\gamma+1}{n-k-\gamma}\right), \\
\frac{2n(k_{\Omega_i}+1)}{n-k_{\Omega_i}-2} - \frac{2n(k_\Omega+1)}{n-k_\Omega-2} &= -\frac{2n(n-1)}{(n-k-1)(n-k-2)}„ \\
\log(c_{\Omega_i}^{-\gamma}) - \log(c_\Omega^{-\gamma}) &= -\gamma\log\left(\frac{n-k+1}{n-k}\right)
\end{aligned} \tag{A.3}$$

Consequently, from (A.2) and (A.3), (7) can be derived.

A.2. Proof of (8)

Notice that for any positive values a and b,

$$\begin{aligned}
f(x) = a\log(1+bx) &\iff f^{-1}(x) = \frac{1}{b}\left\{\exp\left(\frac{x}{a}\right) - 1\right\}, \\
f(x) = x(1+bx)^{-1} &\iff f^{-1}(x) = x(1-bx)^{-1}.
\end{aligned}$$

From the above inverse functions and the definitions of g in (7), (8) can be derived.

A.3. Proof of Independence

Notice that $t_M^{(j)} = \boldsymbol{y}^\top\boldsymbol{\eta}_M^{(j)}$, $s_\Omega^2 = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{y}/(n-k)$ and

$$\begin{aligned}
\boldsymbol{z}_M^{(j)} &= \boldsymbol{y} - \boldsymbol{w}_M^{(j)}t_M^{(j)} = \boldsymbol{y} - \frac{1}{\|\boldsymbol{\eta}_M^{(j)}\|^2}\boldsymbol{\eta}_M^{(j)}\boldsymbol{\eta}_M^{(j)\top}\boldsymbol{y} = \left\{\boldsymbol{I}_n - \boldsymbol{\eta}_M^{(j)}\left(\boldsymbol{\eta}_M^{(j)\top}\boldsymbol{\eta}_M^{(j)}\right)^{-1}\boldsymbol{\eta}_M^{(j)\top}\right\}\boldsymbol{y} = \boldsymbol{H}_M^{(j)}\boldsymbol{y}, \\
\boldsymbol{r}_i^\top\boldsymbol{z}_M^{(j)} &= \boldsymbol{r}_i^\top\boldsymbol{H}_M^{(j)}\boldsymbol{y}.
\end{aligned}$$

It follows from elementary linear algebra and the equation $\boldsymbol{P}_\Omega\boldsymbol{X}_M = \boldsymbol{X}_M$ that

$$\boldsymbol{H}_M^{(j)}\boldsymbol{\eta}_M^{(j)} = \boldsymbol{0}_n, \quad (\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{\eta}_M^{(j)} = (\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{X}_M(\boldsymbol{X}_M^\top\boldsymbol{X}_M)^{-1}\boldsymbol{e}_j = \boldsymbol{0}_n. \tag{A.4}$$

The right-hand side of the above equation implies that

$$(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{H}_M^{(j)} = (\boldsymbol{I}_n - \boldsymbol{P}_\Omega) - (\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{\eta}_M^{(j)}\left(\boldsymbol{\eta}_M^{(j)\top}\boldsymbol{\eta}_M^{(j)}\right)^{-1}\boldsymbol{\eta}_M^{(j)\top} = \boldsymbol{I}_n - \boldsymbol{P}_\Omega. \tag{A.5}$$

From the definition of $\boldsymbol{r}_i$ in (4) and the result $\Omega_i \subset \Omega$, we can see that

$$(\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{r}_i\boldsymbol{r}_i^\top = (\boldsymbol{I}_n - \boldsymbol{P}_\Omega)(\boldsymbol{P}_\Omega - \boldsymbol{P}_{\Omega_i}) = \boldsymbol{O}_{n,n} \iff (\boldsymbol{I}_n - \boldsymbol{P}_\Omega)\boldsymbol{r}_i = \boldsymbol{0}_n. \tag{A.6}$$

Using (A.5) and (A.6), we have

$$\boldsymbol{r}_i^\top \boldsymbol{H}_M^{(j)}(\boldsymbol{I}_n - \boldsymbol{P}_\Omega) = \boldsymbol{r}_i^\top(\boldsymbol{I}_n - \boldsymbol{P}_\Omega) = \boldsymbol{0}_n^\top. \tag{A.7}$$

Independence can be proved from (A.4), (A.7) and Cochran's theorem.