Consistent Model Selection Methods Using Wald Statistics and Estimation of Selection Probability

Tetsuro Sakurai[*], Yasunori Fujikoshi[**] and Takayuki Yamada[***]

[*]*School of General and Management Studies, Suwa University of Science, 5000-1 Toyohira, Chino, Nagano 391-0292, Japan,*

[**]*Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan*

[***]*Faculty of Data Science, Kyoto Women's University, 35 Kitahiyoshi-cho, Imakumano, Higashiyama-ku, Kyoto 605-8501, Japan,*

Abstract

In this paper, we explore variable selection methods in the context of generalized linear models and Cox proportional hazards models. In large-dimensional (LD) setting where the number of explanatory variables is large, we propose the simple knock-one-out (KOO) method based on Wald statistic, and show that the limiting probability of selecting the true model is equal to 1. Additionally, we examine the application of the Bootstrap method to estimate the selection probabilities. Numerical experiments and applications to real datasets are presented to demonstrate the effectiveness of the proposed approach.

AMS 2000 Subject Classification: primary 62H15; secondary 62H10

Key Words and Phrases: Model Selection Criterion, KOO Method, Bootstrap Method, Consistency, Nonlinear regression models, Cox proportional hazard model,

1. Introduction

In this paper, we address the problem of variable selection in generalized linear models and Cox proportional hazards models. Traditional methods such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are commonly used for variable selection. Previous studies by Fujikoshi et al. (2014) and Yanagihara et al. (2015) have demonstrated the consistency of AIC and BIC under certain conditions in the context of multivariate linear regression models. Here, "consistency" refers to the property that the selection criterion identifies the true model with probability 1 in an asymptotic framework.

However, variable selection using criteria like AIC and BIC becomes computationally intensive as the number of explanatory variables increases. Specifically, with k explanatory variables including an intercept, there are $2^k - 1$ candidate models, making exhaustive calculation challenging. One solution to this problem is the Knock-One-Out (KOO) method, introduced by Bai et al. (2024). This method builds on the work of Nishii et al. (1988) and Zhao et al. (1986), who applied BIC in multivariate discriminant analysis and canonical correlation analysis within a large-sample asymptotic framework. Recent advancements by Bai et al. (2024) have extended these results to multivariate regression models under non-normal conditions.

In this paper, we propose a consistent variable selection method using the KOO approach, named the KOO-Wald Method. This method utilizes Wald-type statistics based on the full model with all explanatory variables for variable selection. Unlike traditional KOO-based methods, which require deriving a consistency threshold for each selection criterion, the proposed method overcomes this challenge by providing a unified approach to variable selection that does not require recalculating the threshold. Moreover, the method can be applied to generalized linear models and Cox proportional hazards models, which have not been previously addressed. However, for the proposed method to be consistent, certain conditions such as the asymptotic normality of the maximum likelihood estimator must be satisfied.

The advantages and disadvantages of the proposed KOO-Wald Method are summarized as follows:

- Advantage 1: No need to derive a new threshold d.
- Advantage 2: The method is uniformly applicable across various models.
- Advantage 3: It can be applied to a broader range of analyses, including generalized linear models and Cox proportional hazards models.
- Disadvantage 1: The criterion is fundamentally based on the maximum likelihood estimator, requiring that the estimator satisfies asymptotic normality.

Additionally, we propose a method for estimating selection probabilities using the Bootstrap method. We present numerical experiments and applications to real data to illustrate these concepts. Estimating selection probabilities allows us to assess the importance of each variable, analogous to how weather forecasts provide not just predictions of rain but also the probability of precipitation, helping people decide whether to carry an umbrella.

Furthermore, we discuss the application of the proposed method to more complex model selection problems, along with solutions using the Bootstrap method.

The structure of this paper is as follows: In Section 2, we prepare the notations and symbols used in this paper. In Section 3, we propose the KOO-Wald Method and present the results of numerical experiments. In Section 4, we introduce a method for estimating selection probabilities using the Bootstrap method, along with corresponding numerical results. In Section 5, we apply the proposed methods to real data and present the results. In Section 6, we discuss the application of our method to complex model selection, supported by numerical experiments and real data applications. In Section 7, we conclude with a summary and future research directions.

2. Notations and Preliminaries

In this paper, we consider the problem of variable selection in generalized linear models (GLMs) and Cox proportional hazards models. For both models, let y_i denote the response variable, $\boldsymbol{x}_i : k \times 1$ represent the vector of all explanatory variables, and $\boldsymbol{\beta} : k \times 1$ represent the unknown parameters. Note that the generalized linear model includes an intercept term, whereas the Cox proportional hazards model does not. Specifically, we have:

$$\begin{aligned} \text{Generalized Linear Model:} \quad & \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \\ \text{Cox Proportional Hazards Model:} \quad & \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \end{aligned}$$

For the generalized linear model, the full model M_F using all explanatory variables is expressed as:

$$[M_F] \quad y_i \sim G : \mathrm{E}[y_i|\boldsymbol{x}_i] = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad i = 1, \ldots, n$$

where G denotes an exponential family distribution, and g is the link function.

For the Cox proportional hazards model, the full model M_F using all explanatory variables is expressed as:

$$\begin{aligned} [M_F] \quad y_i \sim G : F(t) = 1 - S(t) = 1 - \int_0^t h(u)\,du, \\ h(t) = h_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad i = 1, \ldots, n \end{aligned}$$

where $F(t)$ is the cumulative distribution function, $S(t)$ is the survival function, $h(t)$ is the hazard function, and $h_0(t)$ is the baseline hazard function.

Let $\boldsymbol{j}$ be a subset of the index set $\omega = \{1, 2, \ldots, k\}$, and let $M_{\boldsymbol{j}}$ denote the model using the explanatory variables corresponding to $\boldsymbol{j}$. The true model, which generates the data, is denoted by M_*, and the corresponding set of explanatory variables is indexed by $\boldsymbol{j}_*$. The collection of all subsets $\boldsymbol{j}$ is denoted by $\mathcal{F}$:

$$\mathcal{F} = \{\{1\}, \ldots, \{k\}, \{1, 2\}, \ldots, \{1, \ldots, k\}\}$$

In this paper, we assume that the true model M_* is included among the candidate models, i.e., $\boldsymbol{j}_* \in \mathcal{F}$. Under this assumption, we consider a variable selection method such that the selected model $\hat{\boldsymbol{j}}$ satisfies:

$$\lim_{n\to\infty} P(\hat{\boldsymbol{j}} = \boldsymbol{j}_*) = 1.$$

3. Main result

The variable selection method proposed in this paper is as follows. Consider the Wald-type statistic T_i for the parameter β_i in the full model M_F:

$$T_i = \frac{\hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}_i)}}}, \quad i = 1, \ldots, k,$$

where $\hat{\beta}_i$ is the maximum likelihood estimator (MLE) of β_i, and $\widehat{\mathrm{Var}(\hat{\beta}_i)}$ is the estimator of $\mathrm{Var}(\hat{\beta}_i)$.

The decision rule is given by:

$$T_i > n^{1/4} \Rightarrow \beta_i > 0, \quad -n^{1/4} \leq T_i \leq n^{1/4} \Rightarrow \beta_i = 0, \quad T_i < -n^{1/4} \Rightarrow \beta_i < 0,$$

where n denotes the sample size. This decision rule allows us to determine not only the inclusion of a variable but also its sign. The selected model $\hat{\boldsymbol{j}}$ is thus given by:

$$\hat{\boldsymbol{j}} = \{j \in \omega \mid |T_j| > n^{1/4}\}.$$

This method is consistent under the following assumptions A0 to A5:

A0 Model Assumption: The true model is included within the full model.

A1 Sample Size and Parameter Assumptions:

$n \to \infty, \quad k = O(n^a), \quad 0 \leq a < \frac{1}{2}.$

A2 Consistency of the Estimator (1): Under A1,

$\hat{\beta}_i = \beta_i + O_p(n^{-1/2}).$

A3 Consistency of the Estimator (2): Under A1, for any $\varepsilon > 0$,

$P(|\hat{\beta}_i - \beta_i| \geq \varepsilon) = O(n^{-1/2}).$

A4 Variance of the Estimator: Under A1,

$\mathrm{Var}(\hat{\beta}_i) = O(n^{-1}).$

A5 Wald-type Probability Evaluation: Under A1,

$$P\left(\frac{\hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}_i)}}} \geq x\right) = P\left(\frac{\hat{\beta}_i}{\sqrt{\mathrm{Var}(\hat{\beta}_i)}} \geq x\right) + O(n^{-1/2}).$$

Here, k is the number of variables considered in the selection process. The assumptions A2 to A5 are properties that the MLE satisfies under large sample sizes. For more details on the properties of the MLE in generalized linear models and other contexts under large samples, see Dobson et al. (2008).

Under these assumptions, the following theorem holds:

Theorem 3.1. *Under assumptions A0 to A5, the selected model* $\hat{\boldsymbol{j}} = \{j \in \omega \mid |T_j| > n^{1/4}\}$ *satisfies:*

$$\lim_{n \to \infty} P(\hat{\boldsymbol{j}} = \boldsymbol{j}_*) = 1.$$

3.1. Numerical Experiments: Verification of Consistency

In this section, we verify the consistency of the proposed variable selection method through numerical experiments using regression models, logistic regression models, Poisson regression models, and Cox proportional hazards models. The models are specified as follows. Let $\boldsymbol{y} : n \times 1$ denote the sample of the response variable, where the i-th observation is y_i. Let $\boldsymbol{\beta}_1 : k_1 \times 1$ represent the non-zero coefficients in the true model, and let the full model's coefficients be $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \mathbf{0}')' : k \times 1$. Let $\mathbf{X} : n \times k$ denote the matrix of explanatory variables for the full model, with $\boldsymbol{x}_i$ representing the explanatory variables corresponding to y_i. The models can be expressed as follows:

$$
\begin{aligned}
\text{Regression Model:} \quad & y_i \sim N(\mu_i, \sigma^2), \\
& \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \ldots, n \\
\text{Logistic Regression Model:} \quad & y_i \sim \mathrm{Bern}(p_i), \\
& p_i = \frac{1}{1 + \exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}, \quad i = 1, \ldots, n \\
\text{Poisson Regression Model:} \quad & y_i \sim \mathrm{Po}(\lambda_i), \\
& \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad i = 1, \ldots, n \\
\text{Cox Proportional Hazards Model:} \quad & y_i \sim F(t) = 1 - S(t) = 1 - \int_0^t h(u)\, du, \\
& h(t) = h_0(t) \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad i = 1, \ldots, n
\end{aligned}
$$

Here, $\mathrm{Bern}(p)$ denotes the Bernoulli distribution with success probability p, and $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with mean λ. In the Cox proportional hazards model, $F(t)$ represents the cumulative distribution function, $S(t)$ the survival function, $h(t)$ the hazard function, and $h_0(t)$ the baseline hazard function.

The proposed variable selection method is based on the Wald statistic:

$$
T_i = \frac{\hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}}, \quad i = 1, \ldots, k,
$$

and the decision rule is as follows:

$$
T_i > n^{1/4} \Rightarrow \beta_i > 0, \quad -n^{1/4} < T_i < n^{1/4} \Rightarrow \beta_i = 0, \quad T_i < -n^{1/4} \Rightarrow \beta_i < 0,
$$

where n denotes the sample size. Under these conditions, we confirm the consistency of the method by demonstrating that the probability of selecting the true model approaches 1 as the sample size increases.

As a comparison, we also examine the consistency of variable selection methods using the KOO method, stepwise selection (both forward and backward with AIC and BIC), and lasso regression. The KOO method's decision rules using AIC and BIC are as follows:

$$\text{KOO method:}\quad \mathrm{AIC}_{-j} - \mathrm{AIC} > 0 \Rightarrow \beta_j \neq 0, \quad \mathrm{BIC}_{-j} - \mathrm{BIC} > 0 \Rightarrow \beta_j \neq 0.$$

For stepwise selection, we conducted numerical experiments using the full model as the starting point, iteratively removing one explanatory variable and selecting the model that minimizes the criterion until no further reduction in AIC or BIC is possible.

Lasso regression, which performs both estimation and variable selection, was also evaluated. To determine the penalty parameter λ, 10-fold cross-validation was used. We examined the regression coefficients for two values of λ: λ_{min}, where the error was minimized, and λ_{1se}, where λ was within one standard error of the minimum error. The lasso computations were performed using the R package 'glmnet', which provides lasso regression for logistic, Poisson, and Cox proportional hazards models. For details on the glmnet method, see Friedman et al. (2010).

The numerical experiments were conducted under the following conditions:

- Number of simulations: 10^4
- Values of $\boldsymbol{\beta}$: The values were evenly divided between 1 and 2 across k_1 components, and those values alternated in sign, i.e., $\beta_j = (-1)^{j-1}\left(1 + \frac{j-1}{k_1-1}\right)$
- Values of explanatory variables: The values of $\mathbf{X} : n \times k$ were generated from a uniform distribution on $(-1, 1)$
- For the Cox proportional hazards model, random numbers were generated from a Weibull distribution $W(m, \eta)$. The shape parameter was

set to $m = 3$, and $\eta = \exp(-1/m\boldsymbol{x}'\boldsymbol{\beta})$. It is known that the hazard function in this case is expressed as:

$$h(t) = m\left(\frac{t}{\eta}\right)^m = m\left(\frac{t}{\boldsymbol{x}'\boldsymbol{\beta}}\right)^m.$$

The results for each model are presented below.

The following are the results for the regression model.

Regression			Wald				KOO		Step		Lasso	
n	k	k_1	$\beta > 0$	$\beta < 0$	$\beta = 0$	all	AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}
100	10	0	0.00	0.00	0.98	0.98	0.15	0.66	0.00	0.00	0.00	0.00
	10	5	1.00	1.00	0.99	0.99	0.39	0.81	0.39	0.82	0.07	0.66
	10	10	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
500	10	0	0.00	0.00	1.00	**1.00**	0.17	0.87	0.00	0.00	0.00	0.00
	10	5	1.00	1.00	1.00	**1.00**	0.43	0.93	0.43	0.93	0.07	0.93
	10	10	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
200	50	0	0.00	0.00	0.99	0.99	0.00	0.13	0.00	0.00	0.00	0.00
	50	25	1.00	1.00	0.99	0.99	0.00	0.33	0.00	0.42	0.00	0.00
	50	50	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
1000	50	0	0.00	0.00	1.00	**1.00**	0.00	0.60	0.00	0.00	0.00	0.00
	50	25	1.00	1.00	1.00	**1.00**	0.01	0.76	0.01	0.78	0.00	0.04
	50	50	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

The following are the results for the logistic regression model.

logistic			Wald				KOO		Step		Lasso	
n	k	k_1	$\beta > 0$	$\beta < 0$	$\beta = 0$	all	AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}
100	10	0	0.00	0.00	0.99	0.99	0.16	0.67	0.00	0.00	0.01	0.01
	10	5	0.16	0.11	1.00	0.02	0.30	0.41	0.31	0.46	0.06	0.33
	10	10	0.00	0.00	0.00	0.00	0.42	0.09	0.43	0.10	0.86	0.42
500	10	0	0.00	0.00	1.00	**1.00**	0.17	0.88	0.00	0.00	0.00	0.00
	10	5	0.99	0.87	1.00	0.86	0.42	0.93	0.42	0.93	0.03	0.79
	10	10	0.57	0.34	0.00	0.20	**1.00**	0.99	**1.00**	0.99	**1.00**	**1.00**
200	50	0	0.00	0.00	0.97	0.97	0.00	0.12	0.00	0.00	0.00	0.00
	50	25	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	50	0	0.00	0.00	1.00	**1.00**	0.00	0.60	0.00	0.02	0.00	0.00
	50	25	0.10	0.05	1.00	0.01	0.01	0.71	0.01	0.76	0.00	0.00
	50	50	0.00	0.00	0.00	0.00	0.99	0.82	0.99	0.82	**1.00**	**1.00**

In the case of the logistic regression model, the convergence was slow, so additional numerical experiments were conducted with an increased sample size n.

logistic			Wald				KOO		Step		Lasso	
n	k	k_1	$\beta > 0$	$\beta < 0$	$\beta = 0$	all	AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}
1000	10	0	0.00	0.00	1.00	**1.00**	0.18	0.91	0.00	0.00	0.00	0.00
	10	5	1.00	1.00	1.00	**1.00**	0.41	0.96	0.42	0.96	0.03	0.87
	10	10	0.95	0.84	0.00	0.81	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
5000	50	0	0.00	0.00	1.00	**1.00**	0.00	0.83	0.00	0.01	0.00	0.00
	50	25	1.00	1.00	1.00	**1.00**	0.01	0.91	0.01	0.91	0.00	0.00
	50	50	0.93	0.88	0.00	0.82	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

The following are the results for the poisson regression model.

poisson			Wald				KOO		Step		Lasso	
n	k	k_1	$\beta > 0$	$\beta < 0$	$\beta = 0$	all	AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}
100	10	0	0.00	0.00	0.98	0.98	0.18	0.71	0.00	0.00	0.00	0.00
	10	5	1.00	1.00	0.99	0.99	0.45	0.86	0.43	0.85	0.02	0.15
	10	10	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99
500	10	0	0.00	0.00	1.00	**1.00**	0.19	0.88	0.00	0.00	0.00	0.00
	10	5	1.00	1.00	1.00	**1.00**	0.43	0.94	0.42	0.94	0.02	0.56
	10	10	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
200	50	0	0.00	0.00	0.98	0.98	0.00	0.28	0.00	0.00	0.00	0.00
	50	25	1.00	1.00	1.00	**1.00**	0.05	0.65	0.02	0.65	0.00	0.00
	50	50	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.01	0.00
1000	50	0	0.00	0.00	1.00	**1.00**	0.00	0.64	0.00	0.00	0.00	0.00
	50	25	1.00	1.00	1.00	**1.00**	0.03	0.82	0.02	0.82	0.00	0.00
	50	50	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.03	0.03

The following are the results for the Cox proportional hazards model.

Cox			Wald				KOO		Step		Lasso	
n	k	k_1	$\beta > 0$	$\beta < 0$	$\beta = 0$	all	AIC	BIC	AIC	BIC	$\lambda_{\min}$	$\lambda_{1\text{se}}$
100	10	0	0.00	0.00	0.97	0.97	0.14	0.63	0.15	0.70	0.46	0.60
	10	5	0.96	1.00	0.99	0.95	0.38	0.81	0.38	0.82	0.02	0.48
	10	10	0.96	0.99	0.00	0.94	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
500	10	0	0.00	0.00	1.00	**1.00**	0.17	0.87	0.18	0.88	0.51	0.69
	10	5	1.00	1.00	1.00	**1.00**	0.41	0.93	0.41	0.93	0.01	0.69
	10	10	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
200	50	0	0.00	0.00	0.90	0.90	0.00	0.07	0.00	0.29	0.42	0.56
	50	25	0.99	1.00	0.96	0.96	0.00	0.31	0.00	0.41	0.00	0.00
	50	50	0.99	0.99	0.00	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
1000	50	0	0.00	0.00	1.00	**1.00**	0.00	0.56	0.00	0.64	0.47	0.65
	50	25	1.00	1.00	1.00	**1.00**	0.01	0.76	0.01	0.78	0.00	0.00
	50	50	1.00	1.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

In the table, n represents the sample size, k denotes the total number of explanatory variables, and k_1 is the number of parameters for which $\beta_i \neq 0$. The column "Wald $\beta > 0$" indicates the probability that all positive coefficients were correctly identified as positive, "Wald $\beta < 0$" indicates the probability that all negative coefficients were correctly identified as negative, and "Wald $\beta = 0$" shows the probability that all zero coefficients were correctly identified as zero. The column "Wald all" represents the probability of selecting the true model.

The columns labeled "KOO" represent the probability of selecting the true model using AIC and BIC in the KOO method, while "Step" refers to the probability of selecting the true model using AIC and BIC in stepwise selection. The "Lasso" column shows the probability of selecting the true model when using lasso regression with hyperparameters $\lambda_{\min}$ and $\lambda_{1\text{se}}$. Here, $\lambda_{\min}$ is the value of λ that minimizes the error in 10-fold cross-validation, and $\lambda_{1\text{se}}$ is the largest λ within one standard error of the minimum error. Boldface in the table indicates a probability of 1 for selecting the true model.

These results demonstrate that consistency is achieved as the sample size increases for each model. However, in the case of the logistic regression model, convergence is slower compared to the other models, requiring a larger sample size for consistent results.

4. Estimation of Selection Probabilities

In this section, we propose a method for estimating the selection probabilities of variables using the variable selection method introduced in this paper. The probability that a coefficient β_i is determined to be non-zero by the proposed method can be expressed as follows:

$$p_i = P\left(|T_i| > n^{1/4}\right), \quad i = 1, \dots, k.$$

This probability p_i can be estimated using the Bootstrap method as follows. Let B denote the number of bootstrap samples, and let $T_{ij}^{\#}$ represent the value of T_i computed from the j-th bootstrap sample. Then, the estimator $\hat{p}_i$ for p_i is given by:

$$\hat{p}_i = \frac{1}{B}\sum_{j=1}^{B} I(|T_{ij}^{\#}| > n^{1/4}),$$

where the indicator function $I(|T_{ij}^{\#}| > n^{1/4})$ is defined as:

$$I(|T_{ij}^{\#}| > n^{1/4}) = \begin{cases} 1 & \text{if } |T_{ij}^{\#}| > n^{1/4}, \\ 0 & \text{if } |T_{ij}^{\#}| \le n^{1/4}. \end{cases}$$

In other words, $\hat{p}_i$ corresponds to the proportion of bootstrap samples in which the coefficient β_i is judged to be non-zero.

Next, we introduce the following assumptions:

B1 Asymptotic Normality: Under A1,

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}} \xrightarrow{d} N(0,1).$$

B2 Bootstrap Distribution: Under A1,

$$P\left(\frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}} < x\right) = P\left(\frac{\hat{\beta}_i^{\#} - \hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}} < x \;\middle|\; \boldsymbol{X}\right) + O(n^{-1}).$$

Assumption B1 is a property satisfied by the maximum likelihood estimator under large samples. Assumption B2 is commonly assumed when applying the Bootstrap method in large samples. For more details on the Bootstrap method in large samples, see Hall (1992).

Under these conditions, the following theorem holds:

Theorem 4.1. *Under assumptions A0–A5 and B1–B2, the median m of $\hat{p}_i$ approximates the selection probability p_i. Specifically, it satisfies:*

$$P(\hat{p}_i \leq p_i) = P(\hat{p}_i \geq p_i) = \frac{1}{2} + o(1),$$

where the order term $o(1)$ is related to n and B.

4.1. Numerical Experiment: Estimation of Selection Probabilities

In this section, we conduct numerical experiments to estimate selection probabilities under the regression model, logistic regression model, Poisson regression model, and Cox proportional hazards model, following the same approach as in the verification of consistency. The numerical experiments were conducted under the following conditions:

- Number of simulations: 10^4
- Values of $\boldsymbol{\beta}$: To observe the behavior of selection probabilities, small values were chosen, specifically $\beta_j = (-1)^{j-1}\frac{1}{4}\left(1 + \frac{j-1}{k_1-1}\right)$
- Values of explanatory variables: The matrix $\mathbf{X} : n \times k$ was generated from a uniform distribution on $(-1, 1)$.
- Cox proportional hazards model: The same setup as in the consistency verification.
- Number of bootstrap samples: 10^3

The results obtained under these settings are presented below.

The following are the results for the regression model.

regression: $k = 10, k_1 = 5$

		$n = 200, k = 10, k_1 = 5$				$n = 1000, k = 10, k_1 = 5$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.39	0.42	0.39	0.29	0.99	0.94	0.99	0.11
β_2	-0.31	0.11	0.19	0.10	0.22	0.53	0.52	0.53	0.29
β_3	0.38	0.22	0.30	0.22	0.26	0.88	0.80	0.89	0.22
β_4	-0.44	0.40	0.43	0.40	0.29	0.99	0.95	0.99	0.10
β_5	0.50	0.58	0.56	0.58	0.29	1.00	0.99	1.00	0.03
β_6	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_7	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_8	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_9	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_{10}	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00

Here, β_i represents the value of the coefficient for $i = 1$ to $i = k$. When the coefficient β_i is zero, it indicates that the corresponding explanatory variable is unnecessary. The probability p_i represents the proportion of numerical experiments in which T_i was determined to be non-zero, thus providing an estimate of the true selection probability. Additionally, "mean" refers to the average value of $\hat{p}_i$ estimated from the bootstrap samples, "median" refers to the median of $\hat{p}_i$, and "sd" refers to the standard deviation of $\hat{p}_i$.

The average and median of $\hat{p}_i$ are approximately as follows:

$$\text{Mean of } \hat{p}_i \approx \mathrm{E}[\hat{p}_i], \qquad \text{Median of } \hat{p}_i = m \approx P(\hat{p}_i \leq m) = \frac{1}{2}.$$

These results suggest that the bootstrap estimator $\hat{p}_i$ tends to exhibit unbiasedness with respect to the median rather than the expectation.

The results of the numerical experiments conducted with larger values of k and k_1 are as follows.

regression:$k = 50, k_1 = 25$

		$n = 200$				$n = 1000$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.25	0.32	0.26	0.26	0.98	0.93	0.98	0.12
β_2	-0.26	0.03	0.10	0.04	0.14	0.16	0.25	0.17	0.24
β_3	0.27	0.04	0.11	0.05	0.15	0.21	0.29	0.21	0.25
β_4	-0.28	0.04	0.12	0.05	0.16	0.28	0.33	0.27	0.27
β_5	0.29	0.05	0.13	0.06	0.16	0.33	0.38	0.33	0.28
β_6	-0.30	0.06	0.14	0.07	0.17	0.41	0.44	0.41	0.29
β_7	0.31	0.07	0.16	0.08	0.19	0.47	0.48	0.47	0.29
β_8	-0.32	0.08	0.16	0.09	0.19	0.54	0.53	0.55	0.29
β_9	0.33	0.09	0.18	0.10	0.20	0.63	0.59	0.63	0.28
β_{10}	-0.34	0.10	0.19	0.11	0.20	0.69	0.63	0.69	0.28
β_{11}	0.35	0.11	0.21	0.13	0.21	0.75	0.68	0.75	0.26
β_{12}	-0.36	0.12	0.22	0.14	0.22	0.81	0.73	0.81	0.25
β_{13}	0.38	0.14	0.23	0.16	0.23	0.85	0.77	0.85	0.24
β_{14}	-0.39	0.16	0.25	0.18	0.23	0.89	0.81	0.89	0.21
β_{15}	0.40	0.17	0.26	0.19	0.24	0.92	0.84	0.92	0.20
β_{16}	-0.41	0.20	0.28	0.21	0.24	0.94	0.87	0.95	0.18
β_{17}	0.42	0.22	0.30	0.23	0.25	0.96	0.89	0.96	0.16
β_{18}	-0.43	0.24	0.32	0.26	0.26	0.97	0.92	0.98	0.14
β_{19}	0.44	0.26	0.33	0.28	0.26	0.98	0.93	0.98	0.12
β_{20}	-0.45	0.29	0.35	0.29	0.26	0.99	0.95	0.99	0.10
β_{21}	0.46	0.31	0.37	0.32	0.27	0.99	0.96	0.99	0.08
β_{22}	-0.47	0.35	0.39	0.35	0.27	1.00	0.97	1.00	0.07
β_{23}	0.48	0.37	0.41	0.37	0.28	1.00	0.98	1.00	0.06
β_{24}	-0.49	0.39	0.43	0.40	0.28	1.00	0.98	1.00	0.05
β_{25}	0.50	0.42	0.45	0.42	0.28	1.00	0.99	1.00	0.04
β_{26}	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{50}	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00

Here, the results for β_{27} through β_{49} are omitted as they were similar to those for β_{26} and β_{50}.

The following are the results for the logistic regression model.

logistic:$k = 10, k_1 = 5$

		$n = 200, k = 10, k_1 = 5$				$n = 1000, k = 10, k_1 = 5$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.02	0.10	0.04	0.14	0.03	0.11	0.04	0.15
β_2	-0.31	0.01	0.05	0.02	0.10	0.00	0.02	0.00	0.06
β_3	0.38	0.01	0.07	0.03	0.12	0.01	0.05	0.01	0.10
β_4	-0.44	0.02	0.10	0.04	0.14	0.03	0.11	0.04	0.15
β_5	0.50	0.03	0.13	0.07	0.16	0.10	0.20	0.12	0.21
β_6	0.00	0.00	0.02	0.00	0.03	0.00	0.00	0.00	0.00
β_7	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_8	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_9	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_{10}	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00

logistic: $k = 50, k_1 = 25$

		$n = 200$				$n = 1000$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.02	0.61	0.61	0.20	0.02	0.11	0.04	0.15
β_2	-0.26	0.00	0.55	0.54	0.19	0.00	0.01	0.00	0.04
β_3	0.27	0.00	0.56	0.55	0.20	0.00	0.02	0.00	0.04
β_4	-0.28	0.00	0.55	0.54	0.19	0.00	0.02	0.00	0.05
β_5	0.29	0.01	0.56	0.55	0.20	0.00	0.02	0.00	0.05
β_6	-0.30	0.01	0.56	0.55	0.19	0.00	0.02	0.00	0.06
β_7	0.31	0.01	0.57	0.56	0.20	0.00	0.03	0.00	0.07
β_8	-0.32	0.01	0.57	0.56	0.20	0.00	0.03	0.01	0.07
β_9	0.33	0.01	0.57	0.57	0.20	0.00	0.04	0.01	0.08
β_{10}	-0.34	0.01	0.58	0.57	0.20	0.00	0.04	0.01	0.08
β_{11}	0.35	0.01	0.58	0.57	0.20	0.00	0.05	0.01	0.09
β_{12}	-0.36	0.01	0.58	0.58	0.20	0.00	0.05	0.01	0.10
β_{13}	0.38	0.01	0.59	0.58	0.20	0.00	0.06	0.02	0.11
β_{14}	-0.39	0.01	0.59	0.58	0.20	0.01	0.07	0.02	0.12
β_{15}	0.40	0.01	0.59	0.59	0.20	0.01	0.07	0.02	0.12
β_{16}	-0.41	0.01	0.60	0.60	0.20	0.01	0.08	0.03	0.13
β_{17}	0.42	0.01	0.60	0.60	0.20	0.01	0.09	0.03	0.13
β_{18}	-0.43	0.02	0.60	0.60	0.20	0.01	0.10	0.04	0.14
β_{19}	0.44	0.01	0.61	0.61	0.20	0.02	0.11	0.05	0.15
β_{20}	-0.45	0.02	0.61	0.62	0.20	0.02	0.12	0.05	0.16
β_{21}	0.46	0.02	0.62	0.62	0.20	0.02	0.13	0.06	0.17
β_{22}	-0.47	0.02	0.63	0.63	0.20	0.03	0.14	0.07	0.17
β_{23}	0.48	0.02	0.62	0.63	0.20	0.04	0.16	0.09	0.18
β_{24}	-0.49	0.03	0.63	0.64	0.20	0.04	0.17	0.09	0.19
β_{25}	0.50	0.03	0.64	0.65	0.20	0.05	0.19	0.11	0.20
β_{26}	0.00	0.00	0.51	0.50	0.18	0.00	0.00	0.00	0.00
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{50}	0.00	0.00	0.52	0.50	0.18	0.00	0.00	0.00	0.00

The following are the results for the poisson regression model.

poisson:$k = 10, k_1 = 5$

		$n = 200, k = 10, k_1 = 5$				$n = 1000, k = 10, k_1 = 5$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.40	0.36	0.30	0.30	0.99	0.96	1.00	0.10
β_2	-0.31	0.20	0.28	0.21	0.25	0.86	0.78	0.87	0.23
β_3	0.38	0.40	0.44	0.41	0.28	0.99	0.96	0.99	0.09
β_4	-0.44	0.64	0.60	0.63	0.28	1.00	1.00	1.00	0.02
β_5	0.50	0.81	0.74	0.82	0.25	1.00	1.00	1.00	0.00
β_6	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00
β_7	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00
β_8	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00
β_9	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00
β_{10}	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00

poisson:$k = 50, k_1 = 25$

		$n = 200$				$n = 1000$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.07	0.05	0.02	0.10	0.95	0.83	0.92	0.21
β_2	-0.26	0.09	0.23	0.17	0.20	0.87	0.79	0.87	0.22
β_3	0.27	0.10	0.25	0.19	0.21	0.92	0.83	0.91	0.20
β_4	-0.28	0.12	0.26	0.20	0.21	0.95	0.87	0.95	0.17
β_5	0.29	0.14	0.28	0.23	0.22	0.97	0.91	0.97	0.15
β_6	-0.30	0.16	0.30	0.25	0.23	0.98	0.93	0.98	0.12
β_7	0.31	0.18	0.32	0.27	0.23	0.99	0.96	0.99	0.09
β_8	-0.32	0.21	0.34	0.30	0.24	1.00	0.97	1.00	0.07
β_9	0.33	0.24	0.36	0.32	0.25	1.00	0.98	1.00	0.05
β_{10}	-0.34	0.27	0.38	0.35	0.25	1.00	0.99	1.00	0.04
β_{11}	0.35	0.29	0.40	0.36	0.25	1.00	0.99	1.00	0.03
β_{12}	-0.36	0.34	0.43	0.40	0.25	1.00	1.00	1.00	0.02
β_{13}	0.38	0.37	0.45	0.43	0.26	1.00	1.00	1.00	0.02
β_{14}	-0.39	0.40	0.47	0.46	0.26	1.00	1.00	1.00	0.01
β_{15}	0.40	0.44	0.50	0.49	0.26	1.00	1.00	1.00	0.01
β_{16}	-0.41	0.47	0.51	0.52	0.26	1.00	1.00	1.00	0.01
β_{17}	0.42	0.51	0.54	0.55	0.26	1.00	1.00	1.00	0.00
β_{18}	-0.43	0.54	0.56	0.58	0.26	1.00	1.00	1.00	0.00
β_{19}	0.44	0.59	0.59	0.61	0.26	1.00	1.00	1.00	0.00
β_{20}	-0.45	0.61	0.61	0.64	0.25	1.00	1.00	1.00	0.00
β_{21}	0.46	0.64	0.63	0.66	0.25	1.00	1.00	1.00	0.00
β_{22}	-0.47	0.68	0.65	0.70	0.25	1.00	1.00	1.00	0.00
β_{23}	0.48	0.72	0.67	0.73	0.24	1.00	1.00	1.00	0.00
β_{24}	-0.49	0.74	0.69	0.74	0.24	1.00	1.00	1.00	0.00
β_{25}	0.50	0.77	0.71	0.77	0.23	1.00	1.00	1.00	0.00
β_{26}	0.00	0.00	0.03	0.02	0.04	0.00	0.00	0.00	0.00
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{50}	0.00	0.00	0.03	0.02	0.05	0.00	0.00	0.00	0.00

The following are the results for the Cox proportional hazards model.

Cox:$k = 10, k_1 = 5$

		$n = 200, k = 10, k_1 = 5$				$n = 1000, k = 10, k_1 = 5$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.04	0.14	0.06	0.17	0.13	0.22	0.15	0.23
β_2	-0.31	0.10	0.22	0.14	0.22	0.50	0.51	0.52	0.29
β_3	0.38	0.21	0.33	0.26	0.26	0.86	0.79	0.87	0.22
β_4	-0.44	0.37	0.45	0.42	0.28	0.99	0.94	0.99	0.11
β_5	0.50	0.55	0.58	0.61	0.28	1.00	0.99	1.00	0.04
β_6	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_7	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_8	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_9	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00
β_{10}	0.00	0.00	0.02	0.00	0.04	0.00	0.00	0.00	0.00

Cox:$k = 50, k_1 = 25$

		$n = 200$				$n = 1000$			
	β_i	p_i	mean	median	sd	p_i	mean	median	sd
β_1	0.25	0.06	0.27	0.21	0.21	0.13	0.25	0.17	0.24
β_2	-0.26	0.06	0.28	0.23	0.21	0.17	0.29	0.22	0.25
β_3	0.27	0.08	0.30	0.24	0.22	0.22	0.34	0.27	0.27
β_4	-0.28	0.08	0.31	0.26	0.22	0.27	0.38	0.34	0.28
β_5	0.29	0.09	0.33	0.28	0.23	0.33	0.43	0.41	0.28
β_6	-0.30	0.10	0.34	0.29	0.23	0.40	0.48	0.47	0.29
β_7	0.31	0.12	0.35	0.31	0.24	0.47	0.53	0.54	0.29
β_8	-0.32	0.12	0.36	0.32	0.24	0.53	0.58	0.60	0.28
β_9	0.33	0.14	0.38	0.34	0.24	0.60	0.62	0.67	0.28
β_{10}	-0.34	0.16	0.40	0.37	0.24	0.68	0.68	0.74	0.26
β_{11}	0.35	0.17	0.41	0.38	0.24	0.74	0.72	0.79	0.25
β_{12}	-0.36	0.19	0.43	0.41	0.25	0.79	0.76	0.84	0.24
β_{13}	0.38	0.21	0.44	0.42	0.25	0.84	0.79	0.87	0.22
β_{14}	-0.39	0.22	0.46	0.44	0.25	0.88	0.83	0.91	0.20
β_{15}	0.40	0.26	0.48	0.47	0.25	0.91	0.86	0.93	0.18
β_{16}	-0.41	0.26	0.49	0.48	0.25	0.94	0.89	0.95	0.16
β_{17}	0.42	0.29	0.51	0.51	0.25	0.95	0.90	0.97	0.15
β_{18}	-0.43	0.32	0.53	0.53	0.25	0.97	0.93	0.98	0.13
β_{19}	0.44	0.34	0.54	0.55	0.25	0.98	0.94	0.99	0.11
β_{20}	-0.45	0.36	0.56	0.58	0.25	0.99	0.95	0.99	0.09
β_{21}	0.46	0.38	0.57	0.59	0.25	0.99	0.97	1.00	0.08
β_{22}	-0.47	0.41	0.59	0.61	0.25	0.99	0.97	1.00	0.07
β_{23}	0.48	0.43	0.60	0.63	0.25	1.00	0.98	1.00	0.05
β_{24}	-0.49	0.45	0.61	0.64	0.25	1.00	0.99	1.00	0.04
β_{25}	0.50	0.48	0.64	0.67	0.24	1.00	0.99	1.00	0.04
β_{26}	0.00	0.00	0.11	0.07	0.10	0.00	0.00	0.00	0.00
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{50}	0.00	0.00	0.11	0.07	0.10	0.00	0.00	0.00	0.00

From these results, it is evident that as the sample size increases, the bootstrap-based estimation of selection probabilities asymptotically becomes a median-unbiased estimator. Among the models, the regression model consistently converged faster than the others. The Poisson regression model followed, showing relatively fast convergence. In the Cox proportional hazards model, the convergence was comparable to that of the Poisson regression model when k was small; however, as k increased, the performance was not as strong as that of the Poisson regression model. Consistent with the results for consistency, the logistic regression model exhibited slower convergence compared to the other models.

5. Application to Real Data

In this section, we describe the results of applying the proposed methods to real datasets using linear regression, logistic regression, Poisson regression, and Cox proportional hazards models. For each model, we identified the selected explanatory variables using the following methods:

- The Wald-type KOO method proposed in this paper
- The KOO method using AIC or BIC
- Stepwise selection using AIC or BIC
- Lasso regression with λ_{min} or λ_{1se}

For the Wald-type KOO method, we also estimated the selection probabilities of the coefficients.

5.1. Linear Regression Model

We applied the linear regression model to the Boston housing dataset, which is provided by the MASS library in R (Venables and Ripley, 2002). This dataset contains information on housing prices in Boston with $n = 506$ observations. The response variable is the housing price, and the following explanatory variables were used in the linear regression model:

- crim: Crime rate
- zn: Proportion of residential land zoned for large lots

- indus: Proportion of non-retail business acres per town
- chas: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- nox: Nitric oxides concentration (parts per 10 million)
- rm: Average number of rooms per dwelling
- age: Proportion of owner-occupied units built before 1940
- dis: Weighted distances to five Boston employment centers
- rad: Index of accessibility to radial highways
- tax: Full-value property-tax rate per $10,000
- ptratio: Pupil-teacher ratio by town
- black: Proportion of Black residents
- lstat: Percentage of lower status of the population

The results are as follows.

$n = 506$	Wald	KOO		Step		Lasso		Wald
$k = 14$		AIC	BIC	AIC	BIC	λ_{min}	λ_{1se}	prob.
Intercept	24.47	36.34	36.34	36.34	36.34	34.36	17.53	0.93
crim		-0.11	-0.11	-0.11	-0.11	-0.10	-0.02	0.04
zn		0.05	0.05	0.05	0.05	0.04		0.09
indus								0.00
chas		2.72	2.72	2.72	2.72	2.68	1.90	0.16
nox		-17.38	-17.38	-17.38	-17.38	-16.25	-3.38	0.42
rm	4.22	3.80	3.80	3.80	3.80	3.87	4.26	0.99
age								0.00
dis	-0.55	-1.49	-1.49	-1.49	-1.49	-1.39	-0.32	1.00
rad		0.30	0.30	0.30	0.30	0.25		0.43
tax		-0.01	-0.01	-0.01	-0.01	-0.01		0.03
ptratio	-0.97	-0.95	-0.95	-0.95	-0.95	-0.93	-0.79	1.00
black		0.01	0.01	0.01	0.01	0.01	0.01	0.10
lstat	-0.67	-0.52	-0.52	-0.52	-0.52	-0.52	-0.52	1.00

A blank entry indicates that the variable was not selected. From these results, it is evident that the Wald-type KOO method constructs the smallest model, and the coefficients selected by this method are always included in the models selected by the other methods.

5.2. Poisson Regression Model

Here, we applied the Poisson regression model to the dataset of doctoral students' publication records from Long (1997). This dataset consists of $n = 915$ doctoral students majoring in biochemistry. The response variable is the number of papers published by each student by the time they completed their doctoral program. The following explanatory variables were used in the Poisson regression model:

- male: Dummy variable indicating gender (1 = male, 0 = female)
- married: Dummy variable indicating marital status (1 = married, 0 = not married)
- kids: Number of children under age 6
- prestige: Prestige of the graduate program
- mentor: Number of papers published by the student's mentor

The results are as follows.

$n = 915$	Wald	KOO		Step		Lasso		Wald
$k = 6$		AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}	prob.
Intercept				0.12	0.19	0.17	0.45	0.00
male		0.27	0.38	0.23	0.24	0.19		0.15
married		0.23		0.15		0.10		0.01
kids		-0.19	-0.12	-0.18	-0.14	-0.14		0.23
prestige								0.00
mentor	0.04	0.03	0.03	0.03	0.03	0.02	0.01	1.00

A blank entry indicates that the variable was not selected. These results demonstrate that the Wald-type KOO method constructs the smallest model, and the coefficients selected by this method are always included in the models selected by the other methods.

5.3. Cox Proportional Hazards Model

In this section, we applied the Cox proportional hazards model to the kidney dataset, which is available in the 'survival' package in R. This dataset includes information on the recurrence times of infections in kidney patients using portable dialysis equipment (Therneau and Grambsch, 2000). The analysis was conducted on $n = 76$ recurrence time data points using the following explanatory variables in the Cox proportional hazards model:

- age: Age
- sex: Gender (1 = male, 2 = female)
- GN: Dummy variable for disease type (1 = glomerulonephritis)
- AN: Dummy variable for disease type (1 = acute nephritis)
- PKD: Dummy variable for disease type (1 = polycystic kidney disease)
- frail: Estimated frailty score from the original study

Here, if all dummy variables GN, AN, and PKD are zero, the patient is classified under "Other" diseases. The results obtained are as follows.

$n = 76$	Wald	KOO		Step		Lasso		Wald
$k = 6$		AIC	BIC	AIC	BIC	$\lambda_{\min}$	λ_{1se}	prob.
age						0.01		0.02
sex	-1.89	-2.11	-1.89	-2.11	-1.89	-1.92	-1.06	0.99
GN						0.08		0.02
AN		0.73		0.73		0.54	0.06	0.08
PKD	-2.11	-2.06	-2.11	-2.06	-2.11	-1.98	-0.93	0.65
frail	1.66	1.78	1.66	1.78	1.66	1.65	1.08	1.00

A blank entry indicates that the variable was not selected. These results demonstrate that the Wald-type KOO method constructs the most parsimonious model, and the coefficients selected by this method are consistently included in the models chosen by other variable selection methods.

5.4. Logistic Regression Model

In this section, we applied a logistic regression model to the MNIST dataset, a standard dataset for handwritten digit recognition provided by LeCun et al. (1998), to distinguish between the digits "7" and "9." The MNIST dataset consists of 28×28 pixel grayscale images, with each pixel represented by an integer value ranging from 0 to 255. To avoid biases in the analysis, we focused only on variables where the median pixel value was between 100 and 150. The following results were obtained using the training data.

$n = 12214$	Wald	KOO		Step		Lasso		Wald
$k = 19$		AIC	BIC	AIC	BIC	$\lambda_{\min}$	$\lambda_{1\mathrm{se}}$	prob.
Intercept		-0.496	-0.518	-0.496	-0.518	-0.476	-0.280	0.00
X209	0.006	0.006	0.006	0.006	0.006	0.006	0.003	1.00
X214	0.005	0.004	0.004	0.004	0.004	0.004	0.003	1.00
X235	-0.007	-0.006	-0.006	-0.006	-0.006	-0.006	-0.003	1.00
X244	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.004	1.00
X262		-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.00
X297	-0.007	-0.009	-0.009	-0.009	-0.009	-0.008	-0.006	1.00
X318	0.006	0.004	0.004	0.004	0.004	0.004	0.003	1.00
X319		0.003	0.003	0.003	0.003	0.003	0.002	0.00
X353		0.003	0.003	0.003	0.003	0.003	0.001	0.07
X408	0.010	0.010	0.009	0.010	0.009	0.009	0.007	1.00
X439		0.002	0.002	0.002	0.002	0.002	0.001	0.00
X491		-0.001		-0.001		-0.001		0.00
X494		0.003	0.003	0.003	0.003	0.003	0.001	0.00
X519	-0.007	-0.005	-0.006	-0.005	-0.006	-0.005	-0.004	0.53
X549		-0.001		-0.001		0.000		0.00
X575		-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	0.00
X576		-0.001	-0.002	-0.001	-0.002	-0.001		0.00
X603								0.00

A blank entry indicates that the variable was not selected. The variable names such as X209 represent the pixel numbers in the 28×28 pixel image data. The pixel numbering starts at the top left corner of the image with X001 and increases row by row from left to right, with higher numbers corresponding to pixels closer to the bottom right corner. These results show that the Wald-type KOO method constructs the smallest model, and the coefficients selected by this method are consistently included in the models chosen by other selection methods. The variable compression rates are as follows:

$n = 12214$ $k = 19$	Wald	KOO AIC	KOO BIC	Step AIC	Step BIC	Lasso λ_{min}	Lasso λ_{1se}
Number of variables	8	18	16	18	16	19	16
Compression rate	42%	95%	84%	95%	84%	100%	84%

This shows that the Wald-type variable selection method achieves the highest compression rate. Next, we examine the accuracy of the logistic regression model in correctly distinguishing between "7" and "9" for each of the models selected by the different methods.

	All	Wald	KOO AIC	KOO BIC	Step AIC	Step BIC	Lasso λ_{min}	Lasso λ_{1se}
Training (n=12214)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Test (n=2037)	0.82	0.82	0.83	0.83	0.83	0.83	0.82	0.82

The "All" model uses all 19 variables, while "Training" refers to the data used to build the model, and "Test" refers to the remaining data. The results show that the accuracy of the model selected by the Wald-type method does not significantly differ from the accuracy of models selected by other methods. It is important to note that the accuracy is not 1 because the purpose of the proposed method is to select the true model with a probability of 1, not to achieve perfect classification accuracy.

6. Application: Extending the Method

In this section, we propose a method that can be applied to a broader range of contexts.

6.1. General Form of the Wald-type KOO Method

Here, we describe the Wald-type KOO method in a form that can be applied not only to generalized linear models but also to other settings.

Consider a more general framework. Assume that the sample data are generated from a distribution with parameters $\boldsymbol{\theta}$:

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}).$$

Here, $\boldsymbol{\theta}$ represents the parameters of interest in the analysis, similar to how μ and σ^2 characterize a normal distribution, or how regression coefficients β_i are the focus in multiple regression analysis.

Given $\boldsymbol{\theta}$ as a $k \times 1$ vector, we examine each parameter θ_i for $i = 1, \ldots, k$ to determine which of the following conditions it satisfies relative to a reference value θ_{0i}:

$$\theta_i < \theta_{0i}, \quad \theta_i > \theta_{0i}, \quad \theta_i = \theta_{0i}.$$

Each parameter θ_i must satisfy one of these conditions. For example, in multiple regression analysis, $\theta_i = \beta_i$ and $\theta_{0i} = 0$, so we assess whether each regression coefficient β_i is zero or not.

We categorize the parameters into three groups: those greater than θ_{0i} (Greater), those less than θ_{0i} (Less), and those equal to θ_{0i} (Equal). This can be expressed as:

$$\begin{aligned} M_* : \theta_{\ell_1} < \theta_{0\ell_1}, \ldots, \theta_{\ell_L} < \theta_{0\ell_L}, \theta_{g_1} > \theta_{0g_1}, \ldots, \theta_{g_G} > \theta_{0g_G}, \theta_{e_1} = \theta_{0e_1}, \ldots, \theta_{e_E} = \theta_{0e_E}, \\ L + G + E = k, \quad \{\ell_1, \ldots, \ell_L, g_1, \ldots, g_G, e_1, \ldots, e_E\} = \{1, \ldots, k\}. \end{aligned}$$

For consistency with terminology used in variable selection, we refer to M_* as the true model. To facilitate mathematical notation, we define the value J_i for each index i as follows:

$$\text{In } M_* : \quad \theta_i < \theta_{0i} \Rightarrow J_i = -1, \quad \theta_i > \theta_{0i} \Rightarrow J_i = 1, \quad \theta_i = \theta_{0i} \Rightarrow J_i = 0.$$

We then collect these into a vector $\boldsymbol{J} = (J_1, \ldots, J_k)'$. This vector $\boldsymbol{J}$ contains the state of each parameter in M_*. For example, in multiple regression analysis, $\boldsymbol{J}$ indicates whether each regression coefficient β_i is positive, negative, or zero when $\theta_i = \beta_i$ and $\theta_{0i} = 0$. Variable selection can thus be seen as identifying the non-zero elements of $\boldsymbol{J}$.

The Wald-type KOO method proposed in this paper makes individual decisions for each parameter based on a test statistic T_i. The decision is made using a threshold $d > 0$ as follows:

$$T_i < -d \Rightarrow \theta_i < \theta_{0i}, \quad T_i > d \Rightarrow \theta_i > \theta_{0i}, \quad -d \leq T_i \leq d \Rightarrow \theta_i = \theta_{0i}.$$

The results of these decisions are then summarized in $\hat{J}_i$ as follows:

$$T_i < -d \Rightarrow \hat{J}_i = -1, \quad T_i > d \Rightarrow \hat{J}_i = 1, \quad -d \leq T_i \leq d \Rightarrow \hat{J}_i = 0.$$

The overall decision vector is then $\hat{\boldsymbol{J}} = (\hat{J}_1, \ldots, \hat{J}_k)$. In this paper, we focus on the property that $\hat{\boldsymbol{J}}$ converges in probability to $\boldsymbol{J}$ under an asymptotic framework, which we refer to as consistency. For example, in multiple regression analysis, the test statistic T_i could be expressed as the difference between AIC values, $T_i = \mathrm{AIC}_{-i} - \mathrm{AIC}_{\mathrm{Full}}$, where AIC_{-i} is the AIC when excluding β_i, and the threshold d is set to 0.

6.2. Proposed Test Statistic

In this context, we consider the following test statistic for making decisions about each parameter:

$$T_i = \frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}}}, \quad i = 1, \ldots, k,$$

where $\hat{\theta}_i$ is the maximum likelihood estimator of θ_i. The statistic T_i is known as the Wald-type statistic. The threshold d is defined as:

$$d = n^{1/4}.$$

We impose the following assumptions on the sample size n, the number of parameters k, the threshold d, and the estimator $\hat{\theta}_i$:

A1' $n \to \infty, \quad d = n^{1/4}, \quad k = O(n^a), \quad 0 \leq a < \frac{1}{2}$,

A2' Consistency of the estimator (1): Under A1'

$\hat{\theta}_i = \theta_i + O_p(n^{-1/2})$,

A3' Consistency of the estimator (2): Under A1'

For any $\varepsilon > 0$, $P(|\hat{\theta}_i - \theta_i| \geq \varepsilon) = O(n^{-1/2})$,

A4' Variance of the estimator: Under A1'

$\mathrm{Var}(\hat{\theta}_i) = O(n^{-1})$,

A5' Wald-type probability evaluation: Under A1'

$$P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}}} \geq x\right) = P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\mathrm{Var}(\hat{\theta}_i)}} \geq x\right) + O(n^{-1/2}).$$

Under these assumptions, the following result holds:

Theorem 6.1. *If assumptions A0' through A5' are satisfied, then:*

$$\lim_{n \to \infty} P(\hat{\boldsymbol{J}} = \boldsymbol{J}) = 1$$

Note that Theorem 3.1 corresponds to the case where $\theta_{0i} = 0$.

6.3. Estimation of Standard Errors Using Bootstrap

In the method proposed in this paper, the state of each parameter is determined using T_i, which requires the prior estimation of the standard error of the estimator $\hat{\theta}_i$, denoted as $\sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}}$. In cases where the standard error is not provided, we propose a method to construct T_i by estimating the standard error using the Bootstrap method.

Let B denote the number of bootstrap samples, and let $\hat{\theta}_{ij}^{\#}$ represent the estimator obtained from the j-th bootstrap sample. The standard error can be estimated as follows:

$$\hat{\sigma}_i^{\#} = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\hat{\theta}_{ij}^{\#} - \hat{\theta}_i)^2}, \quad i = 1, \ldots, k.$$

Using this bootstrap-estimated standard error, the test statistic T_i is then constructed as:

$$T_i = \frac{\hat{\theta}_i - \theta_{0i}}{\hat{\sigma}_i^{\#}}.$$

We assume the following relationship between the standard error and its bootstrap estimate:

$$\text{C1} \quad \text{Under A1'}: \quad \hat{\sigma}_i^{\#} = \sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}} + O_p(n^{-1/2}).$$

Under this assumption, the following theorem holds:

Theorem 6.2. *Under assumptions A0' through A5' and C1, the test statistic T_i constructed using the bootstrap method satisfies:*

$$\lim_{n\to\infty} P(\hat{\boldsymbol{J}} = \boldsymbol{J}) = 1.$$

6.4. Numerical Experiment

In this section, we conducted numerical experiments using the generalized Wald-type KOO method. The experiment focuses on the structure of means across three groups. The samples from the three groups are given as follows:

$$\begin{aligned}
&\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1} \overset{\text{i.i.d.}}{\sim} G : \mathrm{E}[\boldsymbol{x}_1] = \boldsymbol{\mu}_1, \mathrm{Var}(\boldsymbol{x}_1) = \Sigma_1,\\
&\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2} \overset{\text{i.i.d.}}{\sim} G : \mathrm{E}[\boldsymbol{x}_2] = \boldsymbol{\mu}_2, \mathrm{Var}(\boldsymbol{x}_2) = \Sigma_2,\\
&\boldsymbol{x}_{31}, \ldots, \boldsymbol{x}_{3n_3} \overset{\text{i.i.d.}}{\sim} G : \mathrm{E}[\boldsymbol{x}_3] = \boldsymbol{\mu}_3, \mathrm{Var}(\boldsymbol{x}_3) = \Sigma_3,\\
&\text{where } \boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}, \boldsymbol{x}_{3j} \text{ are mutually independent.}
\end{aligned}$$

Here, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$. The goal is to determine the extent to which the components μ_{ij} of the mean vectors of each group differ. The difference between the means of groups j_1 and j_2 in the i-th component is defined and assessed as follows:

$$
\begin{aligned}
&T_{ij_1j_2} = \frac{\bar{x}_{ij_1} - \bar{x}_{ij_2}}{\hat{\sigma}^{\#}_{ij_1j_2}}, \quad i = 1, \ldots, k, \\
&T_{ij_1j_2} > d_{j_1j_2} \Rightarrow \mu_{ij_1} - \mu_{ij_2} > 0, \\
&T_{ij_1j_2} < -d_{j_1j_2} \Rightarrow \mu_{ij_1} - \mu_{ij_2} < 0, \\
&-d_{j_1j_2} \le T_{ij_1j_2} \le d_{j_1j_2} \Rightarrow \mu_{ij_1} - \mu_{ij_2} = 0, \\
&d_{j_1j_2} = n^{1/4}_{j_1j_2}, \quad n_{j_1j_2} = \min(n_{j_1}, n_{j_2}), \\
&\hat{\sigma}^{\#}_{j_1j_2} = \sqrt{\frac{1}{B}\sum_{j=1}^{B}\Big((\bar{x}^{\#}_{ij_1} - \bar{x}^{\#}_{ij_2}) - (\bar{x}_{ij_1} - \bar{x}_{ij_2})\Big)^2}.
\end{aligned}
$$

Here, $\bar{x}^{\#}_{ij}$ represents the mean from the resampled data.

The simulation was conducted under the following settings:

$$
\begin{aligned}
&\boldsymbol{\mu}_1 = (\mathbf{1}_{p_1}, \mathbf{1}_{p_2}, -\mathbf{1}_{p_3}, \mathbf{0}_{p_4}, \mathbf{0}_{p_5})', \\
&\boldsymbol{\mu}_2 = (\mathbf{0}_{p_1}, \mathbf{0}_{p_2}, \mathbf{0}_{p_3}, \mathbf{0}_{p_4}, \mathbf{0}_{p_5})', \\
&\boldsymbol{\mu}_3 = (-\mathbf{1}_{p_1}, \mathbf{0}_{p_2}, \mathbf{1}_{p_3}, \mathbf{1}_{p_4}, \mathbf{0}_{p_5})', \\
&\quad p = p_1 + p_2 + p_3 + p_4 + p_5, \quad p_1 = p_2 = p_3 = p_4 = p_5,
\end{aligned}
$$

Σ_i : Diagonal elements are generated from a uniform distribution on (1, 2),
and off-diagonal elements from a uniform distribution on (0, 1).

$\boldsymbol{x}_{ij} = \boldsymbol{\mu}_i + \Sigma_i^{1/2}\boldsymbol{z}_{ij}, \quad i = 1, \ldots, n, \quad j = 1, 2, 3,$
where the components of $\boldsymbol{z}_{ij}$ are independently and identically distributed.

The distribution of $\boldsymbol{z}_{ij}$ was considered under the following cases:

Standard normal distribution: $z_{ij} \sim N(0, 1)$,

Standardized uniform distribution: $z_{ij} = \dfrac{u_{ij} - 1/2}{\sqrt{1/12}}, \quad u_{ij} \sim U(0, 1)$,

Standardized binomial distribution: $z_{ij} = \dfrac{x_{ij} - 1/2}{\sqrt{1/4}}, \quad x_{ij} \sim \mathrm{Bin}(1, 1/2)$,

Standardized Poisson distribution: $z_{ij} = x_{ij} - 1, \quad x_{ij} \sim \mathrm{Pois}(1)$.

The mean structures for each dimension are represented as follows:

$$p_1 : \mu_1 > \mu_2 > \mu_3,$$
$$p_2 : \mu_1 > \mu_2 = \mu_3,$$
$$p_3 : \mu_1 < \mu_2 < \mu_3,$$
$$p_4 : \mu_1 = \mu_2 < \mu_3,$$
$$p_5 : \mu_1 = \mu_2 = \mu_3.$$

The results for the true model selection probabilities across each dimension and overall are presented below.

Distribution	n	p	p_1	p_2	p_3	p_4	p_5	all
Normal	10	5	0.95	0.85	0.95	0.86	0.76	0.53
	50	5	1.00	0.99	1.00	0.99	0.97	0.96
	100	100	1.00	0.96	1.00	0.96	0.89	0.82
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
Uniform	10	5	0.96	0.86	0.96	0.85	0.75	0.54
	50	5	1.00	0.99	1.00	0.99	0.97	0.95
	100	100	1.00	0.96	1.00	0.96	0.90	0.83
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
Binomial	10	5	0.96	0.86	0.96	0.86	0.76	0.55
	50	5	1.00	0.99	1.00	0.99	0.97	0.95
	100	100	1.00	0.96	1.00	0.96	0.89	0.83
	500	100	1.00	1.00	1.00	1.00	1.00	1.00
Poisson	10	5	0.94	0.85	0.95	0.85	0.75	0.51
	50	5	1.00	0.99	1.00	0.99	0.97	0.95
	100	100	1.00	0.96	1.00	0.97	0.90	0.83
	500	100	1.00	1.00	1.00	1.00	1.00	1.00

Here, the selection probability for p_i represents the probability of correctly identifying the true structure in all dimensions corresponding to p_i, and "all" indicates the probability of correctly identifying the true structure across all dimensions.

These results suggest that, given a sufficiently large sample size, the probability of selecting the true structure is high, regardless of the distribution. The differences in distribution did not have a significant impact on the selection probabilities in this study.

6.5. Application to Real Data

In this section, we analyze data from three species of penguins, as presented in Gorman et al. (2014). This dataset, known as the Penguin Dataset, is available on Kaggle and is also incorporated into the Python data visualization library seaborn. After removing entries with missing values, we analyzed $n = 333$ samples (Adelie: $n_1 = 146$, Chinstrap: $n_2 = 68$, Gentoo: $n_3 = 119$) to examine the mean structure of the following measurements:

$$\begin{array}{rl} \text{bill length} & : \text{Length of the penguin's bill (mm)} \\ \text{bill depth} & : \text{Depth of the penguin's bill (mm)} \\ \text{flipper length} & : \text{Length of the penguin's flipper (mm)} \\ \text{body mass} & : \text{Body mass of the penguin (g)} \end{array}$$

Using the test statistics as conducted in the numerical experiments, the mean structure was analyzed, yielding the following results:

	Adelie	Chinstrap	Gentoo	Selected Mean Structure
bill length	38.8	48.8	47.6	Adelie<Chinstrap = Gentoo
bill depth	18.3	18.4	15.0	Adelie = Chinstrap>Gentoo
flipper length	190.1	195.8	217.2	Adelie = Chinstrap<Gentoo
body mass	3706.2	3733.1	5092.4	Adelie = Chinstrap<Gentoo

7. Concluding Remarks

In this paper, we proposed a simple and unified method that possesses consistency in variable selection. Through applications to regression models, logistic regression models, Poisson regression models, and Cox proportional hazards models, we found that the proposed method exhibits consistency in variable selection, provided that the sample size is sufficiently large. However, for logistic regression models, the convergence is slower, necessitating a larger sample size compared to other models. Additionally, we presented a method for estimating the selection probability, which is asymptotically median unbiased. Similar to the consistency results, the logistic regression model showed slower convergence, requiring more samples than the other models. We also demonstrated the applicability of this method to more complex model selection scenarios.

For future work, it should be noted that the results presented here were obtained under conditions where the number of variables was not excessively large relative to the sample size. However, the numerical results suggest that the theoretical conditions provided could be relaxed under certain circumstances. Further refinement and evaluation may allow for a relaxation of these conditions. Moreover, given the slower convergence observed in the logistic regression model, exploring faster methods for achieving convergence is another potential direction for future research.

Appendix: Proof of Consistency and Selection Probability

A1 Proof of Consistency

In this section, we provide a proof of the consistency of the proposed test statistic. We specifically extend Theorem 3.1 to Theorem 6.1. First, we prove the consistency of the test statistic T_i without using bootstrap standard errors. Consistency of T_i with respect to the threshold d can be rewritten as follows:

$$\begin{aligned}
&\lim_{n\to\infty} P(\hat{\boldsymbol{J}} = \boldsymbol{J}) = 1 \\
\Leftrightarrow &\lim_{n\to\infty} P\left(\left\{\bigcap_{i=1}^{L} T_{\ell_i} < -d\right\} \cap \left\{\bigcap_{i=1}^{G} T_{g_i} > d\right\} \cap \left\{\bigcap_{i=1}^{E} -d \leq T_i \leq d\right\}\right) = 1 \\
\Leftrightarrow &\lim_{n\to\infty} P\left(\left\{\bigcap_{i=1}^{L} T_{\ell_i} < -d\right\} \cap \left\{\bigcap_{i=1}^{G} T_{g_i} > d\right\} \cap \left\{\bigcap_{i=1}^{E} |T_i| \leq d\right\}\right) = 1 \\
\Leftrightarrow &\lim_{n\to\infty} P\left(\left\{\bigcup_{i=1}^{L} T_{\ell_i} \geq -d\right\} \cup \left\{\bigcup_{i=1}^{G} T_{g_i} \leq d\right\} \cup \left\{\bigcup_{i=1}^{E} |T_{e_i}| > d\right\}\right) = 0
\end{aligned}$$

To establish this, it suffices to show the following:

$$\lim_{n\to\infty} \sum_{i=1}^{L} P\left(T_{\ell_i} \geq -d\right) + \lim_{n\to\infty} \sum_{i=1}^{G} P\left(T_{g_i} \leq d\right) + \lim_{n\to\infty} \sum_{i=1}^{E} P\left(|T_{e_i}| > d\right) = 0$$

This is derived from the upper bound given by the complement probability and the sum of disjoint events.

First, we consider the evaluation of $P(T_{\ell_i} \geq -d)$. To simplify the notation, we denote the index i as $i \in \{\ell_i\}$. Note that in the true model, $\theta_i < \theta_{0i}$. We have:

$$\begin{aligned}
P\left(T_i \geq -d\right) &= P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}}} \geq -d\right) \\
&= P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\mathrm{Var}(\hat{\theta}_i)}} \geq -d\right) + O(n^{-1/2}) \\
&= P\left(\hat{\theta}_i - \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2})
\end{aligned}$$

Next, we rearrange the inequality inside the probability as follows:

$$\begin{aligned}
&\hat{\theta}_i - \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d \\
\Leftrightarrow &\hat{\theta}_i - \theta_i \geq \theta_{0i} - \theta_i - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d \\
= &\theta_{0i} - \theta_i - O(n^{-1/2})n^{1/4} = \theta_{0i} - \theta_i - O(n^{-1/4})
\end{aligned}$$

Since $\theta_i < \theta_{0i}$, i.e., $\theta_{0i} - \theta_i > 0$, we can choose a sufficiently large n such that:

$$\theta_{0i} - \theta_i - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d > 0$$

In this context, it is well-known that for a random variable X and a constant $a > 0$, the following inequality holds:

$$P(X \geq a) \leq P(X \leq -a \cup X \geq a) = P(|X| \geq a)$$

Thus, for sufficiently large n, we can evaluate $P(T_i \geq -d)$ as follows:

$$\begin{aligned}
P\left(T_i \geq -d\right) &= P\left(\hat{\theta}_i - \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&= P\left(\hat{\theta}_i - \theta_i \geq \theta_{0i} - \theta_i - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&\leq P\left(|\hat{\theta}_i - \theta_i| \geq \theta_{0i} - \theta_i - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&= O(n^{-1/2}) + O(n^{-1/2}) = O(n^{-1/2})
\end{aligned}$$

From this, we have:

$$\sum_{i=1}^{L} P\left(T_{\ell_i} \geq -d\right) \leq \sum_{i=1}^{L} O(n^{-1/2}) \leq k \times O(n^{-1/2}) = O(n^{a}) \times O(n^{-1/2}) = O(n^{a-1/2})$$

where a satisfies $0 \leq a < 1/2$, implying that $a - 1/2 < 0$, so this term converges to zero as n increases.

Next, consider the evaluation of $P(T_{g_i} \leq d)$. This can be shown in a similar manner as before. To simplify the notation, let $i \in \{g_i\}$ denote the index i. Note that in the true model, $\theta_i > \theta_{0i}$. Therefore,

$$\begin{aligned}
P\left(T_i \leq d\right) &= P\left(-\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}}} \geq -d\right) \\
&= P\left(-\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\mathrm{Var}(\hat{\theta}_i)}} \geq -d\right) + O(n^{-1/2}) \\
&= P\left(-\hat{\theta}_i + \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}).
\end{aligned}$$

The inequality within the probability can be rewritten as follows:

$$
\begin{aligned}
-\hat{\theta}_i + \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d \Leftrightarrow \theta_i - \hat{\theta}_i \geq \theta_i - \theta_{0i} - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d \\
= \theta_i - \theta_{0i} - O(n^{-1/2})n^{1/4} = \theta_i - \theta_{0i} - O(n^{-1/4}).
\end{aligned}
$$

Since $\theta_i > \theta_{0i}$, implying $\theta_i - \theta_{0i} > 0$, we can choose a sufficiently large n such that

$$\theta_i - \theta_{0i} - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d > 0.$$

Therefore, for sufficiently large n:

$$
\begin{aligned}
P\left(T_i \leq d\right) &= P\left(-\hat{\theta}_i + \theta_{0i} \geq -\sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&= P\left(\theta_i - \hat{\theta}_i \geq \theta_i - \theta_{0i} - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&\leq P\left(|\theta_i - \hat{\theta}_i| \geq \theta_i - \theta_{0i} - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&= P\left(|\hat{\theta}_i - \theta_i| \geq \theta_i - \theta_{0i} - \sqrt{\mathrm{Var}(\hat{\theta}_i)}d\right) + O(n^{-1/2}) \\
&= O(n^{-1/2}) + O(n^{-1/2}) = O(n^{-1/2}).
\end{aligned}
$$

From this, we have:

$$\sum_{i=1}^{G} P\left(T_{g_i} \leq d\right) \leq \sum_{i=1}^{G} O(n^{-1/2}) \leq k \times O(n^{-1/2}) = O(n^a) \times O(n^{-1/2}) = O(n^{a-1/2}),$$

where a satisfies $0 \leq a < 1/2$, implying that $a - 1/2 < 0$, so this term converges to zero as n increases.

Finally, consider the evaluation of $P(|T_{e_i}| > d)$. Again, to simplify the notation, let $i \in \{e_i\}$ denote the index i. By Chebyshev's inequality:

$$P\left(|T_i| > d\right) \leq \frac{1}{d^2},$$

which implies:

$$\sum_{i=1}^{E} P\left(|T_{e_i}| > d\right) \leq \frac{k}{d^2} = O(n^{a-1/2}),$$

where $k = n^a$ and $0 < a < 1/2$, so $a - 1/2 < 0$, and this term also converges to zero.

In summary:

$$\begin{aligned}0 &\leq \sum_{i=1}^{L} P\left(T_{\ell_i} \geq -d\right) + \sum_{i=1}^{G} P\left(T_{g_i} \leq d\right) + \sum_{i=1}^{E} P\left(|T_{e_i}| > d\right) \\ &\leq O(n^{a-1/2}) + O(n^{a-1/2}) + O(n^{a-1/2}) = O(n^{a-1/2}),\end{aligned}$$

which implies:

$$\lim_{n\to\infty} \sum_{i=1}^{L} P\left(T_{\ell_i} \geq -d\right) + \lim_{n\to\infty} \sum_{i=1}^{G} P\left(T_{g_i} \leq d\right) + \lim_{n\to\infty} \sum_{i=1}^{E} P\left(|T_{e_i}| > d\right) = 0.$$

Thus, we have shown that $P(\hat{\boldsymbol{J}} = \boldsymbol{J}) \to 1$.

Next, we consider the proof of Theorem 6.2. In proving the consistency when using bootstrap methods, the relationship between the standard error and its bootstrap estimate under assumption C1 is given by

$$\hat{\sigma}_i^{\#} = \sqrt{\widehat{\mathrm{Var}(\hat{\theta}_i)}} + O_p(n^{-1/2}).$$

From this, the following holds:

$$P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\hat{\sigma}_i^{\#}} \geq x\right) = P\left(\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{\mathrm{Var}(\hat{\theta}_i)}} \geq x\right) + O(n^{-1/2}).$$

Thus, the proof proceeds in the same manner as the previous consistency proof.

A2 Estimation of Selection Probability

In this section, for simplicity, we denote $\beta = \beta_i$ and $d = n^{1/4}$.

First, consider the case where $\beta > 0$. The selection probability p can be expressed as follows:

$$
\begin{aligned}
p &= P\left(\left|\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}}\right| > d\right) \\
&= P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} > d\right) + P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} < -d\right).
\end{aligned}
$$

Focusing on the second term, we can evaluate it using assumption A3 as follows:

$$
\begin{aligned}
P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} < -d\right) &= P\left(\hat{\beta} - \beta < -\beta - d\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}\right) \\
&\leq P\left(|\hat{\beta} - \beta| > \beta + d\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}\right) = O(n^{-1/2}).
\end{aligned}
$$

From this, we have:

$$
0 \leq P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} < -d\right) \leq O(n^{-1/2}),
$$

which implies:

$$
P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} < -d\right) = O(n^{-1/2}).
$$

Therefore, the selection probability p is given by:

$$
p = P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} > d\right) + O(n^{-1/2}).
$$

Assumption B1 assumes asymptotic normality of the estimator $\hat{\beta}$, so we have:

$$\begin{aligned} P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} > d\right) &= P\left(\frac{\hat{\beta}-\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} > d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) \\ &= 1 - \Phi\left(d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) + o(1), \end{aligned}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution, given by $\Phi(x) = P(Z \leq x)$, $Z \sim N(0,1)$. Thus, the selection probability p is:

$$p = 1 - \Phi\left(d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) + o(1).$$

Next, consider the estimation of the selection probability $\hat{p}$. Given a sample $\boldsymbol{X} = (X_1, \ldots, X_n)'$, the estimator $\hat{\beta} = \hat{\beta}(\boldsymbol{X})$ is obtained. The distribution of the estimator $\hat{\beta}^{\#} = \hat{\beta}(\boldsymbol{X}^{\#})$ based on the bootstrap sample $\boldsymbol{X}^{\#} = (X_1^{\#}, \ldots, X_n^{\#})'$ is given by assumption B2 as:

$$P\left(\frac{\hat{\beta}-\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < x\right) = P\left(\frac{\hat{\beta}^{\#}-\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < x \,\middle|\, \boldsymbol{X}\right) + O(n^{-1}).$$

Since the distribution of $\hat{\beta}^{\#}$ can be computed using the empirical distribution function, we have:

$$P\left(\frac{\hat{\beta}^{\#}-\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < x \,\middle|\, \boldsymbol{X}\right) = \frac{1}{B}\sum_{j=1}^{B} I\left(\frac{\hat{\beta}_j^{\#}-\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < x\right) + O(B^{-1}).$$

Therefore, the estimated selection probability $\hat{p}$ is:

$$\begin{aligned}
\hat{p} &= \frac{1}{B}\sum_{j=1}^{B} I\left(\left|\frac{\hat{\beta}_j^{\#}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right| > d\right) \\
&= \frac{1}{B}\sum_{j=1}^{B} I\left(\frac{\hat{\beta}_j^{\#}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} > d\right) + \frac{1}{B}\sum_{j=1}^{B} I\left(\frac{\hat{\beta}_j^{\#}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < -d\right) \\
&= P\left(\frac{\hat{\beta}^{\#} - \hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} > d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \;\middle|\; \boldsymbol{X}\right) + P\left(\frac{\hat{\beta}^{\#} - \hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < -d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \;\middle|\; \boldsymbol{X}\right) + O(B^{-1}).
\end{aligned}$$

The first term, using assumptions B1 and B2, evaluates as:

$$P\left(\frac{\hat{\beta}^{\#} - \hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} > d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \;\middle|\; \boldsymbol{X}\right) = 1 - \Phi\left(d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) + o(1).$$

The second term is evaluated as follows:

$$\begin{aligned}
P\left(\frac{\hat{\beta}^{\#} - \hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < -d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \;\middle|\; \boldsymbol{X}\right) &= P\left(\frac{\hat{\beta} - \beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < -d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) + O(n^{-1}) \\
&= P\left(\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} < -d\right) + O(n^{-1}) \\
&= O(n^{-1/2}).
\end{aligned}$$

Thus, the estimated selection probability $\hat{p}$ is:

$$\hat{p} = 1 - \Phi\left(d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) + o(1).$$

Noting that $\Phi(x)$ is a monotonically increasing function with an inverse function $\Phi^{-1}(x)$, we have:

$$
\begin{aligned}
P(\hat{p} \le p) &= P\left\{1 - \Phi\left(d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) \le 1 - \Phi\left(d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right)\right\} + o(1) \\
&= P\left\{\Phi\left(d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right) \ge \Phi\left(d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right)\right\} + o(1) \\
&= P\left\{d - \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \ge d - \frac{\beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}\right\} + o(1) \\
&= P\left(\frac{\hat{\beta} - \beta}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}} \le 0\right) + o(1) \\
&= \frac{1}{2} + o(1),
\end{aligned}
$$

where the last equality follows from the asymptotic normality assumption in B1, and assuming that the $O(\cdot)$ term can be taken outside the probability expression. Thus,

$$P(\hat{p} \ge p) = 1 - P(\hat{p} \le p) = \frac{1}{2} + o(1).$$

This shows that the median of the estimated selection probability $\hat{p}$ is a good approximation of the true selection probability p when $\beta > 0$. The case for $\beta < 0$ can be shown similarly.

References

[1] Bai, Z., Choi, K. P., Fujikoshi, Y. and Hu, J. (2024). KOO approach for scalable variable selection problem in large-dimensional regression. To appear in Statistica Sinica.

[2] Dobson, A.J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models.* CRC Press.

[3] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-typ criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184–200.

[4] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, **33**, 1–22.

[5] Gorman, K. B., Williams, T. D. and Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081

[6] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

[7] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition, in Proceedings of the IEEE, **86** (11), 2278–2324

[8] Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications. http://investigadores.cide.edu/aparicio/data/refs/Long

[9] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Mathematical Journal*, *18*, 451–462.

[10] Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York. https://cran.r-project.org/web/packages/survival/survival.pdf

[11] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth Edition, Springer, New York. https://cran.r-project.org/web/packages/MASS/MASS.pdf

[12] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, **9**, 869–897.

[13] Zhao, L. C. , Krishnaiah, P. R. and Bai, Z. D. (1986). On determination of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.