Dissertation for the degree of Doctor of Philosophy

A study on performance improvement through machine learning method when analyzing survival data including censoring

Seong-Yun Hwang

Department of Statistics
The Graduate School
Jeonbuk National University

August, 2024

Dissertation for the degree of Doctor of Philosophy

A study on performance improvement through machine learning method when analyzing survival data including censoring

Seong-Yun Hwang

Department of Statistics
The Graduate School
Jeonbuk National University

August, 2024

Dissertation for the degree of Doctor of Philosophy

A study on performance improvement through machine learning method when analyzing survival data including censoring

Under the direction of **Seong-Jun Yang**

Seong-Yun Hwang

Department of Statistics
The Graduate School
Jeonbuk National University

August, 2024

A Dissertation submitted to the Graduate school of Jeonbuk National University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics under the direction of **Seong-Jun Yang**.

The dissertation for the degree of Doctor of Philosophy by

Seong-Yun Hwang

has been approved by the committee members.

15th, June, 2024

Chair: _____

Vice Chair: _____

Member: _____

Member: _____

Member: _____

Contents

1 Introduction **1**

1.1 Background and purpose of the study 1

1.2 Research method and composition 2

2 survival data **5**

2.1 survival data 5

2.2 survival analysis 6

2.3 Synthetic response for censored data analysis 11

3 Kernel ridge censored regression analysis **15**

3.1 Multiple regression analysis 16

3.2 Ridge regression analysis 20

3.3 Kernel ridge regression analysis 23

3.4 Kernel ridge censored regression analysis 32

4 Ensemble method **34**

4.1 Bagging 36

4.2 Random forests 39

5 Kernel ridge censored regression analysis using ensemble method 41

5.1 Kernel ridge censored regression analysis using Bagging technique . 42

5.2 Kernel ridge censored regression analysis using Random Forests technique 47

6 Simulation and real data analysis (Kernel ridge censored regression analysis using ensemble method) 49

6.1 Simulation 50

6.2 Real data analysis 74

7 Time-dependent AUC **83**

7.1 ROC curve 83

7.2 Time-dependent AUC 86

8 Cox regression model **89**

8.1 Cox proportional hazard model 89

8.2 Extended Cox regression model 93

9 Calibration **94**

9.1 Calibration 94

9.2 Concordance Index 96

10 **99**

10.1 99

11 **100**

11.1 100

12 **101**

12.1 101

13 **102**

13.1 102

14 **103**

14.1 103

15 **104**

15.1 104

List of Figures

2.1 Survival function of lung cancer data 7

3.1 Kernel-trick method 24

4.1 Bootstrap 37

4.2 Overfitting and Underfitting 38

6.1 $p = 3$, censoring 0% 57

6.2 $p = 3$, censoring 10% 58

6.3 $p = 3$, censoring 30% 59

6.4 $p = 3$, censoring 50% 60

6.5 $p = 5$, censoring 0% 61

6.6 $p = 5$, censoring 10% 62

6.7 $p = 5$, censoring 30% 63

6.8 $p = 5$, censoring 50% 64

6.9 $p = 7$, censoring 0% 65

6.10 $p = 7$, censoring 10% 66

6.11 $p = 7$, censoring 30% 67

6.12 $p = 7$, censoring 50% 68

6.13 $p = 9$, censoring 0% 69

6.14 $p = 9$, censoring 10% 70

6.15 $p = 9$, censoring 30% 71

6.16 $p = 9$, censoring 50% 72

6.17 Real data analysis : UIS data 75

6.18 Real data analysis : PBC data 76

6.19 Real data analysis : Cancer data 77

6.20 Real data analysis : Cancer data 78

6.21 Real data analysis : Bfeed data 79

7.1 ROC curve 85

9.1 Indicator function VS Log-sigmoid lower bound 97

List of Tables

2.1 Representative example of survival analysis method 9

3.1 Various kernel types that satisfy Mercer's theorem 27

6.1 Summary of real data analysis results 81

7.1 Confusion matrix 83

ABSTRACT

A study on performance improvement through machine learning method when analyzing survival data including censoring

SEONG YUN HWANG

DEPARTMENT OF STATISTICS

THE GRADUATE SCHOOL

JEONBUK NATIONAL UNIVERSITY

This paper contains the contents of a total of two studies, all of which are related to the analysis of survival data.

The first study relates to a method that can improve predictive power when regression analysis is performed on data with censoring. Censorship is usually caused by internal or external causes, such as the death of a patient due to a factor other than the disease being studied in the survival data related to the patient's survival time, which appears frequently in the medical field. will do The main purpose of analyzing survival data is to determine which factors have a significant effect on the patient's survival time and to predict the patient's survival

time through this. In the case of survival data including such censoring, since the survival time, which is the subject of estimation, is only partially observed, a synthetic response can be created to analyze the data. However, these artificial variables have a characteristic that the conditional variance when an explanatory variable is given tends to be larger than the conditional variance of the original survival time, and the width increases as the survival time increases. Because of this, the stability of the estimator is poor, which can be a problem. To compensate for this problem, in this study, when constructing a regression model for artificial variables, the data in the explanatory variable space is moved to a high-dimensional characteristic space by using an appropriate mapping function for complex nonlinear data without specifying a transformation function. When there is a problem of the kernel trick method and multicollinearity, the applicable ridge regression method is applied. In addition, we would like to propose a method for improving the predictive power of survival time by reducing the variance of the estimator by applying ensemble methods such as bagging and random forest. Through computer simulation, various situations were assumed and the predictive power of explanatory variables was compared and analyzed in data including censoring. Through this, it was confirmed that the method proposed in this study showed overall superior predictive power compared to the general method.

The second study relates to a method that can improve the overall time-dependent AUC that can be calculated when analyzing survival data. Continue to explain...

keywords : survival data, synthetic response, ridge regression, machine

learning, kernel trick method, ensemble method, time-dependent AUC, Cox proportional hazard model

1 Introduction

1.1 Background and purpose of the study

Survival data is data mainly used in the medical field to check and study the survival time of patients. Recently, in addition to the medical field, it is widely used in various research fields such as analyzing the survival rate of companies or the unemployment rate of workers in the economy and management fields, and its use is also increasing. In particular, one of the greatest characteristics of survival data is that censoring is included. Censoring is caused by various causes, such as when a patient dies due to a cause other than the disease being studied, when a hospitalized patient is transferred to another hospital, when the researcher arbitrarily adjusts the observation time, and when the ongoing research is stopped. Therefore, in the survival data, the actual patient's survival time to be estimated is only partially observed when censoring has not occurred. Based on these characteristics of survival data, various analysis methods such as Kaplan-Meier estimation, Nelson-Aalen estimation, Cox proportional hazard model, Parametric model, and Accelerated failure time model have been proposed. For details on the various survival analysis methods, see Kleinbaum, D.G. and Klein, M. (2010).

However, in the recent big-data era, the data we actually encounter has a

nonlinear relationship between a response variable and explanatory variables, or in most cases, standardization is not done. Therefore, there are many cases in which the researcher has to process the data in advance so that the data has an appropriate form for analysis in consideration of various methods such as variable transformation. In addition, the problem of multicollinearity caused by the association between explanatory variables included in the data is one of the important issues to be overcome. This problem is no exception in survival data dealt with in the medical field. Therefore, in this paper, as the first research result, when analyzing survival data including censoring, we propose a method that can overcome these problems to some extent.

두 번째 연구결과 설명...

1.2 Research method and composition

In this paper, two major studies have been conducted, and all of them contain information on how to obtain more improved results by applying methods such as machine learning to the analysis of survival data.

First, the contents of the first study will be explained as a whole. If you want to predict the survival time from survival data with censorship, you can use a synthetic response defined to replace the actual patient's survival time, which is only partially observed, when building a model. For the data conversion method using this synthetic response, see Buckley, J. and James, I. (1979), Koul et al. (1981), Leurgans, S. (1987), et al. And when regression analysis is performed

with this synthetic response as a response variable, kernel ridge censored regression method, which is combined the advantages of the kernel trick method and ridge regression method, is applied. Kernel trick method is widely mentioned in the field of machine learning, which is a hot topic recently, as a method that can automatically perform an appropriate transformation without applying an appropriate transformation function to the data to be analyzed in advance. In particular, this method is widely used in various algorithms such as support vector machine (SVM), which is frequently used in classification problems. And the ridge regression method is a kind of penalized regression method that can be used to supplement the multicollinearity problem. Plus, an ensemble method that can significantly reduce the variance of the estimator is additionally applied to build a more stable and reliable prediction model. In this study, bagging that can reduce variance and reduce the risk of overfitting by averaging the results from multiple bootstrap samples selected through iterative sampling with replacement using raw data and a random forest that enables more accurate estimation by reducing the association between explanatory variables through the process of selecting only some of the explanatory variables without including all explanatory variables during bootstrap sampling are applied. A method of increasing the predictive power for survival time through this process is proposed as the first research result.

Next, the contents of the second study will be introduced as a whole. An important key of the second study is the calibration of the time-dependent AUC (Area Under Curve). In other words, a method was studied to increase the overall value of time-dependent AUC, which can be calculated when analyzing time-

dependent survival data and is an index to evaluate the accuracy of the analysis result. 두 번째 연구결과 요약 설명 계속...

This dissertation consists of a total of 13 chapters. Chapter 1 outlines the background of the research presented in this paper and the research method accordingly. Chapters 2 to 6 describe the case of applying the kernel trick method and ensemble method in regression analysis on survival data including censoring, which can improve predictive power to predict survival time. Chapter 2 briefly describes the characteristics of survival data and the commonly used survival analysis method when analyzing it, and then mentions how a synthetic response for censored data analysis can be defined. Chapter 3 describes kernel ridge censored regression analysis, which is the core of the first study, and chapter 4 introduces the ensemble method. Next, Chapter 5 explains how to apply the ensemble method to kernel ridge censored regression analysis. And in Chapter 6, we present and evaluate the results of simulation and real data analysis conducted to prove that the kernel ridge censored regression analysis using ensemble method has overall better predictive power than the normal method. Next, Chapters 7 to 12 describe the study of the calibration of time-dependent AUC with Cox regression model. 두번째 연구 종료 후 추가... Finally, Chapter 13 summarizes the research method presented in this paper as a whole and concludes the dissertation.

2 survival data

2.1 survival data

Survival data are often used in the medical field to analyze and study the survival time of patients. As mentioned in the introduction, the biggest characteristic of this data is censoring that it has been caused by internal or external factors, such as a patient's death due to a cause other than the disease being studied or the researcher's arbitrary observation time adjustment. These data include explanatory variables $(x_1, x_2, \ldots, x_p)$, observed survival time $t = \min(y, c)$, and $\delta = I(y \leq c)$, which is an indicator variable indicating whether censoring or not, is included by default. Here, y is the time until the event of interest occurs (usually the patient's actual survival time in the medical field), and c is the censoring time. That is, if the patient's actual survival time is observed by satisfying $y \leq c$, $t = y$ and $\delta = 1$. On the contrary, when censoring is performed by satisfying $y > c$, $t = c$ and $\delta = 0$. In other words, in the actual survival data, the actual survival time of all patients cannot be confirmed, and only in the case of $y \leq c$, the survival time of the actual patients can be partially observed. Therefore, a special method for analyzing survival data is required, and it is collectively called a survival analysis method.

2.2 survival analysis

In survival analysis, the following functions are basically defined based on the observed survival time t.

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx = 1 - P(T \leq t) = 1 - F(t) \tag{2.1}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log[S(t)] \tag{2.2}$$

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{\frac{d}{du}[1 - S(u)]}{S(u)} du = -\log[S(t)] \tag{2.3}$$

Here, $S(t)$ is a survival function, $h(t)$, $H(t)$ are hazard function and cumulative hazard function, respectively, and $f(t)$ and $F(t)$ are distribution function and cumulative distribution function for survival time, respectively. And Figure 2.1 shows the graph of estimated survival function $\hat{S}(t)$, which is according to explanatory variable sex by applying Kaplan-Meier estimation based on 'cancer' data, which is related to patients with lung cancer. This graph was drawn using the program R 4.1.1 version, and the 'cancer' data is embedded in 'survival' package, which is used for survival analysis. And the p-value $p = 0.0013$ displayed in this graph represents the results of the log-rank test whether there is a difference in the survival function according to gender. After all, since the p-value is much smaller than the general significance level $\alpha = 0.05$, it can be interpreted that the survival rate varies according to gender. In addition, the number below the survival function graph represents the number of surviving patients over time.

As can be seen from Figure 2.1, in general, the survival function $S(t)$ starts from 1 and gradually decreases to 0 as time passes. This means that the survival

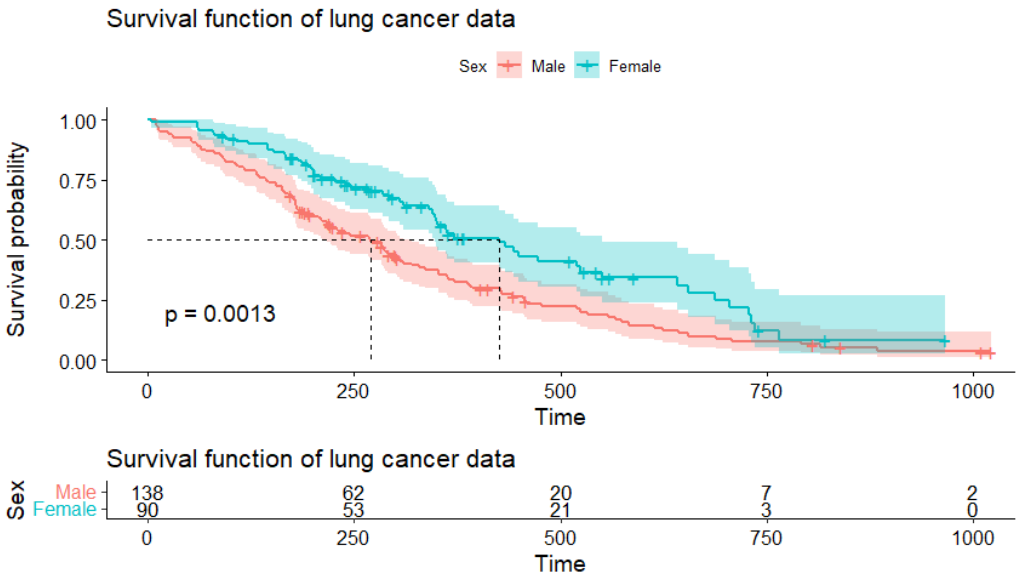


Figure 2.1: Survival function of lung cancer data

rate of patients decreases as time goes by. Of course, like the survival cure model that additionally considers a cure object, there is a case in which a situation in which there is an individual surviving over time due to a special cause is considered. However, as shown in Figure 2.1, a survival function $S(t)$ is usually used that starts at 1 and converges to 0. For details on the survival cure model, please refer to Sposto, R. (2002) et al. And as can be seen from the above equation, the survival function $S(t)$, the hazard function $h(t)$, and the distribution function of survival time $f(t)$ are related to each other. Therefore, if the survival function $S(t)$ is estimated from the survival data, the hazard function $h(t)$ and the distribution function of survival time $f(t)$ can be estimated naturally using this. In general, as shown in Table 2.1, in survival analysis, Kaplan-Meier estimation for estimating the survival function $S(t)$ with a non-parametric method, Nelson-Aalen estima-

tion for estimating the cumulative hazard function $H(t)$ with a non-parametric method, Cox proportional hazard model that models the hazard function $h(t)$ with a semi-parametric method under the assumption that the hazard between groups to be compared is uniformly proportional during the follow-up period, and Parametric model that models the hazard function $h(t)$ as a parametric method assuming a specific distribution are mainly used. In addition, various survival analysis methods exist, and detailed descriptions of them are provided by Sabin, C. and Petrie, A. (2019) and Chen et al. (2017) et al.

Kaplan-Meier Estimation	$\hat{S}(t) = \begin{cases} 1, & \text{if } t < t_1 \\ \prod_{t_i \le t}[1 - \frac{d_i}{Y_i}], & \text{if } t \ge t_1 \end{cases}, i = 1, 2, \ldots, D$	
Nelson-Aalen Estimation	$\tilde{H}(t) = \begin{cases} 0, & \text{if } t < t_1 \\ \Sigma_{t_i \le t} \frac{d_i}{Y_i}, & \text{if } t \ge t_1 \end{cases}, i = 1, 2, \ldots, D$	
Cox Proportional Hazard Model	$h(t\|X) = h_0(t)\exp(X\beta)$, $S(t\|X) = \{\exp[-H_0(t)]\}^{\exp(X\beta)}$	
Parametric Model	Exponential	$h(t) = \lambda_0 > 0$, $S(t) = \exp(-\lambda_0 t)$
	Weibull	$h(t) = \lambda_0\lambda_1 t^{\lambda_1 - 1}$, $S(t) = \exp(-\lambda_0 t^{\lambda_1})$, $(\lambda_0 > 0, \lambda_1 > 0)$
	Rayleigh	$h(t) = \lambda_0 + 2\lambda_1 t$, $S(t) = \exp[-(\lambda_0 t + \lambda_1 t^2)]$, $(\lambda_0 > 0, \lambda_1 \ge 0)$
	Gompertz	$h(t) = \exp(\lambda_0 + \lambda_1 t)$, $S(t) = \exp[\frac{1}{\lambda_1}\{\exp(\lambda_0) - \exp(\lambda_0 + \lambda_1 t)\}]$, $(\lambda_1 > 0)$
	Lognormal	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp[-\frac{\{\log(t)-\mu\}^2}{2\sigma^2}]$, $S(t) = 1 - \Phi[\frac{\log(t)-\mu}{\sigma}]$

Table 2.1: Representative example of survival analysis method

In addition, among the parts for the Kaplan-Meier estimator and the Nelson-Aalen estimator in Table 2.1, d_i is the number of events that occurred at time t_i, and Y_i is the number of individuals at risk at time t_i . And $\Phi(\cdot)$ of the Lognormal distribution part means cumulative distribution function of standard normal distribution. Additionally, in the case of the parametric model, using the hazard function $h(t)$ and the distribution function of survival time $f(t)$ presented in Table 2.1, and Equation 2.1 $\sim$ 2.3, we can derive $h(t)$, $f(t)$, $S(t)$, and the cumulative hazard function $H(t)$. Detailed proof of this will be omitted.

2.3 Synthetic response for censored data analysis

As mentioned in Section 2.1, the only variable related to survival time included in the survival data is the observed survival time for patients, $t = \min(y, c)$. Therefore, it is not possible to know the exact survival time for all patients, that is, the time it takes for the event of interest to occur. Therefore, when constructing a regression model for predicting survival time, it is somewhat unreasonable to set the observed survival time t as a response variable. Therefore, in this study, we used the synthetic response from Koul et al. (1981) to transform the data, and the form is as follows. For details, see Kim, J. (2018) and Lee, S. (2018).

$$Y^S = \frac{\delta t}{1 - G(T)} \tag{3.4}$$

$$G(t) = P(C \leq t), \delta = I(Y \leq C), T = \min(Y, C) \tag{3.5}$$

Here, the function $1 - G(T)$ is estimated using the Kaplan-Meier estimator on the assumption that the censoring variable T does not depend on the explanatory variable X. If there is a dependency, the following type of estimator proposed by Beran, R. (1981) should be used. In this estimator, W_0 means Nadaraya-Watson weight, h_0 means bandwidth, and K_0 means kernel function.

$$1 - \hat{G}(t|x) = \prod_{i=1}^{n} [1 - \frac{(1 - \delta_i) I_{\{\phi(T_i) \leq t\}} W_{0i}(x, h_0)}{\Sigma_{j=1}^{n} I_{\{\phi(T_i) \leq \phi(T_j)\}} W_{0j}(x, h_0)}] \tag{3.6}$$

$$W_{0i}(x, h_0) = \frac{K_0(\frac{X_i - x}{h_0})}{\Sigma_{j=1}^{n} K_0(\frac{X_j - x}{h_0})} \tag{3.7}$$

In this study, the Kaplan-Meier estimator of the following form was used based on the assumption that the censoring variable T does not depend on the explanatory

variable X.

$$1 - \hat{G}(t) = \prod_{i=1}^{n} [1 - \frac{(1-\delta_i) I(T_i \leq t)}{\Sigma_{j=1}^{n} I(T_j \geq T_i)}] \tag{3.8}$$

One thing to note is that when generating this synthetic response Y^S, the value of the estimator $1 - \hat{G}(t)$ for the denominator almost approaches 0, that is, in the case of $1 - \hat{G}(t) \approx 0$, the value of Y^S is infinitely divergent or undefined. In this study, a method of determining the truncation point was used in case such a situation occurs. If the value of the observed survival time t is greater than the point at which the cumulative probability becomes 0.98, $Y^S = 0$ is set.

The synthetic response Y^S has the following properties assuming some suitable conditions. Here, some suitable conditions are that the patient's actual survival time Y and censoring time C are independent of each other ($Y \perp C$), and the probability that Y is less than or equal to C is not depend on the explanatory variable X ($P(Y \leq C|X, Y) = P(Y \leq C|Y)$).

****

If $Y \perp C$ and $P(Y \leq C|X, Y) = P(Y \leq C|Y)$, then $E(Y|X = x) = E(Y^S|X = x)$

proof)

$E(Y^S|X = x) = E(\frac{\delta T}{1-G(T)}|X = x) = E(\frac{\delta Y}{1-G(Y)}|X = x)$

$= E(\frac{Y \times I(Y \leq C)}{1-G(Y)}|X = x) \approx E(Y\frac{1-G(Y)}{1-G(Y)}|X = x) = E(Y|X = x)$

$\because$ If $Y \leq C$, then $\delta = 1, T = Y$. So $Y^S = \frac{1 \times Y}{1-G(Y)} = \frac{\delta Y}{1-G(Y)}$

And if $Y > C$, then $\delta = 0, T = C$. So $Y^S = \frac{0 \times C}{1-G(C)} = \frac{\delta Y}{1-G(Y)}$

And $E(I(Y \leq C)|X = x) = P(Y \leq C|X = x, Y = y) = P(Y \leq C|Y) = 1 - G(Y)$

****

That is, if the patient's actual survival time Y and censoring time C are inde-

pendent of each other, and the probability that censoring does not occur does not depend on the explanatory variable X, the conditional mean of the synthetic response Y^S is coincides with the conditional mean of patient's actual survival time Y. Therefore, instead of putting the variable Y as a response variable, it can be said that it is reasonable to build a regression model with synthetic response Y^S as a new response variable. However, in the case of conditional variance, the case of synthetic response Y^S is larger than the case of variable Y as follows.

****

If $Y \perp C$ and $P(Y \leq C|X, Y) = P(Y \leq C|Y)$, then

$Var(Y^S|X = x) = Var(Y|X = x) + E(\frac{G(T)}{1-G(T)}Y^2|X = x)$

proof)

$Var(Y^S|X = x) = E((Y^S)^2|X = x) - \{E(Y^S|X = x)\}^2$

$= E[(\frac{\delta T}{1-G(T)})^2|X = x] - \{E(Y|X = x)\}^2$

$= E[(\frac{\delta T}{1-G(T)})^2|X = x] - E(Y^2|X = x) + Var(Y|X = x)$

$= E[\frac{\delta}{1-G(Y)}\frac{Y^2}{1-G(T)} - Y^2|X = x] + Var(Y|X = x)$ $(\because \delta^2 = \delta)$

$= E[\frac{I(Y \leq C)}{1-G(Y)}\frac{Y^2}{1-G(T)} - Y^2|X = x] + Var(Y|X = x)$

$\approx E[\frac{1-G(Y)}{1-G(Y)}\frac{Y^2}{1-G(T)} - Y^2|X = x] + Var(Y|X = x)$

$= Var(Y^S|X = x) = Var(Y|X = x) + E(\frac{G(T)}{1-G(T)}Y^2|X = x)$

****

Because of this property, when a regression model to predict the patient's actual survival time Y is constructed using the synthetic response Y^S, as the variance of the estimator increases, the volatility also increases, which reduces the stability and reliability of the model. There are disadvantages. To compensate for this

problem, the ridge regression method and the ensemble method can be applied, and these will be introduced in Chapters 3 and 4, respectively.

3 Kernel ridge censored regression analysis

This chapter describes the Kernel ridge censored regression analysis method. The main purpose of this method is to predict the time it takes for an event of interest to occur in a patient through survival data, and a multiple regression model is fitted with the synthetic response described in Section 3 in Chapter 2 as a response variable. Here, in the case of estimating the regression coefficient, we can more apply the ridge regression method that can supplement the multicollinearity problem based on regulation through the penalty function, and the kernel trick method that complements non-linear relationship between the response variable and the explanatory variable using space transformation. The Kernel ridge censored regression analysis is a method to build a model with stronger predictive power through this. Through this, if the survival data is analyzed and predicted the time taken until the event of interest occurs by fitting a model, good prediction results can be obtained based on higher accuracy than the normal method. Before introducing the Kernel ridge censored regression analysis method in earnest, Section 3.1 will first explain the multiple regression analysis that is the basis of this method.

3.1 Multiple regression analysis

Multiple regression analysis is a statistical analysis method proposed to analyze data based on the assumption that when there is a specific linear relationship between one response variable and two or more explanatory variables. For details on this, refer to Han, S. (2016) and Hwang, S. (2017). Let's look at an example to help you understand. If there are data (y_i, x_i), $i = 1, 2, \ldots, n$ including n observations and k explanatory variables, in multiple regression analysis, the data is analyzed with the basic assumption of the following model formula. As can be seen from this model, the multiple regression analysis method basically assumes that a linear relationship exists between the response variable and the explanatory variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \ldots, n \tag{1.1}$$

Here, ϵ_i is the error assumed by the multiple regression model, and normally assumed ϵ_i follow a normal distribution with mean 0 and variance σ^2 and are independent (iid: identical and independently distributed). The above regression model is usually expressed in the form of the following matrix and vector, and I think that this type of expression is a better way of expression in terms of computational and readability for theoretical proof.

$$y = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n) \tag{1.2}$$

Here, $y = (y_1, y_2, \ldots, y_n)^T$ is a response variable vector of length n, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is an error vector of length n. And $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a regression coefficient

vector of length $p(= k + 1)$, and X in Equation 1.3 is a data matrix that $n \times p$ dimension.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \tag{1.3}$$

Also, 0 is a zero vector with a length of n, and I_n is an identity matrix that $n \times n$ dimension. The estimator of the regression coefficient vector $\hat{\beta}$ is usually calculated using least squares estimation. Of course, the maximum likelihood estimation assuming the normality of the error can also be used. However, in the case of multiple regression analysis, the shape of the estimator calculated by these two methods is the same as a result. The calculation process to apply least squares estimation is as the following equation, and the equation obtained through this process is called the normal equation.

****

$Q = \Sigma_{i=1}^{n} \epsilon_i^2 = (y - X\beta)^T (y - X\beta)$

$\frac{\partial Q}{\partial \beta} = -2X^T(y - X\beta) = 0$

$(X^T X)\hat{\beta} = X^T y$

****

For the normal equation above, if the matrix $X^T X$ has an inverse matrix, that is, if the matrix X satisfies the full-rank property, then $\hat{\beta}$, which is the estimator of β, is determined to be a unique solution of the form. And the expected value and variance of this estimator $\hat{\beta}$ can be proved as follows. In conclusion, it can be confirmed that the estimator $\hat{\beta}$ is an unbiased estimator of the regression coeffi-

cient vector β.

****

$\hat{\beta} = (X^T X)^{-1} X^T y$

$E(\hat{\beta}) = E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T X \beta = \beta$

$Var(\hat{\beta}) = Var[(X^T X)^{-1} X^T y]$

$= (X^T X)^{-1} X^T (\sigma^2 I_n)[(X^T X)^{-1}]^T$

$= (X^T X)^{-1} X^T X[(X^T X)^{-1}]^T \sigma^2$

$= (X^T X)^{-1} X^T X[(X^T X)^T]^{-1} \sigma^2$

$= (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2$

$= (X^T X)^{-1} \sigma^2$

$\because y \sim N_n(X\beta, \sigma^2 I_n)$

****

For reference, calculating the $\hat{\beta}^{MLE}$, which is the estimator of regression coefficient vector β using maximum likelihood estimation is as follows.

****

If $x \sim N_r(\mu, \Sigma)$, then probability distribution function(pdf) of vector x is

$f(x|\mu, \Sigma) = (2\pi)^{-\frac{r}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}, x \in R^r$

And $y = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$.

Therefore, pdf of vector ϵ is

$f(\epsilon) = (2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\epsilon^T (\sigma^2 I_n)^{-1} \epsilon\}, \epsilon \in R^n$

Hence, likelihood function of β and σ^2 is

$L(\beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} exp\{-\frac{1}{2}(y - X\beta)^T (\sigma^2 I_n)^{-1}(y - X\beta)\}$

$\log L(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)$

$\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2) = -\frac{1}{\sigma^2} X^T (y - X\beta) = 0$

$\frac{\partial}{\partial \sigma^2} \log L(\beta, \sigma^2) = -\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\beta)^T (y - X\beta) = 0$

So, if matrix X is full-rank, maximum likelihood estimator of β and σ^2 are

$\hat{\beta}^{MLE} = (X^T X)^{-1} X^T y$ and

$\hat{\sigma^2}^{MLE} = \frac{1}{n} (y - X\hat{\beta}^{MLE})^T (y - X\hat{\beta}^{MLE})$

****

3.2 Ridge regression analysis

Ridge regression analysis is a type of penalized regression method represented together with Lasso regression (Least Absolute Shrinkage Selector Operator). This method is frequently used in various situations when applying the multiple regression analysis introduced in Section 3.1 when analyzing data, such as there is a multicollinearity problem that occurs because there is a correlation between explanatory variables, or when the estimator is not uniquely determined because the number of explanatory variables is greater than the number of observations. The essence of this method is that a certain amount of bias is allowed to obtain the estimator of the regression coefficient, and instead of giving up the advantage of the unbiased estimator, a more reliable estimator is obtained by appropriately regulating the variance to significantly reduce the variance. The estimator is calculated using a method similar to least squares estimation as follows. However, unlike general multiple regression analysis, to obtain an estimator using ridge regression analysis, penalty function with quadratic form $\lambda||\beta||^2 = \lambda\beta^T\beta$ multiplied by a positive real λ is added. This is a great feature of Ridge regression analysis compared with normal multiple regression analysis.

$$\hat{\beta} = \min_{\beta}\{||y - X\beta||^2 + \lambda||\beta||^2\} = \min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta\}, (\lambda > 0) \quad (2.4)$$

Here, λ means the ridge parameter and plays an important role in properly setting the ratio of bias and variance of the estimator calculated by ridge regression analysis. If the value of λ approaches 0, the bias of the estimator approaches 0 and the variance increases. Conversely, if the value of λ increases, the bias increases

but the variance decreases. Therefore, in order to calculate the estimator through ridge regression analysis, the appropriate parameter λ should be determined in advance. In general, using the method of cross validation, parameter λ, in which makes the most desirable condition that the estimated value of the test MSE (mean squared error) is calculated to the minimum, is selected. The $\hat{\beta}$, which is the estimator of β, obtained through this method as follows.

****

$$Q^* = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

$$\frac{\partial Q^*}{\partial \beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0$$

$$(X^TX + \lambda I_n)\hat{\beta} = X^Ty$$

$$\hat{\beta} = (X^TX + \lambda I_n)^{-1}X^Ty, (\lambda > 0)$$

****

The expected value of this estimator $\hat{\beta}$ is different from the regression coefficient vector β. Therefore, this estimator does not satisfy unbiased for β. However, in the above expression, the matrix $X^TX + \lambda I_n$ must have an inverse matrix by λ. Therefore, the estimator $\hat{\beta}$ is determined as the only one value according to the value of λ. For reference, the form of the penalty function used in Lasso regression analysis is $\lambda||\beta||$, and there are some cases where the value of the regression coefficient estimated according to the value of λ becomes 0. Because of these characteristics, lasso regression analysis includes the function of variable selection, and to calculate the estimator $\hat{\beta}$, soft-thresholding method is used, which is in order to estimate the regression coefficient for a specific explanatory variable, assume that the regression coefficients for the remaining explanatory variables are given and

estimates the regression coefficient. For this, please refer to Kim, J. (2018). Also, there is Elastic-net regression, a hybrid method created by combining Ridge regression analysis and Lasso regression analysis, which is extremely preferred when the number of variables p is greater than the number of observations n or strong multicollinearity problem is exist. For more details on penalized regression, see Friedman et al. (2007) and Hastie et al. (2011) and summarizes the core of each method as follows.

****

Ridge regression :

$\min_{\beta=(\beta_0,\ldots,\beta_k)^T}[\Sigma_{i=1}^{n}(y_i-\beta_0-\Sigma_{j=1}^{k}\beta_j x_{ij})^2+\lambda\Sigma_{j=1}^{k}\beta_j^2]$

Lasso regression :

$\min_{\beta=(\beta_0,\ldots,\beta_k)^T}[\Sigma_{i=1}^{n}(y_i-\beta_0-\Sigma_{j=1}^{k}\beta_j x_{ij})^2+\lambda\Sigma_{j=1}^{k}|\beta_j|]$

Elastic-net regression :

$\min_{\beta=(\beta_0,\ldots,\beta_k)^T}[\Sigma_{i=1}^{n}(y_i-\beta_0-\Sigma_{j=1}^{k}\beta_j x_{ij})^2+\lambda_1\Sigma_{j=1}^{k}\beta_j^2+\lambda_2\Sigma_{j=1}^{k}|\beta_j|]$

****

3.3 Kernel ridge regression analysis

The Ridge regression analysis method introduced in Section 3.2 proceeds with the analysis based on the assumption that there is a linear relationship between the response variable and the explanatory variables. However, since most of the data we actually collect and analyze is not organized for smooth processing, it is necessary for the researcher to organize it according to the purpose of the analysis in advance. Since it is common to show an almost nonlinear relationship structure, it is necessary to consider an appropriate transformation function, and in some cases, it is often necessary to consider the interaction effect between explanatory variables. The Kernel trick method is a method that can be applied when there are these problems. By applying this method, if an appropriate mapping function Φ, which is according to the characteristics of this data, is applied to the nonlinear data with complex structure, the data existing in the p dimension explanatory variable space can be transformed according to the characteristics. And the result is placed in a high-dimensional Hilbert space or feature space. If the data is transformed through this process, the same result as applying a transformation function suitable for the characteristics of the data can be obtained appropriately even without considering the transformation function in advance, and the analysis can be carried out by fitting a linear model to the transformed data. If the core of the Kernel trick method is expressed briefly, it can be represented as in Figure 3.1 below. In other words, since the original data shows a nonlinear relationship, even if it is a problem that it is difficult to fit an appropriate model when analyzing it as it is, but if it is appropriately transformed and moved to the feature space, the

relationship changes to linear, so it turns into a problem that is easy to fit model. That is the core of the Kernel trick method.

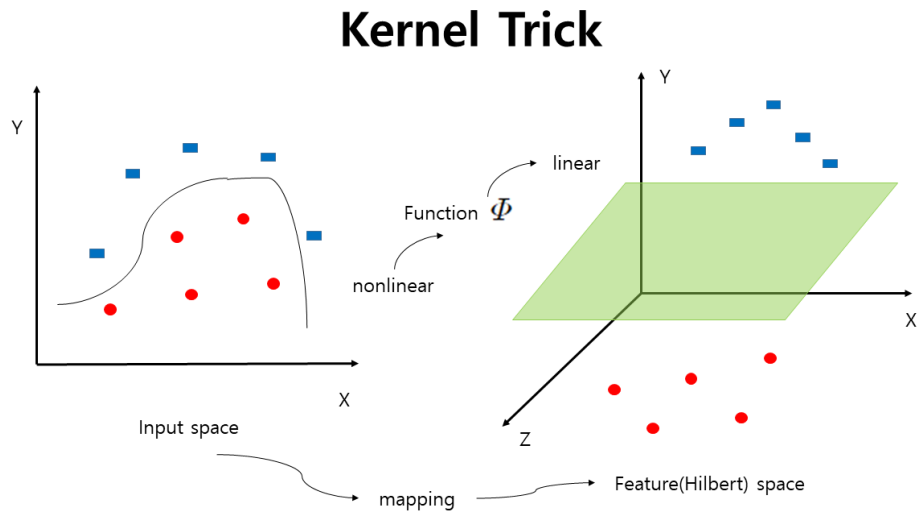


Figure 3.1: Kernel-trick method

From now on, how this Kernel trick method can be applied to ridge regression analysis is described. For more details, see Huh, M. (2015), Lee et al. (2016), Han, S. (2016), and Hwang, S. (2017). Suppose there are training data X with n observations and p dimension explanatory variable space, and n observations in this training data are $(y_i, x_i), i = 1, 2, \ldots, n$. In this regard, by using the mapping function Φ, it is possible to appropriately convert the n explanatory variable data in the training data X through the following method.

$$x_1, x_2, \ldots, x_n \rightarrow \Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n) \tag{3.5}$$

Here, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, i = 1, 2, \ldots, n$. The explanatory variable data

$\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$, which is transformed using the mapping function Φ, is placed on the feature space with high dimension, and the following regression model can be fitted using this.

$$E(y|x_1, \ldots, x_n) = \Phi(x_1)d_1 + \Phi(x_2)d_2 + \cdots + \Phi(x_n)d_n \tag{3.6}$$

Through the process of fitting the above model, a regression coefficient vector $d = (d_1, d_2, \ldots, d_n)^T$ of length n can be set. Space transformation through this kernel trick method is actually performed through calculation by the kernel function $k(\cdot, \cdot)$. Through this process, the projection of $\Phi(X)$, which is related to the linear combination of $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$, $v = d_1\Phi(x_1) + d_2\Phi(x_2) + \cdots + d_n\Phi(x_n)$, is calculated as follows.

$$\Sigma_{j=1}^{n} \langle \Phi(x_i), \Phi(x_j) \rangle d_j = \Sigma_{j=1}^{n} k_{i,j} d_j \tag{3.7}$$

Here, $k_{i,j} = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i)^T \Phi(x_j) = k(x_i, x_j)$ is the (i, j)th element of matrix $K = (K_{i,j})$, $i = 1, \ldots, n$, $j = 1, \ldots, n$. The response variable y is explained using the matrix K and the regression coefficient vector d obtained through this calculation process. For more details on this, see Schölkopf and Smola (2002). The kernel used to calculate the matrix K must satisfy the following Mercer's theorem, according to Minh et al. (2006) and Nguyen, V. (2015).

****

Mercer's theorem

A symmetric function $k_{i,j}$ can be expressed as an inner product

$$k_{i,j} = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i)^T \Phi(x_j)$$

for some Φ if and only if $k_{i,j}$ is positive semi-definite, is equal to,

$$\int k_{i,j} g(x_i) g(x_j) dx_i dx_j \geq 0, \forall g$$

or, equivalently $K = (K_{i,j})$, $i = 1, \ldots, n$, $j = 1, \ldots, n$ is positive semi-definite matrix for any collection $x_1, x_2, \ldots, x_n$.

****

In other words, when the kernel function $k_{i,j}$ is a continuous function in the form of an inner product, if the matrix K made from the value of the kernel function $k_{i,j}$ is a symmetric matrix and a positive semi-definite matrix, there exist Φ that satisfies $k_{i,j} = k_{j,i} = \langle \Phi(x_i), \Phi(x_j) \rangle$. This is the main core of Mercer's theorem. The form of the kernel that satisfies this Mercer's theorem exists in various ways as shown in Table 3.1. For details, see Karatzoglou et al. (2006) and Souza, C. R. (2010) et al.

In this study, Polynomial kernel and Gaussian kernel are applied among the kernels that satisfy Mercer's theorem presented in Table 3.1. These two kernels are as follows.

****

Polynomial kernel : $k_{i,j} = [\alpha(x_i^T x_j) + \beta]^\gamma, \alpha \neq 0, \gamma > 0$

Gaussian kernel : $k_{i,j} = \exp(\sigma ||x_i - x_j||^2), \sigma > 0$

Here, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$

****

In particular, compared to other kernels, Gaussian kernel has a strong characteristic that flexible application is possible even when prior information about the data to be analyzed is not known. The reason is that in the case of the Gaussian

Linear	$k(x,y) = x^T y + c$
Polynomial	$k(x,y) = [\alpha(x^T y) + \beta]^\gamma$
Gaussian (Radial Basis)	$k(x,y) = \exp(-\sigma\|\|x-y\|\|^2)$
Laplace	$k(x,y) = \exp(-\sigma\|\|x-y\|\|)$
ANOVA	$k(x,y) = [\Sigma_{k=1}^n\{-\sigma(x^k - y^k)^2\}]^d$
Sigmoid	$k(x,y) = \tanh[\alpha(x^T y) + \beta]$
Rational Quadratic	$k(x,y) = 1 - \frac{\|\|x-y\|\|^2}{\|\|x-y\|\|^2+c}$
Multiquadratic	$k(x,y) = \sqrt{\|\|x-y\|\|^2 + c^2}$
Inverse Multiquadratic	$k(x,y) = \frac{1}{\sqrt{\|\|x-y\|\|^2+c^2}}$
Bessel	$k(x,y) = \frac{(Bessel)^n_{(\nu+1)}(\sigma\|\|x-y\|\|)}{(\|\|x-y\|\|)^{-n(\nu+1)}}$
Cauchy	$k(x,y) = \frac{1}{1+(\frac{\|\|x-y\|\|}{\sigma})^2}$
Generalized T-Student	$k(x,y) = \frac{1}{1+\|\|x-y\|\|^d}$
Power(conditionally positive definite)	$k(x,y) = -\|\|x-y\|\|^d$
Log(conditionally positive definite)	$k(x,y) = -\log(\|\|x-y\|\|^d + 1)$
Triangular(positive definite in R)	$k(x,y) = 1 - \frac{\|\|x-y\|\|}{2\gamma}, \|\|x-y\|\| < 2\gamma$

Table 3.1: Various kernel types that satisfy Mercer's theorem

kernel, it can be expressed as the sum of infinite series by expanding and expressing the expression of the function through Taylor series expansion as shown in the following equation. This means that it can be expressed as the inner product of the transformed explanatory variables $\Phi(x_i)$ and $\Phi(x_j)$ in the form of vectors with infinite dimensions. That is, if the Gaussian kernel is applied, the data to be analyzed is appropriately transformed and moved to the feature space of infinite space.

****

$$k_{i,j} = \exp(-\sigma||x_i - x_j||^2)$$
$$= \{\exp(-||x_i - x_j||^2)\}^\sigma$$
$$= [\exp(-||x_i||^2)\exp(-||x_j||^2)\Sigma_{r=1}^{\infty}\frac{(x_i^T x_j)^r}{r!}]^\sigma$$
$$= \Phi(x_i)^T\Phi(x_j) = \langle\Phi(x_i), \Phi(x_j)\rangle$$

****

In this study, when a polynomial kernel is applied, the degree parameter γ that determines the flexibility of the boundary created by the kernel is fixed as $\gamma = 3$. The reason is that the larger the number of explanatory variables, the greater the complexity, so to compensate for this, the size of γ should be reduced. It was determined that fixing $\gamma = 3$ would be a sufficient solution. Based on the same logic, the scale parameter α was set to $\alpha = \frac{1}{p^2}$, and the offset parameter β was fixed to $\beta = 1$. And in the case of applying the Gaussian kernel, the tuning parameter σ that determines the flexibility of the boundary created by the kernel was set $\sigma = \frac{1}{p}$ in order to set the boundary in a simple form as the number of explanatory variables increases. Here, p means the dimension of the explanatory variable space

where the original data in the state before conversion through the kernel function is placed. Using this kernel transformation, the following regression model can be obtained.

$$y = Kd + \epsilon \tag{3.8}$$

And it is worth noting that the inverse matrix for the matrix K obtained through transformation does not always exist. To compensate for this, the regression coefficient vector d is estimated by applying the penalty function $\lambda d^T K d$ in the form of ridge regression analysis. The regression coefficient vector $\hat{d}$ estimated through this is calculated as follows.

****

$Q^* = (y - Kd)^T (y - Kd) + \lambda d^T K d$

$\frac{\partial Q^*}{\partial d} = -2K(y - Kd) + 2\lambda K d = 0$

$\hat{d} = \min_d [(y - Kd)^T (y - Kd) + \lambda d^T K d]$

$= (K + \lambda I_n)^{-1} y, (\lambda > 0)$

****

Here, the optimal value of the ridge parameter λ is selected through k-fold cross validation. In this study, $k = 5$ was set, and the optimal λ value was determined and selected as the optimal situation when the mean value of RMSE (Root Mean Square Error) was the minimum. In this way, in order to proceed with the evaluation of the test data X^* through the training data X, several procedures are necessary. First, compute the projection of $\Phi(X^*)$, which is related to $v = d_1\Phi(x_1) + d_2\Phi(x_2) + \cdots + d_n\Phi(x_n)$ that the linear combination of transformed explanatory variable data $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$, that is transformed using ker-

nel function $k(\cdot, \cdot)$. Then, based on this, we need to calculate the matrix K^* to be used for test data evaluation. The calculation process for this is as follows. Here, n means the number of observations in the training data, and u means the number of observations in the test data.

****

$\Sigma_{j=1}^{n} \langle \Phi(x_i^*), \Phi(x_j) \rangle d_j = \Sigma_{j=1}^{n} k_{i,j}^* d_j$

$k_{i,j}^* = \langle \Phi(x_i^*), \Phi(x_j) \rangle = \Phi(x_i^*)^T \Phi(x_j) = k(x_i^*, x_j)$

****

And in this study, the following types of Polynomial kernel and Gaussian kernel are applied.

****

Polynomial kernel : $k_{i,j}^* = [\alpha(x_i^{*T} x_j) + \beta]^\gamma, \alpha = \frac{1}{p^2}, \beta = 1, \gamma = 3$

Gaussian kernel : $k_{i,j}^* = \exp(\sigma ||x_i^* - x_j||^2), \sigma = \frac{1}{p}$

Here, $i = 1, 2, \ldots, u$, $j = 1, 2, \ldots, n$

****

Here, $k_{i,j}^*$ is the (j, i)th element of matrix $K^* = (k_{i,j}^*)$, $i = 1, 2, \ldots, u$, $j = 1, 2, \ldots, n$ and x_i^* is the explanatory variable data for ith element in test data X^*. Through this process, we can do final evaluation of the test data X^* using the matrix K^* and the estimator $\hat{d}$, which is calculated through the training data X. In other words, using the matrix K^* and the estimator $\hat{d}$, we can calculate $\hat{y^*} = (\hat{y_1^*}, \hat{y_2^*}, \ldots, \hat{y_u^*})^T$, which is the estimator of response variable $y^* = (y_1^*, y_2^*, \ldots, y_u^*)^T$ from test data, using training data as follows.

$$\hat{y^*} = K^{*T} \hat{d} \tag{3.9}$$

Based on this estimator, the following RMSE value is calculated, and the smaller this value is, the higher the predictive power of the built model is judged.

$$\sqrt{MSE} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^* - \hat{y_i^*})^2} \tag{3.10}$$

3.4 Kernel ridge censored regression analysis

Kernel ridge censored regression can be expressed by changing the response variable y_i in the data in the kernel ridge regression analysis described in Section 3.3 to the synthetic response y_i^S introduced in Section 3 of Chapter 2. And the form can be expressed as follows. Here, $i = 1, 2, \ldots, n$.

$$y_i^S = \frac{\delta_i t_i}{1 - G(t_i)} \tag{4.11}$$

$$G(t_i) = P(C \leq t_i), \delta_i = I(y_i \leq c_i), t_i = \min(y_i, c_i) \tag{4.12}$$

In other words, using the observed survival time t_i and censoring indicator variable δ_i in the survival data to be analyzed, calculate the synthetic response y_i^S that replaces patient's real survival time y_i and applied it to kernel ridge regression analysis. In this study, for estimating the function $1 - G(t_i)$, we use Kaplan-Meier estimator based on the assumption that censoring variable does not depend on explanatory variable $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. Other than that, the details are the same as those of kernel ridge regression analysis, so a detailed description will be omitted. However, in Kernel ridge censored regression analysis, synthetic response $y^{S*} = (y_1^{S*}, y_2^{S*}, \ldots, y_u^{S*})^T$ is additionally used. So when evaluating the built model, the evaluation criterion of predictive power should be divided into two cases, synthetic response variable $y^{S*} = (y_1^{S*}, y_2^{S*}, \ldots, y_u^{S*})^T$ and the actual response variable $y^* = (y_1^*, y_2^*, \ldots, y_u^*)^T$. In other words, if the evaluation criterion is synthetic response, the RMSE is calculated as $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^u (y_i^{S*} - \hat{y_i^{S*}})^2}$, and when the evaluation criterion is the actual response variable, the RMSE is calculated as $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^u (y_i^* - \hat{y_i^{S*}})^2}$. However, it is necessary to take

into account that the actual survival time of all patients, that is, the time taken until the event of interest occurs, is not recorded in the actual survival data. In this study, in the simulation study that can generate data arbitrarily, both criteria described above will be used to evaluate the performance of the methodology. But in in real data analysis based on real data, only one criteria, which is related with synthetic response, will be used to evaluate the performance of the methodology.

4 Ensemble method

This chapter introduces the ensemble method, a method that can greatly reduce the variability of the estimator by applying the bootstrap method. The ensemble method is a method derived from the tree-model method, which is a type of machine learning, and is a technique to generate more accurate results by creating multiple predictors or classifiers and then combining them. In this study, among various ensemble methods, bagging and random forests are used. The key shared by these two methodologies is that the average of the results obtained using bootstrap samples extracted independently of each other is used as the estimator to significantly reduce the variance to promote the stability of the prediction. If this method is applied, more accurate and reliable predictions can be made. In addition to bagging and random forests, various boosting algorithms such as AdaBoost (Adaptive Boosting), GBM (Gradient Boosting Machine), XGBoost, and LightGBM also belong to the ensemble method. The key to these methods is to increase the performance of the model by sequentially combining several weak learners. For this, see Freund et al. (1999), Lee, J. (2020), Han, S. (2016), and Hwang, S. (2017).

Let's look at a simple example. If it is assumed that the random n samples $X_1, X_2, \ldots, X_n$ are independent and the variances are all equal to σ^2, the variance

of their sample mean $\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$ will be greatly reduced, like $\frac{\sigma^2}{n}$. That is, the estimator obtained by extracting a large number of samples and averaging them has a very small variance compared to the estimator obtained through the general method without going through such a process, so the variability is greatly reduced. Therefore, the estimator obtained through averaging shows better performance than the general estimator in terms of reliability. And through the form of the formula, it is easy to understand that the number of samples n and the variance $\frac{\sigma^2}{n}$ are inversely proportional to each other. Therefore, if you want to obtain a more preferable estimator in terms of reliability compared to a general estimator, you can extract a large number of samples, obtain an estimator for each, and significantly reduce the variance value through the process of averaging them. However, it is practically impossible to increase the number of samples to the point of being near infinity. Therefore, it can be said that methods such as bagging or random forests that extract and analyze a sufficient number of bootstrap samples from the original data are analysis methods that overcome these limitations to some extent.

4.1 Bagging

Bagging (bootstrap aggregation) is a method of ensemble method that extracts and analyzes several bootstrap samples with the same number of observations compared to the training data by repeatedly sampling with replacement on one training data. All extracted bootstrap samples are analyzed using the same type of algorithm-based predictor or classifier, and it has the advantage of supplementing the problem of overfitting that may occur when the model is fitted. Here, overfitting means a phenomenon in which the training data is over-learned in machine learning, so that the error decreases for the training data, but the error increases for the test data. Conversely, a phenomenon in which an inappropriate model is fitted due to insufficient learning on the training data may occur, which is called underfitting. For details on this, see Hastie et al. (2011) and James et al. (2014). Figure 4.1 shows the bootstrap technique. As can be seen from this figure, the core of bootstrap is to extract a large number of bootstrap samples from the training data, obtain all the estimators based on each sample, and then properly combine them. The estimator obtained in this way has a better performance than the estimator obtained by using the training data once. And Figure 4.2 simply expresses the concepts of overfitting and underfitting. Through this figure, it can be confirmed that it is important to implement the algorithm to avoid overfitting and underfitting when fitting the model.

For example, suppose there are B estimators calculated for each sample ob-

tained through bootstrap of B times as follows.

$$\hat{g_{bag}^1}(x), \hat{g_{bag}^2}(x), \ldots, \hat{g_{bag}^B}(x) \tag{1.1}$$

The bagging estimator can be obtained by averaging these B estimators. Through this process, the following type of bagging estimator with relatively small variance can be created.

$$\hat{g_{bag}}(x) = \frac{1}{B}\Sigma_{b=1}^{B}\hat{g_{bag}^b}(x) \tag{1.2}$$

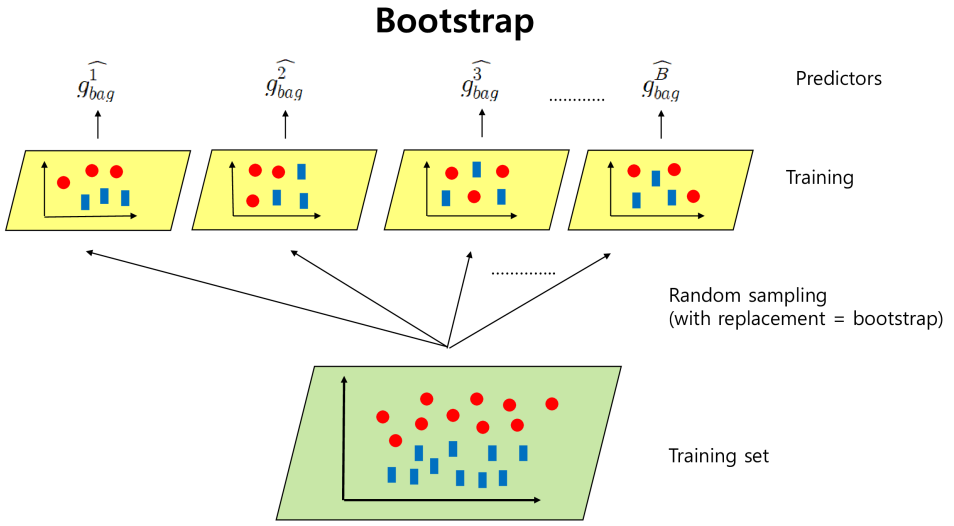


Figure 4.1: Bootstrap

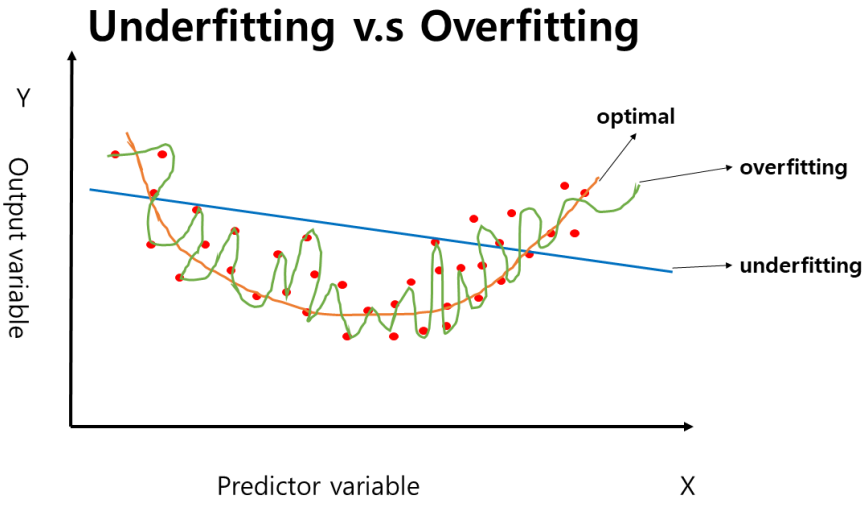


Figure 4.2: Overfitting and Underfitting

4.2 Random forests

Random forests is one of the ensemble methods proposed to compensate for the problem of bagging, and the overall process of obtaining the estimator is similar to the principle of bagging. In this method, similar to the case of bagging introduced in Section 4.1, a final estimator can be obtained by calculating the average of several estimators using the bootstrap method. However, the feature of selecting and using only some of these explanatory variables rather than using all of the explanatory variables for each sample gives a big difference in improving predictive power and reliability compared to bagging that uses all explanatory variables. If the bagging technique to obtain the estimator by including all explanatory variables for the bootstrap sample is used, the accuracy of estimation may be lowered because there may be a correlation between the submodels made by each bootstrap sample. In order to compensate for this shortcoming, the correlation can be greatly reduced by selecting only some of all explanatory variables and applying them to each bootstrap sample. Of course, if you go through this process, bias will inevitably occur. However, since the correlation is greatly reduced, the variance can be reduced to a larger extent, thereby canceling the effect on the bias and reducing it. In this way, more accurate predictions can be made.

Let's take a look at an example. Suppose that there is training data with the number of explanatory variables is p, and using this training data, B samples are extracted through the bootstrap of B times. Here, in the case of random forests, unlike the case of bagging, for each sample extraction, only m explanatory

variables should be randomly selected from all p explanatory variables in the training data and included in the bootstrap sample. In general, in the case of m, $m \approx \frac{p}{3}$ or $m \approx \sqrt{p}$ is set. In this study, $m \approx \sqrt{p}$ is set. When the value of m is not expressed in the form of a natural number, rounding to the nearest decimal point was used to determine the number of explanatory variables to be included in the bootstrap sample. Also, it is necessary to pay attention to the fact that the m explanatory variables to be included in each bootstrap sample must be selected differently each time the bootstrap sample is extracted through random sampling. Through this process, the following B estimators are calculated.

$$\hat{g_{rf}^1}(x), \hat{g_{rf}^2}(x), \ldots, \hat{g_{rf}^B}(x) \tag{2.3}$$

The random forests estimator can be obtained by averaging these B estimators. Through this, it is possible to obtain an excellent random forests estimator of the following form, which is more accurate than the bagging estimator.

$$\hat{g_{rf}}(x) = \frac{1}{B}\Sigma_{b=1}^{B}\hat{g_{rf}^b}(x) \tag{2.4}$$

The random forests estimator obtained through this process shows more desirable performance in terms of accuracy and reliability when compared with the estimator calculated by bagging technique. Chapter 5, which follows, introduces how ensemble methods such as bagging and random forests can be applied to kernel ridge censored regression analysis.

5 Kernel ridge censored regression analysis using ensemble method

This chapter explains how the ensemble method introduced in Chapter 4 can be applied to kernel ridge censored regression analysis. As mentioned in Chapter 4, the ensemble method is a methodology that can greatly increase the accuracy and reliability of the estimation by significantly reducing the correlation between the variance of the estimator and the explanatory variables using a large number of independent bootstrap samples. If this is applied to kernel ridge censored regression analysis, it is possible to construct a model that accurately predicts the patient's survival time, that is, the time it takes until the event of interest occurs.

5.1 Kernel ridge censored regression analysis using Bagging technique

Assume that there is training data with censoring X in which the number of observations is n and the explanatory variable space is p dimension, and the following n observations are included in this data. In this study, function $1-G(t_i)$ is estimated using Kaplan-Meier estimator based on the assumption that censoring variable t_i is not depend on explanatory variable $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

$$(y_i, c_i, x_i, \delta_i, t_i, y_i^S), i = 1, 2, \ldots, n \tag{1.1}$$

$$y_i^S = \frac{\delta_i t_i}{1 - \hat{G}(t_i)} \tag{1.2}$$

$$G(t_i) = P(C \leq t_i), \delta_i = I(y_i \leq c_i), t_i = \min(y_i, c_i) \tag{1.3}$$

Assuming that the bth sample X^b is extracted through bootstrap for this data, suppose we have the following synthetic response $y_i^{Sb} = \frac{\delta_i^b t_i^b}{1-\hat{G}(t_i^b)}$ and explanatory variable $x_i^b = (x_{i1}^b, x_{i2}^b, \ldots, x_{ip}^b)^T$ in this sample. Here, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$.

$$(y_i^{Sb}, x_i^b) \tag{1.4}$$

About this, by using the mapping function Φ, n explanatory variable data for bootstrap training data X^b can be converted as follows.

$$x_1^b, x_2^b, \ldots, x_n^b \to \Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b) \tag{1.5}$$

Explanatory variable data $\Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b)$ converted using mapping function Φ is located in the feature space with high dimension. Using this, the following

regression model can be fitted.

$$E(y^{Sb}|x_1^b,\ldots,x_n^b) = \Phi(x_1^b)d_1^b + \Phi(x_2^b)d_2^b + \cdots + \Phi(x_n^b)d_n^b \tag{1.6}$$

Through the process of fitting the above model, the regression coefficient vector $d^b = (d_1^b, d_2^b, \ldots, d_n^b)^T$, which is length n, is set. Space transformation through this kernel trick is actually performed through the calculation by the kernel function $k(\cdot,\cdot)$. In other words, the projection of $\Phi(X^b)$, which is related with linear combination of $\Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b)$, $v^b = d_1^b\Phi(x_1^b) + \cdots + d_n^b\Phi(x_n^b)$, is obtained through the following calculation.

$$\Sigma_{j=1}^n \left\langle \Phi(x_i^b)\Phi(x_j^b) \right\rangle d_j^b = \Sigma_{j=1}^n k_{i,j}^b d_j^b \tag{1.7}$$

Here, $k_{i,j}^b = \left\langle \Phi(x_i^b)\Phi(x_j^b) \right\rangle = \Phi(x_i^b)^T\Phi(x_j^b) = k(x_i^b, x_j^b)$ is an (i,j)th element of matrix $K^b = (k_{i,j}^b)$, $i = 1,\ldots,n$, $j = 1,\ldots,n$. The synthetic response y^S is explained using the matrix K^b and the regression coefficient vector d^b through this process. Polynomial kernel and Gaussian kernel are applied in this study. These two kernels are as follows.

****

Polynomial kernel : $k_{i,j}^b = [\alpha(x_i^{bT}x_j^b) + \beta]^\gamma, \alpha = \frac{1}{p^2}, \beta = 1, \gamma = 3$

Gaussian kernel : $k_{i,j}^b = \exp(\sigma||x_i^b - x_j^b||^2), \sigma = \frac{1}{p}$

Here, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,n$

****

Through this kernel transformation, the following regression model can be obtained.

$$y^{Sb} = K^b d^b + \epsilon^b \tag{1.8}$$

And to compensate for the fact that the inverse matrix for the matrix K^b calculated through transformation does not always exist, we estimate regression coefficient vector d^b by applying a penalty function in the form of ridge regression analysis $\lambda^b d^{bT} K^b d^b$. The regression coefficient vector $\hat{d^b}$ estimated through this is calculated as follows. Here, the value of the optimal ridge parameter λ^b is decided through the out-of-bag (OOB) method in which random B^* bootstrap samples are extracted from the training data. That is, the extracted sample is set as the new training data, and when this sample is extracted, the remaining observations that are not drawn are used as validation data for the sample to go through the process of finding the optimal condition. Here, the optimal condition means the case in which the average value of B^* root mean squared error (RMSE) obtained for each value of λ^b is the smallest.

****

$$Q^{b*} = (y^{Sb} - K^b d^b)^T (y^{Sb} - K^b d^b) + \lambda^b d^{bT} K^b d^b$$

$$\frac{\partial Q^{b*}}{\partial d^b} = -2K^b (y^{Sb} - K^b d^b) + 2\lambda^b K^b d^b = 0$$

$$\hat{d^b} = \min_{d^b} [(y^{Sb} - K^b d^b)^T (y^{Sb} - K^b d^b) + \lambda^b d^{bT} K^b d^b]$$

$$= (K^b + \lambda^b I_n)^{-1} y^{Sb}, (\lambda^b > 0)$$

****

In this way, in order to evaluate the test data X^* based on the bth bootstrap training data X^b extracted from the training data X, compute the projection of $\Phi(X^*)$, which is related with the linear combination of explanatory variable data $\Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b)$ that converted through kernel function $k(\cdot, \cdot)$, $v^b = d_1^b \Phi(x_1^b) + \cdots + d_n^b \Phi(x_n^b)$, and calculate matrix K^{b*} for using test data evaluation.

The calculation process for this is as follows. Here, n is the number of observations in the training data, and u is the number of observations in the test data.

****

$\Sigma_{j=1}^{n} \left\langle \Phi(x_i^*), \Phi(x_j^b) \right\rangle d_j^b = \Sigma_{j=1}^{n} k_{i,j}^{b*} d_j^b$

$k_{i,j}^{b*} = \left\langle \Phi(x_i^*), \Phi(x_j^b) \right\rangle = \Phi(x_i^*)^T \Phi(x_j^b) = k(x_i^*, x_j^b)$

****

And in this study, the following types of Polynomial kernel and Gaussian kernel are applied.

****

Polynomial kernel : $k_{i,j}^{b*} = [\alpha(x_i^{*T} x_j^b) + \beta]^\gamma, \alpha = \frac{1}{p^2}, \beta = 1, \gamma = 3$

Gaussian kernel : $k_{i,j}^{b*} = \exp(\sigma ||x_i^* - x_j^b||^2), \sigma = \frac{1}{p}$

Here, $i = 1, 2, \ldots, u$, $j = 1, 2, \ldots, n$

****

Here, $k_{i,j}^{b*}$ is an (j, i)th element of matrix $K^{b*} = (k_{i,j}^{b*})$, $i = 1, \ldots, u$, $j = 1, \ldots, n$ and x_i^* is the explanatory variable data of ith observation in test data X^*. Using matrix K^{b*} and bootstrap training data X^b obtained through this process, compute estimator $\hat{d^b}$ and use for proceed final evaluation of test data X^*. As a result, using the matrix K^{b*} and the estimator $\hat{d^b}$, we can calculate the estimator of synthetic response $y^{Sb*} = (y_1^{Sb*}, y_2^{Sb*}, \ldots, y_u^{Sb*})^T$ that obtained from bootstrap training data, $y^{\hat{S}b*} = (y_1^{\hat{S}b*}, y_2^{\hat{S}b*}, \ldots, y_u^{\hat{S}b*})^T$ as follows.

$$y^{\hat{S}b*} = K^{b*T} \hat{d^b} \tag{1.9}$$

This process is carried out for each of B bootstrap training data X^b, $b = 1, 2, \ldots, B$. This gives a total of B estimators $y^{\hat{S}b*} = K^{b*T} \hat{d^b}$, $b = 1, 2, \ldots, B$. After obtaining

this, an estimate of the synthetic response using the bagging technique for the test data X^* can be obtained by averaging it, and the form is as follows.

$$y_{bag}^{\hat{S}*} = \frac{1}{B}\Sigma_{b=1}^{B} y^{\hat{S}b*} = (y_{bag,1}^{\hat{S*}}, y_{bag,2}^{\hat{S*}}, \ldots, y_{bag,u}^{\hat{S*}})^T \tag{1.10}$$

In kernel ridge censored regression analysis, synthetic response $y^{S*} = (y_1^{S*}, y_2^{S*}, \ldots, y_u^{S*})^T$ is used when constructing the model. So when evaluating the built model, it is necessary to divide the evaluation criteria of predictive power into two types when evaluating the constructed model, such as using standard of synthetic response $y^{S*} = (y_1^{S*}, y_2^{S*}, \ldots, y_u^{S*})^T$ and using standard of real response variable $y^* = (y_1^*, y_2^*, \ldots, y_u^*)^T$. Accordingly, in this study, when the evaluation criterion is a synthetic response, set RMSE like $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_{bag,i}^{\hat{S*}})^2}$ for evaluate the performance, and when the evaluation criterion is a real response variable, set RMSE like $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^* - y_{bag,i}^{\hat{S*}})^2}$ for evaluate the performance.

5.2 Kernel ridge censored regression analysis using Random Forests technique

As in the case of applying the bagging technique, let's assume that there exist training data X including censoring such that the number of observations is n and the explanatory variable space is p dimension. And suppose that the following n observations are in this data. Of course, here, as in the case of bagging, function $1 - G(t_i)$ is estimated using Kaplan-Meier estimator based on the assumption that censoring variable t_i is not depend on explanatory variable $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

$$(y_i, c_i, x_i, \delta_i, t_i, y_i^S), i = 1, 2, \ldots, n \tag{2.11}$$

$$y_i^S = \frac{\delta_i t_i}{1 - \hat{G}(t_i)} \tag{2.12}$$

$$G(t_i) = P(C \leq t_i), \delta_i = I(y_i \leq c_i), t_i = \min(y_i, c_i) \tag{2.13}$$

Assuming that the bth sample X^b is extracted through bootstrap for this data, suppose we have synthetic response $y_i^{Sb} = \frac{\delta_i^b t_i^b}{1-\hat{G}(t_i^b)}$ and explanatory variable $x_i^b = (x_{i1}^b, x_{i2}^b, \ldots, x_{ip}^b)^T$ in this sample as follows. Here, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$.

$$(y_i^{Sb}, x_i^b) \tag{2.14}$$

In the case of applying the random forests technique here, unlike the bagging technique, from all p explanatory variables, only m of them are selected and included in each B bootstrap training data X^b, $b = 1, 2, \ldots, B$. In this study, $m \approx \sqrt{p}$ was set, and when the value of m does not appear as a natural number,

rounding to the decimal point was used. One thing to note is that all explanatory variables included in each bootstrap training data should be selected differently. Through this process, the correlation between explanatory variables can be greatly reduced, and more accurate predictions can be obtained through this.

As a result, the process of obtaining an estimator of the synthetic response to the test data X^* using the random forests technique has only one difference that when extracting bootstrap training data X^b, $b = 1, 2, \ldots, B$, only some explanatory variables are selected and included, compared with the bagging technique. The rest of the calculation process and the model evaluation process using the random forests technique are the same as those of the bagging technique. Therefore, a detailed description thereof will be omitted. However, in order to distinguish the estimator and the RMSE for performance evaluation obtained through the random forests technique from the estimator and the RMSE for performance evaluation obtained through the bagging technique, they are denoted as follows.

$$y_{rf}^{\hat{S}*} = (y_{rf,1}^{\hat{S}*}, y_{rf,2}^{\hat{S}*}, \ldots, y_{rf,u}^{\hat{S}*})^T \tag{2.15}$$

$$\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_{rf,i}^{\hat{S}*})^2} \tag{2.16}$$

$$\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_{rf,i}^{\hat{S}*})^2} \tag{2.17}$$

The description and interpretation, evaluation of simulation and real data analysis for prove the excellent performance of kernel ridge censored regression analysis using ensemble method proposed in this study compared with other methodologies such as sub-sampling proposed by Huh, M. (2015) will proceed in Chapter 6.

6 Simulation and real data analysis (Kernel ridge censored regression analysis using ensemble method)

This chapter describes the simulation and real data analysis performed to prove that the Kernel ridge censored regression analysis using ensemble method proposed in Chapter 2 to 5 has overall superior predictive power compared to other methodologies. First, in the case of simulation, training data and test data were generated according to the purpose of the study, respectively, and in the case of real data analysis, the survival data prepared by the actual study was arbitrarily divided into train:test=7:3 ratio. And all simulations and real data analysis were performed using the program R 4.1.1 version.

6.1 Simulation

In simulation, the following methodologies were compared and analyzed for training data and test data arbitrarily generated according to the purpose of the methodology to be evaluated.

1) **PKR1** : Polynomial Kernel Ridge Regression with Synthetic Response Y^S

2) **PKRS1** : Polynomial Kernel Ridge Regression with Sub-sampling and Synthetic Response Y^S

3) **PKRB1** : Polynomial Kernel Ridge Regression with Bagging and Synthetic Response Y^S

4) **PKRR1** : Polynomial Kernel Ridge Regression with Random Forest and Synthetic Response Y^S

5) **GKR1** : Gaussian Kernel Ridge Regression with Synthetic Response Y^S

6) **GKRS1** : Gaussian Kernel Ridge Regression with Sub-sampling and Synthetic Response Y^S

7) **GKRB1** : Gaussian Kernel Ridge Regression with Bagging and Synthetic Response Y^S

8) **GKRR1** : Gaussian Kernel Ridge Regression with Random Forest and Synthetic Response Y^S

9) **PKR2** : Polynomial Kernel Ridge Regression with Generated(Original) Response Y

10) **PKRS2** : Polynomial Kernel Ridge Regression with Sub-sampling and Gen-

erated(Original) Response Y

11) **PKRB2** : Polynomial Kernel Ridge Regression with Bagging and Generated(Original) Response Y

12) **PKRR2** : Polynomial Kernel Ridge Regression with Random Forest and Generated(Original) Response Y

13) **GKR2** : Gaussian Kernel Ridge Regression with Generated(Original) Response Y

14) **GKRS2** : Gaussian Kernel Ridge Regression with Sub-sampling and Generated(Original) Response Y

15) **GKRB2** : Gaussian Kernel Ridge Regression with Bagging and Generated(Original) Response Y

16) **GKRR2** : Gaussian Kernel Ridge Regression with Random Forest and Generated(Original) Response Y

Among the 16 methodologies presented above, sub-sampling used in PKRS1, GKRS1, PKRS2, and GKRS2 is a methodology proposed by Huh, M. (2015). The principle of this methodology will be briefly mentioned while explaining the simulation step. In this simulation, random simulation data was created through the following method.

****

$y_i \sim N(\mu_i, 1^2)$

$\mu_i = 1 + \Sigma_{j=1}^{p} (\frac{x_{ij}}{10})^j$, $x_{ij} \sim U(-10, 10)$

$c_i \sim N(a, 1^2)$, $t_i = \min(y_i, c_i)$

$y_i^S = \frac{\delta_i t_i}{1-\hat{G}(t_i)}$, $\delta_i = I(y_i \leq c_i)$

****

Here, $1 - \hat{G}(t_i)$ was calculated using the Kaplan-Meier estimator introduced in Chapter 2. And the response variable y_i is generated to follow a normal distribution with mean $\mu_i = 1 + \Sigma_{j=1}^{p}(\frac{x_{ij}}{10})^j$ and standard deviation 1 to satisfy strong non-linearity. In addition, in the case of the censoring variable c_i, the average a was set appropriately according to the simulation situation, and it was designed to represent the desired censoring rate. And for methodologies related to Synthetic Response (PKR1, PKRS1, PKRB1, PKRR1, GKR1, GKRS1, GKRB1, GKRR1), RMSE(Root Mean Squared Error) is calculated like $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_i^{\hat{S}*})^2}$, and for methodologies related to Original Response (PKR2, PKRS2, PKRB2, PKRR2, GKR2, GKRS2, GKRB2, GKRR2), RMSE is calculated like $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_i^{\hat{S}*})^2}$. The number of explanatory variables for arbitrary training data and test data created for simulation is set to $p = 3, 5, 7, 9$, and the censoring rate is set to $0\%, 10\%, 30\%, 50\%$. And the number of observations for the training data was set to $n = 50, 100, 200$, and applying two kernel functions (Polynomial, Gaussian) when applying the kernel trick method, total of 192 situations were assumed. And in the case of test data, the number of observations was fixed as $u = 1000$. The simulation steps carried out in this study were as follows, and 100 repetitions were performed for each methodology.

1) PKR1, GKR1, PKR2, GKR2

Step1) Create training data and test data for simulation, respectively.

Step2) Select the optimal value of the ridge parameter λ through 5-fold CV (cross-validation).

Step3) Using the optimal value of λ selected in Step2) and training data, calculate the estimator of the regression coefficient vector $\hat{d}$, and based on this, decide final synthetic response estimator $y^{\hat{S}*}$ for test data and calculate test RMSE $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_i^{\hat{S}*})^2}$ or $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_i^{\hat{S}*})^2}$.

2) PKRS1, GKRS1, PKRS2, GKRS2 (Sub-sampling)

Step1) Create training data and test data for simulation, respectively.

Step2) Among the observations in the training data, 70% was randomly selected to set the new training data, and the remaining 30% was used as validation data to obtain the optimal value of λ through 5-fold CV. Then calculate the test RMSE. Repeat this process 50 times to select the new training data that makes the value of the test RMSE the smallest and the optimal value of λ obtained through this training as the final evaluation criteria.

Step3) Using the optimal value of λ and new training data selected in Step2), calculate regression coefficient vector estimator $\hat{d}$, and based on this, decide final sub-sampling synthetic response estimator $y_{ss}^{\hat{S}*}$ of test data, and calculate test RMSE $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_{ss,i}^{\hat{S}*})^2}$ or $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_{ss,i}^{\hat{S}*})^2}$.

3) PKRB1, GKRB1, PKRB2, GKRB2 (Bagging)

Step1) Create training data and test data for simulation, respectively.

Step2) Using the training data, generate 50 bootstrap samples with the same

number of observations with training data through sampling with replacement and set as new training data. In the case of validation data for each bootstrap sample, the observations corresponding to the number excluded during sampling with replacement are collected and set.

Step3) For each bootstrap sample created in Step2), find the value of the regression coefficient vector estimator $\hat{d}$ for each λ value, and calculate the test RMSE using this value.

Step4) Arrange the result of the test RMSE calculated in Step3) according to the value of each λ and averaged to finally determine the value of λ when the result of averaged test RMSE is the smallest. If there are several values of λ corresponding to this, select the largest value among them.

Step5) Based on the value of λ determined in Step4), using test data that is the evaluation target and 100 different bootstrap training data, calculate final bagging synthetic response estimator $y_{bag}^{\hat{S}*}$. Through this process, calculate test RMSE of test data $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_{bag,i}^{\hat{S*}})^2}$ or $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_{bag,i}^{\hat{S*}})^2}$.

4) PKRR1, GKRR1, PKRR2, GKRR2 (Random forests)

Step1) Create training data and test data for simulation, respectively.

Step2) Using the training data, generate 50 bootstrap samples with the same number of observations with training data through sampling with replacement and set as new training data. At this time, for each sample, $m \approx \sqrt{p}$ explanatory variables are randomly extracted and included. The types of explanatory variables are set differently for each sample. In the case of validation for each bootstrap

sample, the observations corresponding to the number excluded during sampling with replacement are collected and set. The types of explanatory variables to be included in these validation data are set to be the same as in the case of the corresponding bootstrap training data.

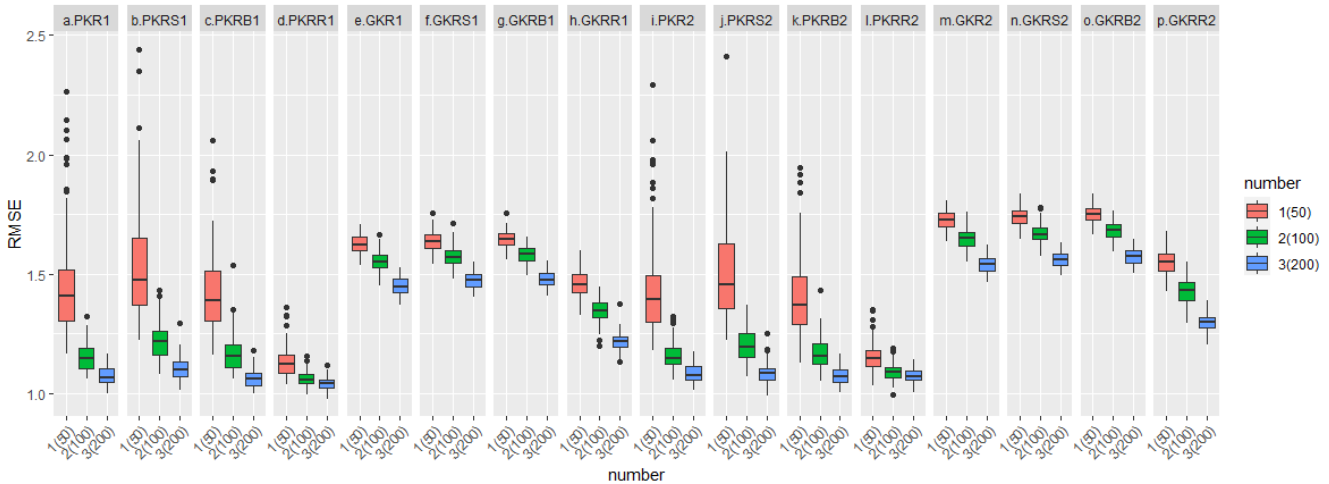
Step3) For each bootstrap sample created in Step2), find the value of the regression coefficient vector estimator $\hat{d}$ for each λ value, and calculate the test RMSE.

Step4) Arrange the result of the test RMSE calculated in Step3) according to the value of each λ and averaged to finally determine the value of λ when the result of averaged test RMSE is the smallest. If there are several values of λ corresponding to this, select the largest value among them.

Step5) Based on the value of λ determined in Step4), using test data that is the evaluation target and 100 different bootstrap training data, calculate final random forests synthetic response estimator $y_{rf}^{\hat{S}*}$. At this time, for each bootstrap training data, $m \approx \sqrt{p}$ explanatory variables are randomly extracted and included. The types of explanatory variables are set differently for each bootstrap training data. And the type of explanatory variable to be included in the test data, which is the evaluation target, should be set according to the standards of the corresponding bootstrap training data. Through this process, calculate test RMSE of test data $\sqrt{MSE1} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{S*} - y_{rf,i}^{\hat{S}*})^2}$ or $\sqrt{MSE2} = \sqrt{\frac{1}{u}\Sigma_{i=1}^{u}(y_i^{*} - y_{rf,i}^{\hat{S}*})^2}$.

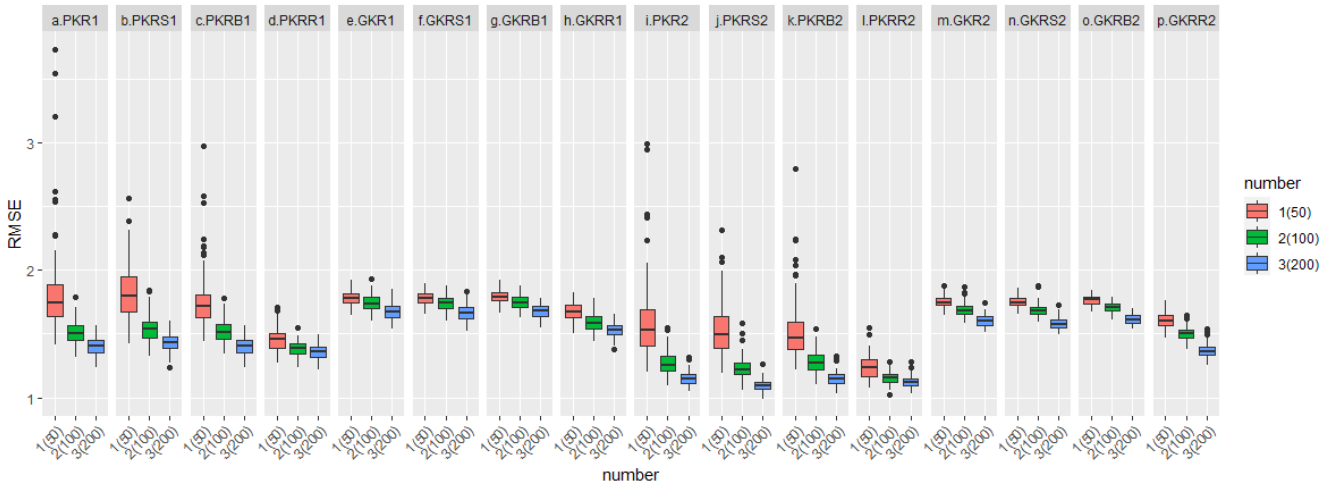
Based on the steps described above, a total of 192 situations were assumed for each methodology and simulation was performed. And the results of the calculated test RMSE are organized in boxplot and table, which can be confirmed in

Figure 6.1 $\sim$ Figure 6.16. Here, the height and length of the boxplot are respectively related to the mean and variance of the test RMSE. Therefore, the lower the boxplot height, the more accurate the predictive power of the corresponding methodology, and the shorter the boxplot length, the better the stability of the corresponding methodology. Based on this fact, it is enough to judge the methodology in which the height and length of the boxplot appear low and short, respectively, as having excellent predictive power. In addition, in the table that related with test RMSE, the results of the method showing the best performance in each simulation situation were displayed in bold letters to make them stand out.



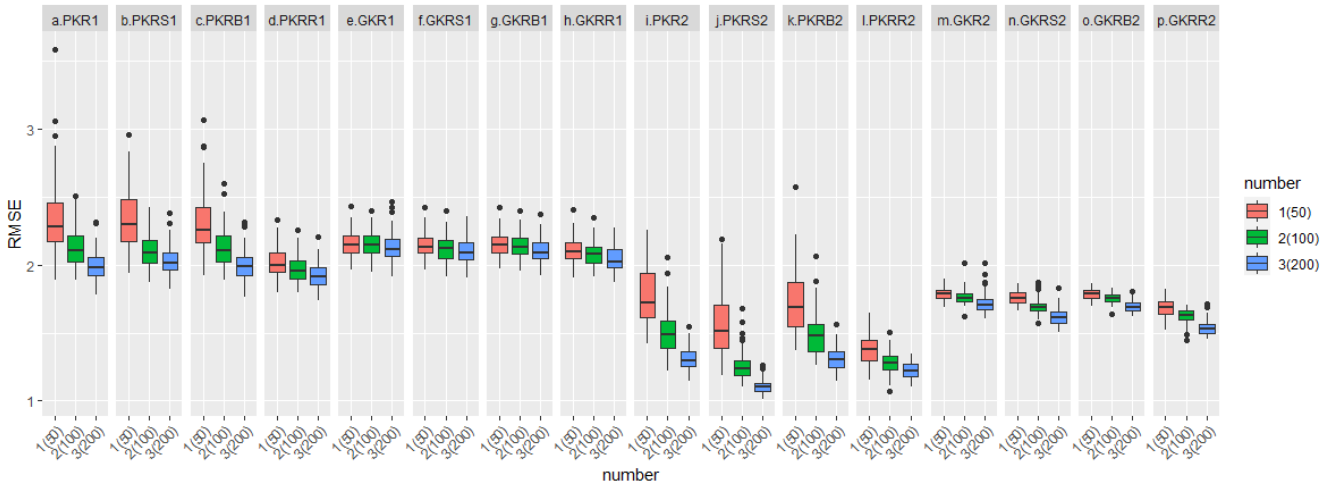
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.453	1.542	1.432	**1.131**	1.627	1.637	1.648	**1.459**	1.442	1.507	1.404	**1.151**	1.727	1.739	1.752	**1.549**
	sd	0.229	0.240	0.170	**0.063**	0.040	0.039	0.035	**0.053**	0.216	0.211	0.162	**0.061**	0.040	0.041	0.036	**0.054**
n=100	mean	1.154	1.223	1.165	**1.060**	1.553	1.573	1.581	**1.346**	1.161	1.204	1.169	**1.090**	1.650	1.667	1.682	**1.429**
	sd	0.058	0.077	0.071	**0.030**	0.042	0.039	0.034	**0.049**	0.058	0.069	0.064	**0.036**	0.043	0.038	0.035	**0.056**
n=200	mean	1.073	1.104	1.062	**1.040**	1.449	1.474	1.477	**1.217**	1.082	1.085	1.073	**1.072**	1.537	1.561	1.573	**1.297**
	sd	0.040	0.046	0.037	**0.027**	0.037	0.033	0.034	**0.037**	0.038	0.042	0.037	**0.030**	0.038	0.034	0.035	**0.036**

Figure 6.1: $p = 3$, censoring 0%



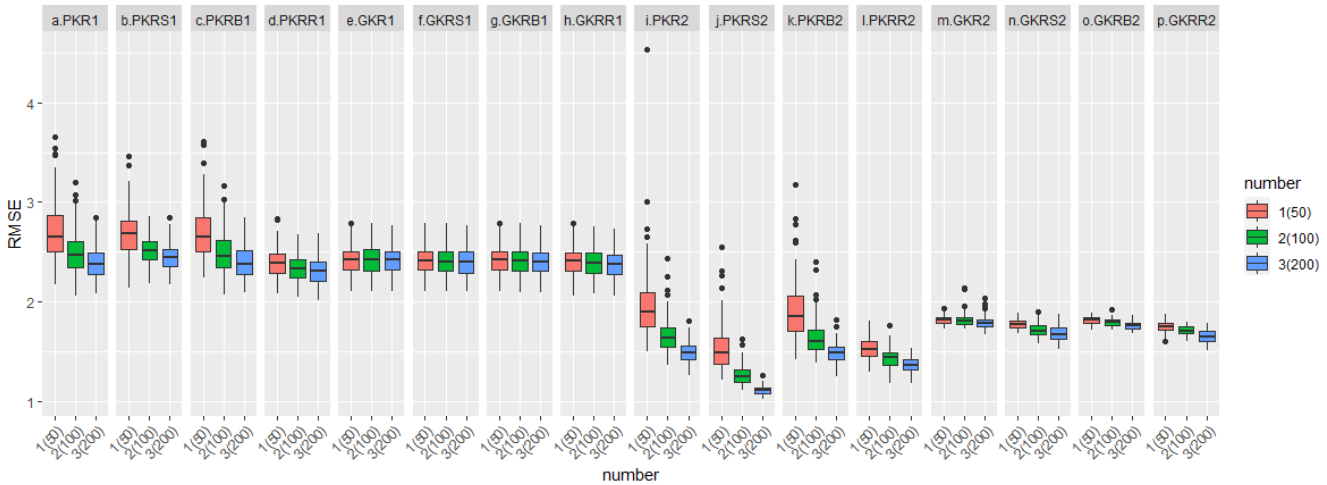
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.832	1.820	1.759	**1.455**	1.782	1.781	1.793	**1.674**	1.610	1.542	1.524	**1.235**	1.749	1.748	1.766	**1.599**
	sd	0.372	0.213	0.236	**0.084**	0.059	0.054	0.056	**0.070**	0.328	0.207	0.237	**0.088**	0.044	0.041	0.038	**0.057**
n=100	mean	1.506	1.547	1.518	**1.381**	1.740	1.736	1.746	**1.593**	1.270	1.224	1.278	**1.155**	1.692	1.685	1.706	**1.497**
	sd	0.088	0.105	0.089	**0.059**	0.064	0.060	0.056	**0.072**	0.095	0.083	0.089	**0.051**	0.051	0.050	0.037	**0.049**
n=200	mean	1.402	1.432	1.399	**1.353**	1.677	1.666	1.679	**1.530**	1.151	**1.094**	1.144	1.119	1.602	1.579	1.615	**1.371**
	sd	0.069	0.065	0.071	**0.057**	0.061	0.064	0.056	**0.059**	0.055	**0.045**	0.055	0.042	0.045	0.044	0.037	**0.056**

Figure 6.2: $p = 3$, censoring 10%



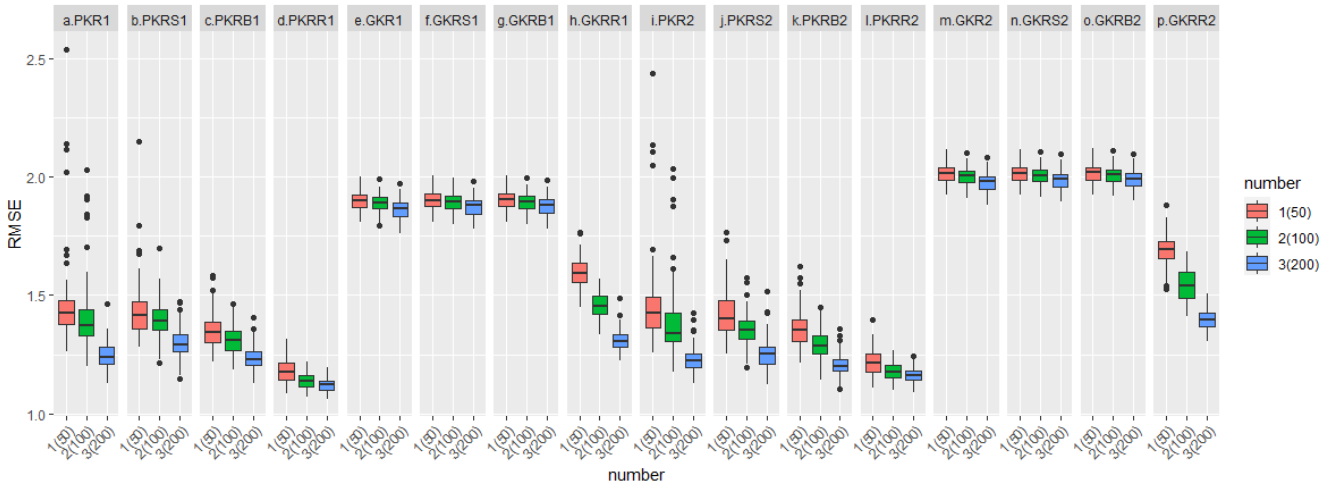
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.345	2.332	2.306	**2.015**	2.156	2.146	2.155	**2.116**	1.770	1.564	1.723	**1.379**	1.784	1.757	1.788	**1.685**
	sd	0.263	0.207	0.206	**0.101**	0.091	0.092	0.090	**0.098**	0.213	0.231	0.226	**0.110**	0.043	0.044	0.039	**0.060**
n=100	mean	2.129	2.112	2.132	**1.962**	2.156	2.120	2.142	**2.082**	1.500	**1.253**	1.483	1.278	1.763	1.697	1.753	**1.621**
	sd	0.139	0.114	0.136	**0.092**	0.093	0.095	0.088	**0.090**	0.151	**0.096**	0.140	0.077	0.051	0.056	0.035	**0.053**
n=200	mean	1.996	2.032	1.999	**1.925**	2.137	2.100	2.112	**2.044**	1.307	**1.107**	1.310	1.227	1.717	1.619	1.695	**1.532**
	sd	0.109	0.101	0.111	**0.093**	0.108	0.100	0.090	**0.094**	0.078	**0.053**	0.077	0.057	0.069	0.061	0.040	**0.049**

Figure 6.3: $p = 3$, censoring 30%



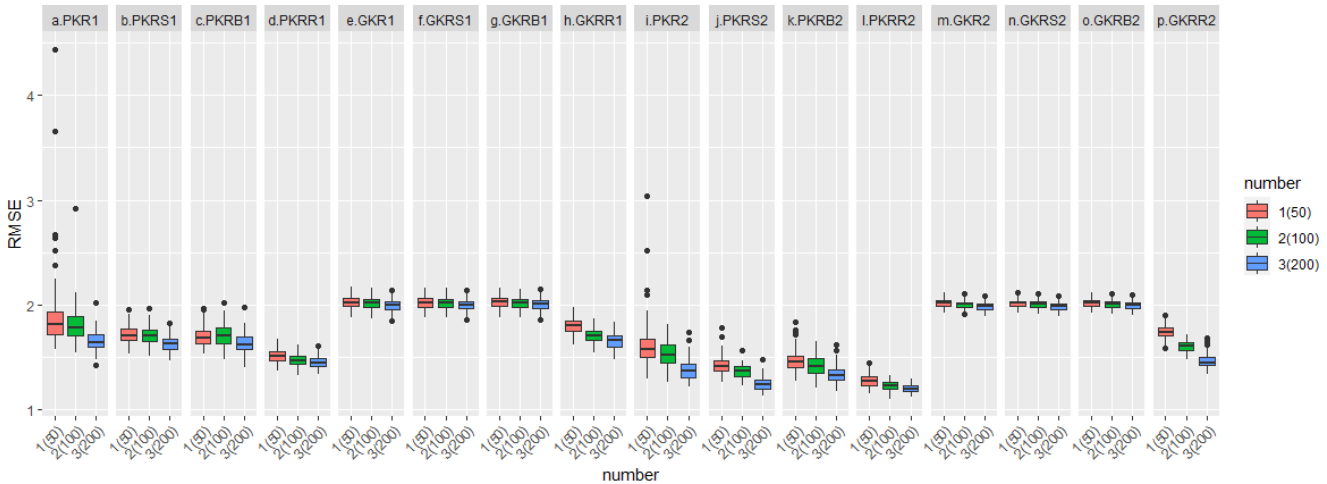
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.710	2.701	2.690	**2.384**	2.422	2.415	2.418	**2.408**	1.987	1.543	1.922	**1.524**	1.816	1.776	1.810	**1.748**
	sd	0.295	0.242	0.279	**0.156**	0.150	0.149	0.151	**0.151**	0.391	0.238	0.307	**0.112**	0.040	0.048	0.037	**0.049**
n=100	mean	2.491	2.521	2.488	**2.336**	2.426	2.405	2.415	**2.396**	1.667	**1.260**	1.646	1.430	1.815	1.716	1.794	**1.708**
	sd	0.218	0.147	0.209	**0.149**	0.155	0.153	0.153	**0.151**	0.189	**0.098**	0.182	0.096	0.064	0.063	0.038	**0.046**
n=200	mean	2.393	2.451	2.395	**2.312**	2.422	2.399	2.406	**2.379**	1.492	**1.110**	1.493	1.366	1.791	1.680	1.760	**1.649**
	sd	0.168	0.139	0.170	**0.152**	0.151	0.148	0.150	**0.149**	0.098	**0.045**	0.097	0.074	0.062	0.080	0.040	**0.064**

Figure 6.4: $p = 3$, censoring 50%



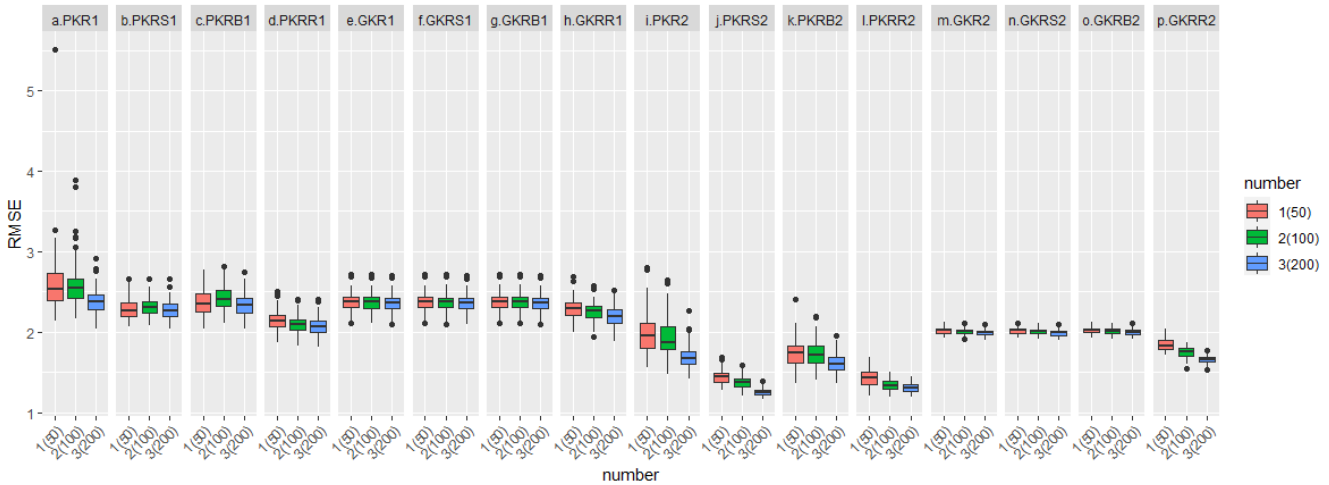
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.460	1.438	1.350	**1.178**	1.897	1.899	1.901	**1.595**	1.458	1.419	1.359	**1.216**	2.011	2.013	2.015	**1.690**
	sd	0.176	0.113	0.067	**0.050**	0.038	0.038	0.038	**0.068**	0.173	0.096	0.075	**0.055**	0.037	0.038	0.037	**0.068**
n=100	mean	1.403	1.398	1.312	**1.138**	1.887	1.891	1.893	**1.457**	1.381	1.357	1.290	**1.178**	2.001	2.005	2.007	**1.543**
	sd	0.145	0.071	0.059	**0.033**	0.037	0.038	0.037	**0.057**	0.150	0.070	0.061	**0.038**	0.037	0.038	0.037	**0.068**
n=200	mean	1.245	1.302	1.234	**1.123**	1.863	1.872	1.877	**1.309**	1.232	1.252	1.207	**1.165**	1.976	1.986	1.991	**1.398**
	sd	0.055	0.062	0.048	**0.028**	0.039	0.038	0.038	**0.044**	0.056	0.067	0.047	**0.030**	0.038	0.038	0.038	**0.043**

Figure 6.5: $p = 5$, censoring 0%



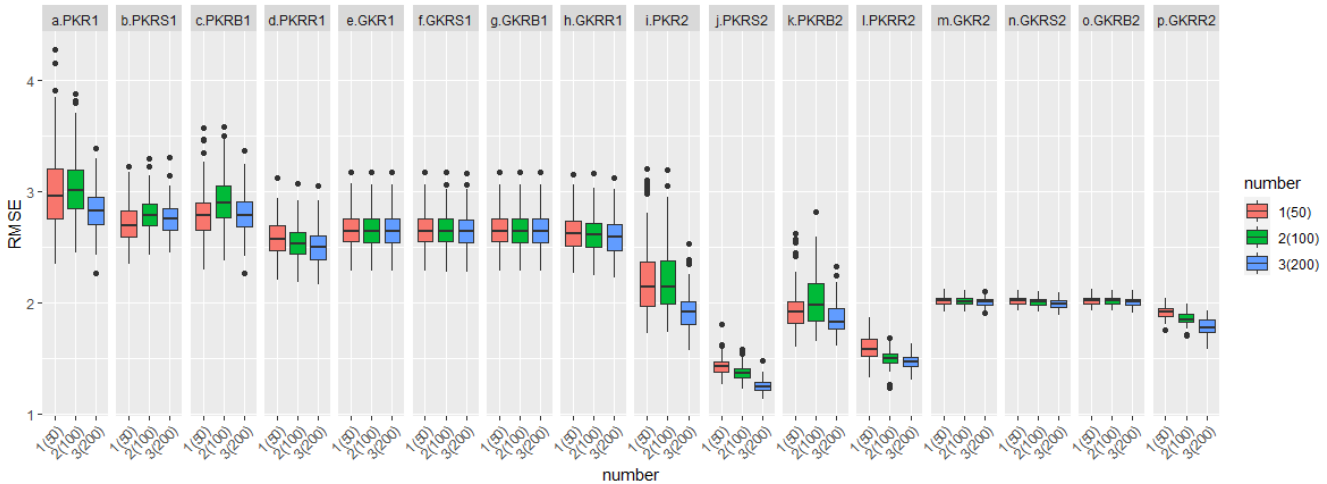
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.897	1.717	1.700	**1.518**	2.022	2.023	2.024	**1.808**	1.629	1.423	1.466	**1.280**	2.013	2.013	2.015	**1.738**
	sd	0.371	0.088	0.096	**0.065**	0.056	0.055	0.056	**0.071**	0.229	0.088	0.103	**0.063**	0.038	0.037	0.037	**0.063**
n=100	mean	1.802	1.711	1.704	**1.471**	2.014	2.015	2.018	**1.706**	1.530	1.365	1.421	**1.229**	2.003	2.005	2.008	**1.611**
	sd	0.167	0.085	0.105	**0.057**	0.056	0.056	0.056	**0.070**	0.125	0.064	0.097	**0.048**	0.038	0.037	0.037	**0.050**
n=200	mean	1.659	1.634	1.633	**1.454**	1.996	2.000	2.006	**1.654**	1.385	1.250	1.337	**1.209**	1.983	1.986	1.994	**1.470**
	sd	0.093	0.070	0.088	**0.056**	0.057	0.056	0.056	**0.076**	0.097	0.063	0.079	**0.037**	0.038	0.038	0.038	**0.074**

Figure 6.6: $p = 5$, censoring 10%



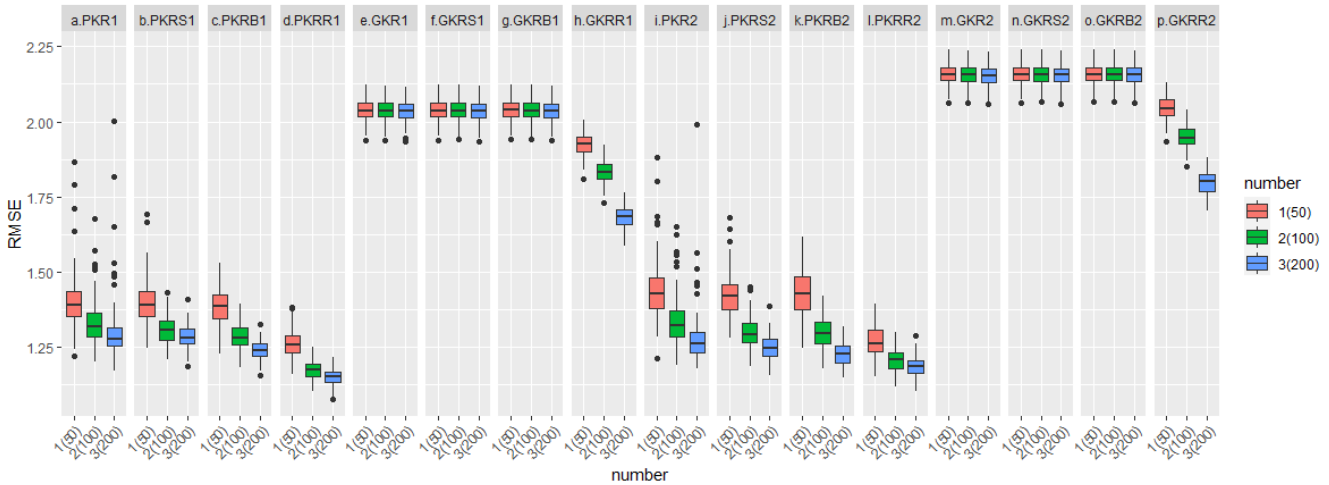
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.592	2.278	2.366	**2.148**	2.373	2.372	2.374	**2.296**	1.974	1.436	1.734	**1.427**	2.016	2.014	2.018	**1.838**
	sd	0.379	0.119	0.161	**0.125**	0.107	0.107	0.107	**0.117**	0.254	0.086	0.174	**0.109**	0.038	0.038	0.038	**0.066**
n=100	mean	2.599	2.316	2.429	**2.093**	2.369	2.368	2.371	**2.253**	1.932	1.372	1.738	**1.341**	2.009	2.006	2.012	**1.747**
	sd	0.282	0.117	0.152	**0.110**	0.107	0.107	0.107	**0.115**	0.228	0.075	0.162	**0.068**	0.038	0.038	0.037	**0.066**
n=200	mean	2.379	2.269	2.342	**2.071**	2.362	2.360	2.364	**2.193**	1.687	**1.254**	1.616	1.311	1.993	1.988	2.001	**1.654**
	sd	0.150	0.118	0.135	**0.112**	0.108	0.108	0.107	**0.121**	0.139	**0.048**	0.118	0.054	0.038	0.038	0.038	**0.047**

Figure 6.7: $p = 5$, censoring 30%



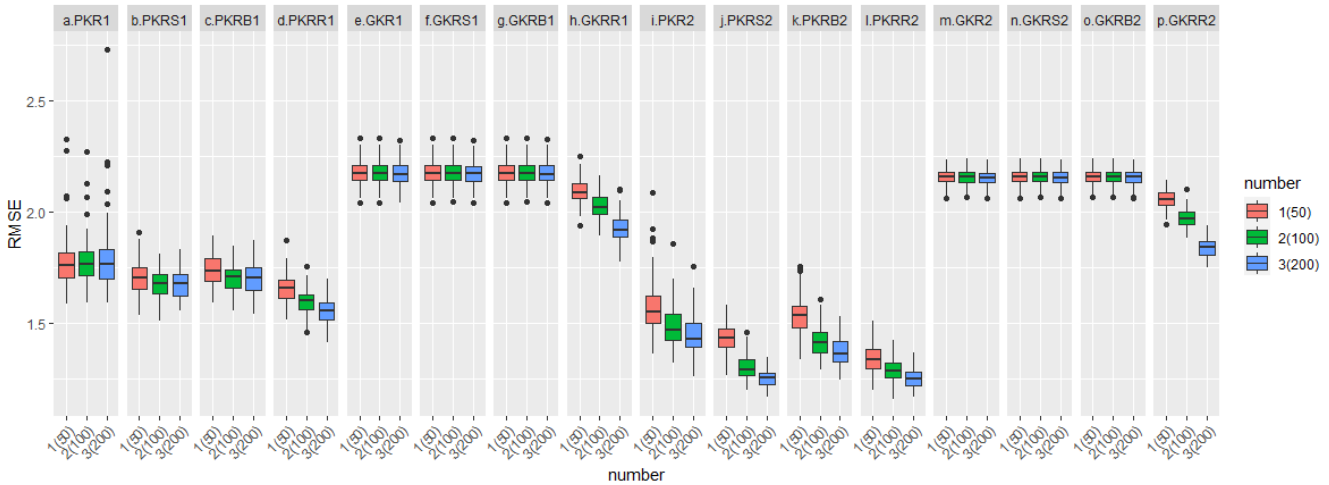
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	3.016	2.712	2.809	**2.572**	2.651	2.650	2.651	**2.630**	2.218	**1.437**	1.963	1.585	2.018	2.015	2.019	**1.913**
	sd	0.362	0.170	0.241	**0.164**	0.168	0.168	0.168	**0.169**	0.341	**0.087**	0.216	0.120	0.038	0.038	0.038	**0.052**
n=100	mean	3.043	2.779	2.906	**2.531**	2.651	2.649	2.651	**2.616**	2.209	**1.375**	2.015	1.506	2.014	2.008	2.016	**1.860**
	sd	0.296	0.169	0.250	**0.163**	0.168	0.169	0.168	**0.168**	0.309	**0.074**	0.232	0.080	0.037	0.038	0.037	**0.055**
n=200	mean	2.825	2.757	2.794	**2.511**	2.648	2.643	2.648	**2.595**	1.939	**1.256**	1.867	1.468	2.006	1.990	2.009	**1.788**
	sd	0.204	0.150	0.190	**0.166**	0.168	0.168	0.168	**0.170**	0.176	**0.059**	0.147	0.074	0.039	0.038	0.039	**0.077**

Figure 6.8: $p = 5$, censoring 50%



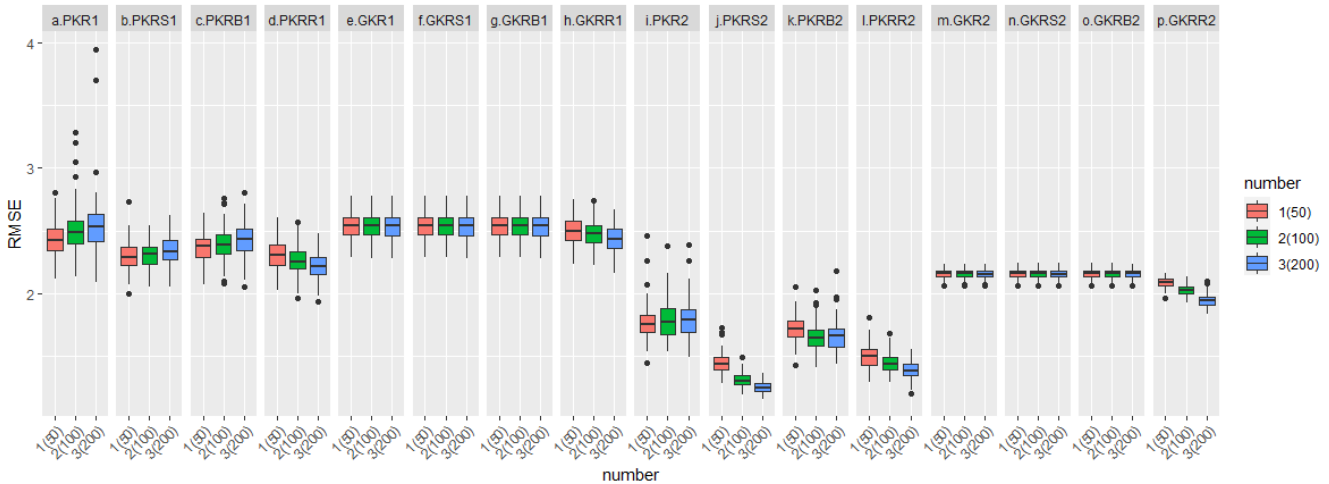
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.404	1.401	1.387	**1.263**	2.039	2.039	2.039	**1.926**	1.443	1.426	1.430	**1.268**	2.159	2.159	2.159	**2.045**
	sd	0.096	0.074	0.059	**0.047**	0.036	0.036	0.036	**0.037**	0.100	0.074	0.079	**0.051**	0.036	0.036	0.036	**0.038**
n=100	mean	1.332	1.308	1.285	**1.176**	2.037	2.038	2.038	**1.831**	1.342	1.300	1.299	**1.207**	2.157	2.158	2.158	**1.949**
	sd	0.077	0.044	0.045	**0.033**	0.036	0.036	0.036	**0.038**	0.088	0.051	0.050	**0.038**	0.036	0.036	0.036	**0.039**
n=200	mean	1.304	1.285	1.242	**1.152**	2.034	2.036	2.036	**1.683**	1.280	1.248	1.227	**1.187**	2.154	2.155	2.156	**1.796**
	sd	0.113	0.040	0.033	**0.027**	0.036	0.036	0.036	**0.037**	0.098	0.040	0.036	**0.034**	0.036	0.037	0.036	**0.038**

Figure 6.9: $p = 7$, censoring 0%



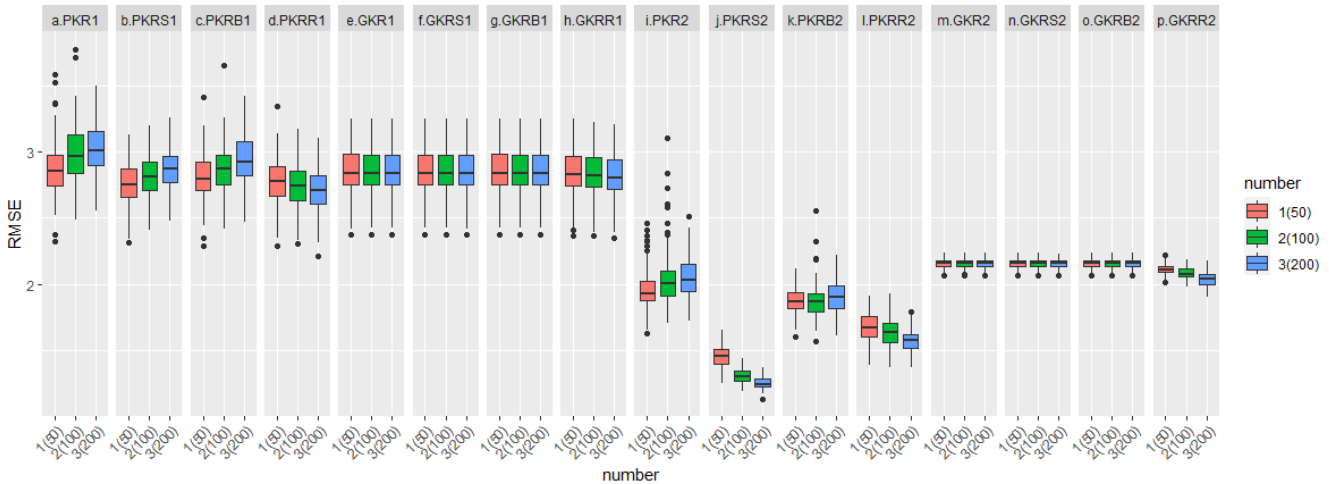
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.783	1.707	1.747	**1.657**	2.178	2.178	2.178	**2.094**	1.568	1.436	1.534	**1.342**	2.159	2.159	2.159	**2.058**
	sd	0.113	0.071	0.072	**0.069**	0.055	0.055	0.055	**0.056**	0.123	0.067	0.085	**0.072**	0.036	0.036	0.036	**0.039**
n=100	mean	1.776	1.674	1.700	**1.594**	2.177	2.177	2.178	**2.025**	1.487	1.302	1.416	**1.287**	2.158	2.158	2.158	**1.973**
	sd	0.104	0.061	0.064	**0.059**	0.055	0.055	0.055	**0.057**	0.089	0.054	0.065	**0.053**	0.036	0.036	0.036	**0.040**
n=200	mean	1.785	1.676	1.700	**1.555**	2.175	2.175	2.176	**1.927**	1.451	**1.252**	1.374	1.253	2.155	2.156	2.157	**1.842**
	sd	0.148	0.064	0.073	**0.055**	0.055	0.054	0.055	**0.060**	0.086	**0.039**	0.063	0.045	0.036	0.036	0.036	**0.042**

Figure 6.10: $p = 7$, censoring 10%



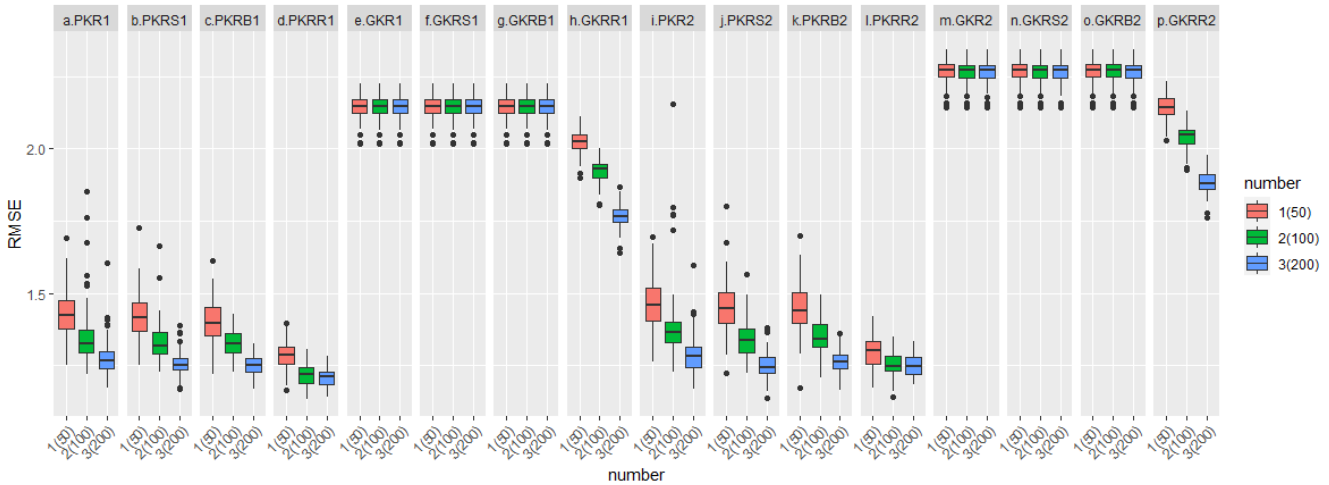
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.429	**2.296**	2.371	2.313	2.536	2.536	2.536	**2.500**	1.771	**1.446**	1.716	1.502	2.159	2.159	2.159	**2.086**
	sd	0.134	**0.119**	0.116	0.115	0.103	0.103	0.103	**0.105**	0.139	**0.082**	0.103	0.106	0.036	0.036	0.036	**0.041**
n=100	mean	2.506	2.310	2.393	**2.259**	2.535	2.535	2.535	**2.474**	1.787	**1.312**	1.656	1.446	2.158	2.158	2.159	**2.027**
	sd	0.188	0.108	0.130	**0.114**	0.103	0.103	0.103	**0.106**	0.150	**0.056**	0.110	0.081	0.036	0.036	0.036	**0.043**
n=200	mean	2.546	2.342	2.431	**2.216**	2.534	2.534	2.535	**2.435**	1.792	**1.253**	1.664	1.391	2.156	2.156	2.157	**1.945**
	sd	0.246	0.116	0.142	**0.112**	0.104	0.104	0.103	**0.106**	0.155	**0.043**	0.123	0.066	0.037	0.036	0.036	**0.050**

Figure 6.11: $p = 7$, censoring 30%



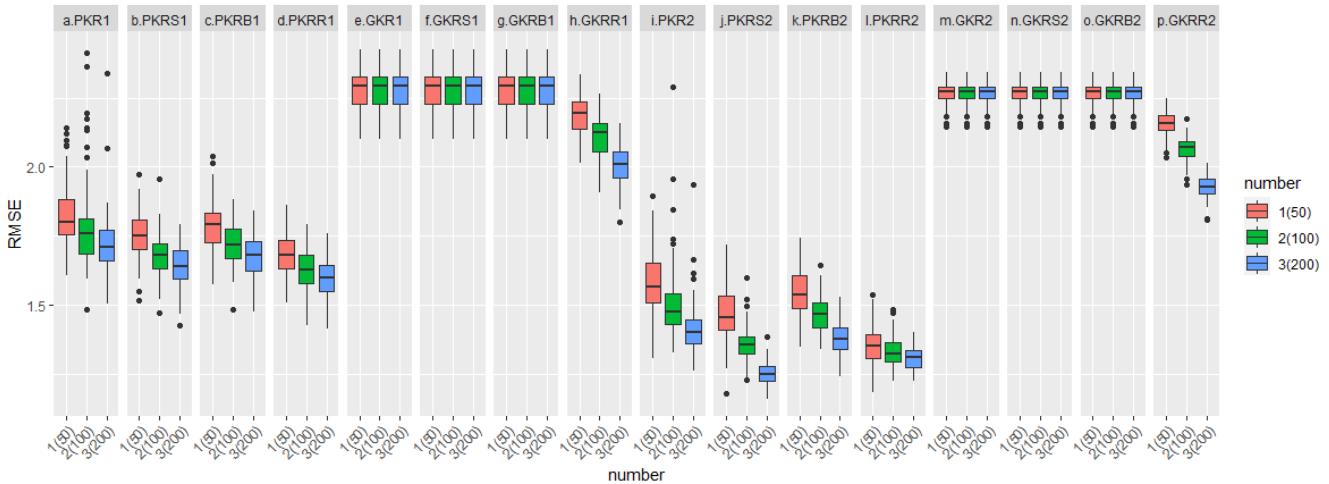
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.884	**2.757**	2.819	2.788	2.851	2.851	2.851	**2.840**	1.959	**1.454**	1.873	1.679	2.159	2.159	2.159	**2.112**
	sd	0.217	**0.172**	0.193	0.179	0.179	0.179	0.179	**0.180**	0.156	**0.074**	0.098	0.117	0.036	0.036	0.036	**0.039**
n=100	mean	2.992	2.814	2.880	**2.747**	2.851	2.851	2.851	**2.833**	2.040	**1.306**	1.876	1.636	2.159	2.158	2.159	**2.084**
	sd	0.226	0.165	0.195	**0.174**	0.179	0.179	0.179	**0.178**	0.228	**0.055**	0.141	0.106	0.036	0.036	0.036	**0.040**
n=200	mean	3.033	2.868	2.941	**2.710**	2.850	2.850	2.850	**2.819**	2.050	**1.251**	1.906	1.576	2.157	2.156	2.158	**2.036**
	sd	0.201	0.168	0.185	**0.177**	0.180	0.179	0.179	**0.179**	0.155	**0.042**	0.127	0.086	0.036	0.036	0.036	**0.052**

Figure 6.12: $p = 7$, censoring 50%



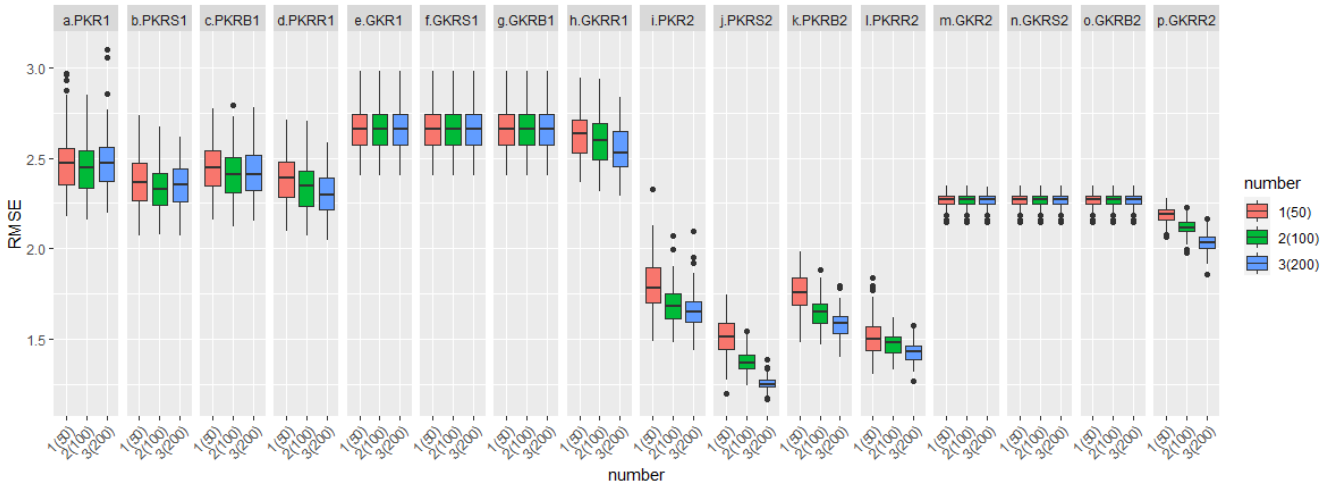
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.428	1.419	1.403	**1.285**	2.145	2.145	2.146	**2.022**	1.460	1.450	1.448	**1.294**	2.267	2.267	2.267	**2.144**
	sd	0.079	0.075	0.070	**0.051**	0.040	0.040	0.040	**0.041**	0.085	0.089	0.086	**0.054**	0.040	0.040	0.040	**0.041**
n=100	mean	1.348	1.331	1.326	**1.217**	2.145	2.145	2.145	**1.923**	1.384	1.339	1.352	**1.254**	2.267	2.267	2.267	**2.042**
	sd	0.097	0.064	0.046	**0.038**	0.040	0.040	0.040	**0.039**	0.125	0.059	0.061	**0.043**	0.040	0.040	0.040	**0.039**
n=200	mean	1.275	1.255	1.249	**1.206**	2.145	2.145	2.145	**1.768**	1.287	1.251	1.263	**1.248**	2.267	2.267	2.267	**1.883**
	sd	0.060	0.038	0.035	**0.032**	0.040	0.040	0.040	**0.039**	0.063	0.045	0.040	**0.035**	0.040	0.040	0.040	**0.039**

Figure 6.13: $p = 9$, censoring 0%



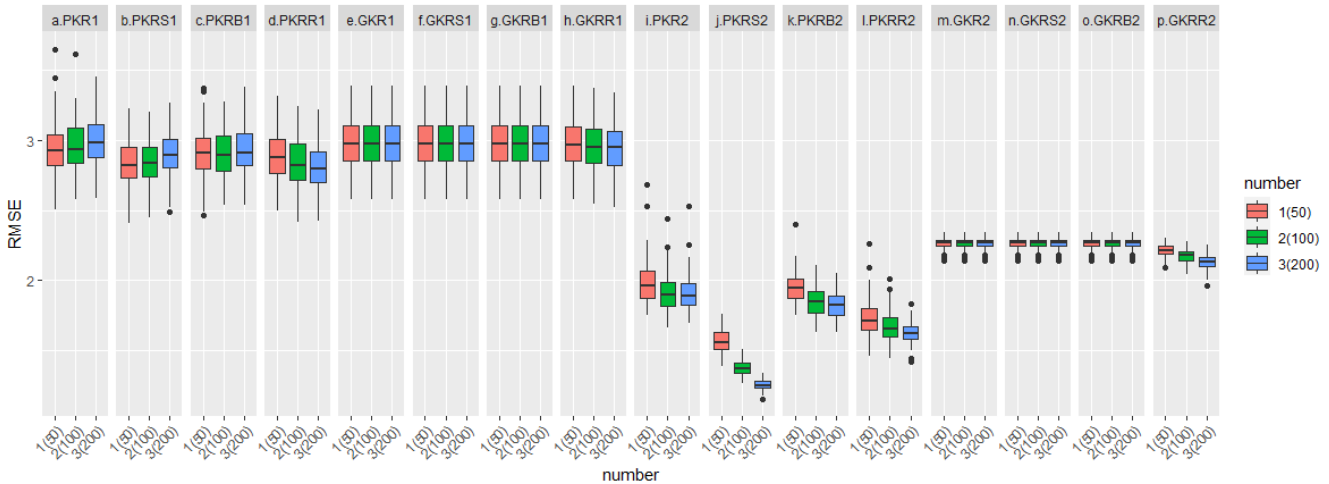
		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	1.821	1.751	1.783	**1.681**	2.278	2.278	2.278	**2.185**	1.582	1.469	1.542	**1.355**	2.267	2.267	2.267	**2.156**
	sd	0.108	0.088	0.090	**0.077**	0.065	0.065	0.065	**0.068**	0.112	0.106	0.089	**0.072**	0.040	0.040	0.040	**0.041**
n=100	mean	1.782	1.677	1.718	**1.625**	2.278	2.278	2.278	**2.112**	1.505	1.360	1.470	**1.330**	2.267	2.267	2.267	**2.066**
	sd	0.153	0.074	0.076	**0.075**	0.065	0.065	0.065	**0.069**	0.127	0.058	0.064	**0.052**	0.040	0.040	0.040	**0.040**
n=200	mean	1.717	1.643	1.678	**1.598**	2.278	2.278	2.278	**2.005**	1.415	**1.253**	1.375	1.307	2.267	2.267	2.267	**1.929**
	sd	0.103	0.073	0.073	**0.069**	0.065	0.065	0.065	**0.070**	0.087	**0.041**	0.054	0.039	0.040	0.040	0.040	**0.040**

Figure 6.14: $p = 9$, censoring 10%



		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.469	**2.362**	2.437	2.379	2.663	2.663	2.663	**2.623**	1.809	**1.513**	1.760	1.517	2.267	2.267	2.267	**2.185**
	sd	0.168	**0.132**	0.130	0.130	0.123	0.123	0.123	**0.124**	0.142	**0.111**	0.106	0.106	0.040	0.040	0.040	**0.043**
n=100	mean	2.439	2.335	2.405	**2.332**	2.663	2.663	2.663	**2.590**	1.686	**1.375**	1.650	1.474	2.267	2.267	2.267	**2.117**
	sd	0.142	0.128	0.135	**0.131**	0.123	0.122	0.123	**0.129**	0.105	**0.063**	0.081	0.067	0.040	0.040	0.040	**0.047**
n=200	mean	2.476	2.351	2.410	**2.301**	2.663	2.663	2.663	**2.548**	1.659	**1.253**	1.581	1.430	2.267	2.267	2.267	**2.031**
	sd	0.158	0.120	0.126	**0.123**	0.123	0.123	0.123	**0.126**	0.098	**0.038**	0.070	0.057	0.040	0.040	0.040	**0.049**

Figure 6.15: $p = 9$, censoring 30%



		PKR 1	PKRS 1	PKRB 1	PKRR 1	GKR 1	GKRS 1	GKRB 1	GKRR 1	PKR 2	PKRS 2	PKRB 2	PKRR 2	GKR 2	GKRS 2	GKRB 2	GKRR 2
n=50	mean	2.936	**2.835**	2.907	2.890	2.980	2.980	2.980	**2.970**	1.985	**1.571**	1.948	1.729	2.267	2.267	2.267	**2.216**
	sd	0.202	**0.180**	0.191	0.188	0.178	0.179	0.179	**0.178**	0.147	**0.083**	0.103	0.129	0.040	0.040	0.040	**0.042**
n=100	mean	2.956	2.852	2.906	**2.842**	2.980	2.980	2.980	**2.959**	1.907	**1.377**	1.850	1.664	2.267	2.267	2.267	**2.175**
	sd	0.180	0.171	0.172	**0.177**	0.179	0.178	0.178	**0.179**	0.133	**0.054**	0.107	0.106	0.040	0.040	0.040	**0.049**
n=200	mean	2.996	2.902	2.939	**2.816**	2.980	2.980	2.980	**2.946**	1.910	**1.254**	1.821	1.628	2.267	2.267	2.267	**2.131**
	sd	0.185	0.172	0.177	**0.177**	0.179	0.179	0.179	**0.177**	0.125	**0.039**	0.088	0.072	0.040	0.040	0.040	**0.051**

Figure 6.16: $p = 9$, censoring 50%

As can be seen from Figure 6.1 $\sim$ Figure 6.16 and the corresponding table about test RMSE, a total of 192 situations were assumed and simulation was performed. As a result, when using kernel ridge censored regression analysis with random forests technique, you can see that predictive power gets better. Of course, in a situation where the number of explanatory variables is large and the censoring ratio is large, there may be cases in which the performance improvement by the random forests method does not occur because the distribution variability increases. However, if we look at the whole, we can see that predictive power is definitely improved when ensemble methods such as bagging or random forests are applied, despite assuming various simulation situations. Accordingly, through this simulation, it was possible to prove that the kernel ridge censored regression analysis method to which the ensemble method is applied has superior predictive power compared to other methodologies.

6.2 Real data analysis

In real data analysis, the 5 real data containing censoring was randomly divided in a ratio of train:test=7:3, and then the analysis was performed to compare the predictive power of the methodologies through the steps introduced in Section 6.1. However, as mentioned in Section 4 in Chapter 3, the analysis was conducted only on the methodology of setting the evaluation criterion as a synthetic response, considering that it is not possible to know the time taken until the event of interest occurs for all observations in the real data including censoring. And, as can be seen from the simulation results in Section 6.1, when the Polynomial kernel is used when the kernel trick method is applied, there is a disadvantage in that the stability of the analysis is poor because the variability is very large and outliers occur frequently. Therefore, when analyzing real data using the Polynomial kernel, it appears often the determinant value that comes out when calculating the inverse matrix is not 0, but it is a positive real number very close to 0. Accordingly, in real data analysis, only Gaussian kernel that showing relatively good flexibility and stability for all data was analyzed. According to these causes, in real data analysis, among 16 methodologies, four methodologies were compared and analyzed: GKR1, GKRS1, GKRB1, and GKRR1. And 100 iterations were performed for each methodology. All data used in this real data analysis are embedded in survival analysis-related packages survival, KMsurv, survMisc, and survMiner of the program R 4.1.1 version.

1) UIS Data

This data is the result of the AIDS Research Unit (UMARU) IMPACT Study (UIS) conducted as a cooperative study for about 5 years from 1989 to 1994 at the University of Massachusetts, USA. The main purpose is to investigate drug abuse in patients, and the predictor variable is the number of days it takes the patient to return to the drug after being tempted. In this study, two different treatment programs were conducted at site A and site B, respectively, and real data analysis was performed only for the case of site A. The number of observations is 398 excluding missing values and outliers, and a total of 8 explanatory variables are used. And the rate of censoring is about 20%. For a detailed description of this data, see Hosmer et al. (2008). As a result of real data analysis, it was

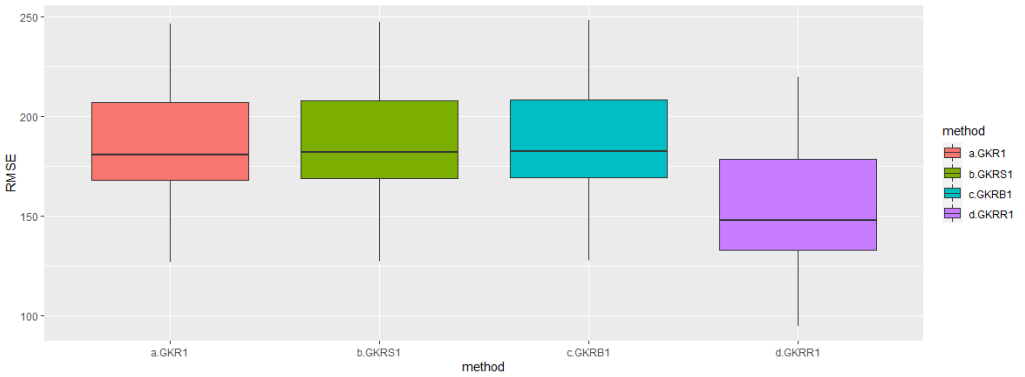


Figure 6.17: Real data analysis : UIS data

confirmed that the average value of the test RMSE was calculated to be smaller when random forests was applied compared to other methodologies. Based on this, it can be judged that the kernel ridge censored regression analysis using ensemble method has good predictive power when compared with other methodologies.

2) PBC Data

This data is the result of a study on patients with primary biliary cholangitis that was conducted for about 10 years from 1974 to 1984 at the Mayo Clinic in the United States. The main objective is to compare the efficacy of placebo and D-penicillamine. The predictor variable is the patient's survival time, and the number of observations is 276, excluding missing values. And there are a total of 17 explanatory variables used, and the censored rate is about 50%. For a detailed description of these data, see Therneau, T. and Grambsch, P. (2000). As a result

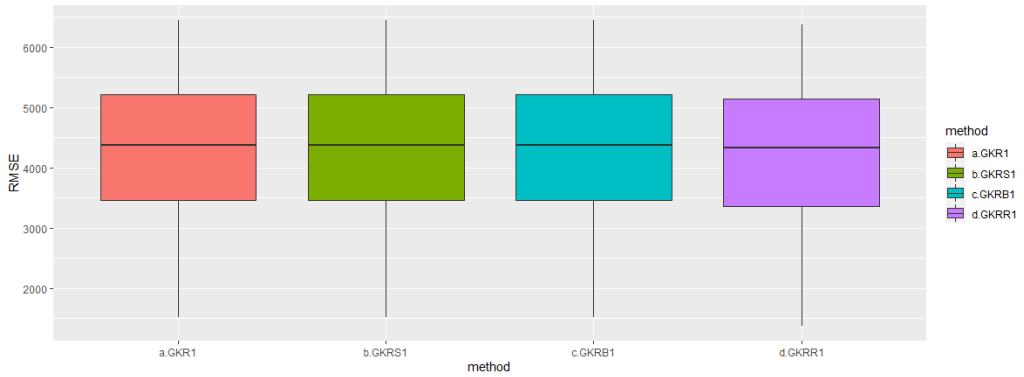


Figure 6.18: Real data analysis : PBC data

of real data analysis, there was no significant improvement in performance when random forests were applied. The number of explanatory variables is 17, and the censoring ratio is also very large, about 50%, so it is judged to be the result of increasing complexity. However, despite these conditions, when kernel ridge censored regression analysis using random forests technique was applied, the average value of test RMSE was calculated to be smaller than that of other methodologies.

Therefore, I think that there is no difficulty in substantiating what we want to claim in this study.

3) Cancer Data

This data is the result of a study on lung cancer patients conducted by the North Central Cancer Treatment Group, a network of cancer experts in North America. The main purpose is to predict the survival time of lung cancer patients, and the number of observations is 167 excluding missing values. And there are a total of 7 explanatory variables used, and the censored rate is about 30%. As a result of

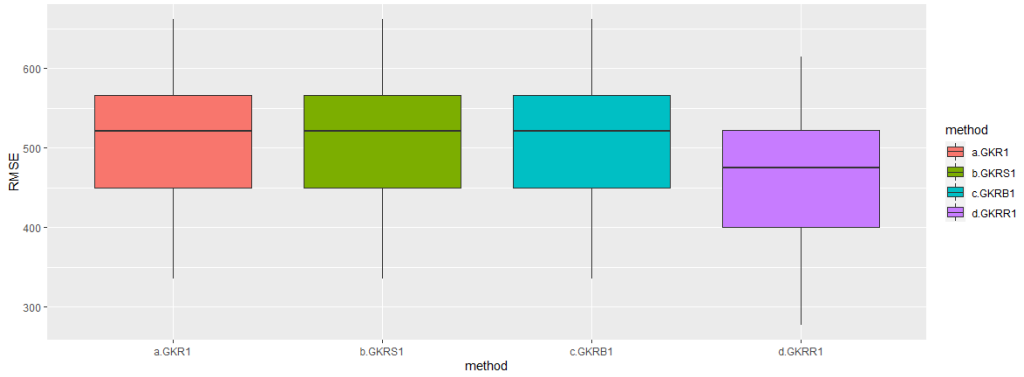


Figure 6.19: Real data analysis : Cancer data

real data analysis, it was confirmed that the average value of the test RMSE was calculated to be smaller when random forests were applied compared to other methodologies as in the case of UIS data. Based on this, it can be judged that the kernel ridge censored regression analysis using ensemble method has good predictive power when compared with other methodologies.

4) Retinopathy Data

This data is the result of a study verifying the effect of laser coagulation as a treatment method to delay diabetic retinopathy. The main purpose is to predict the time it will take to lose sight, and the total number of observations is 394. And there are a total of 6 explanatory variables used, and the censoring rate is about 60%. As a result of real data analysis, there was no significant improvement

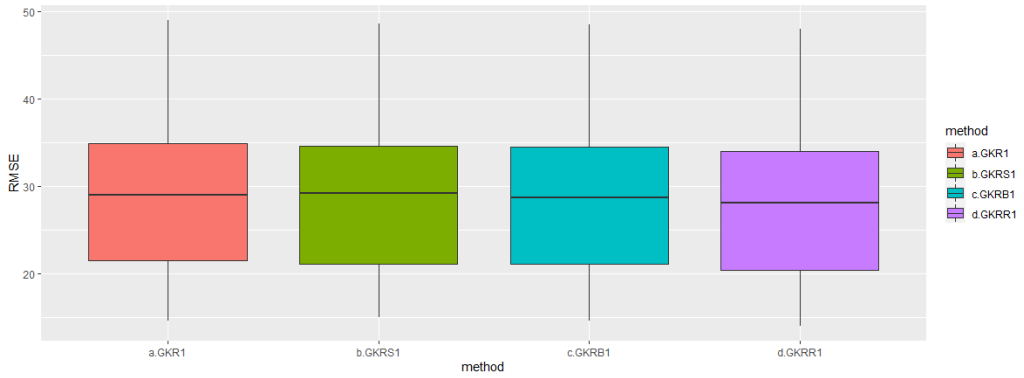


Figure 6.20: Real data analysis : Cancer data

in performance in the case of applying random forests as in the case of PBC data. This is considered to be the result of increasing complexity because the censoring rate is about 60%, which is very large. However, despite these conditions, when kernel ridge censored regression analysis using random forests was applied, the average value of the test RMSE was calculated to be smaller than that of other methodologies. Therefore, I think that there is no difficulty in substantiating what we want to claim in this study.

5) Bfeed Data

These data are the results of a study on the duration of breast feeding of mothers who gave birth to fetuses. The main purpose is to predict the duration of breastfeeding, and the total number of observations is 927. There are a total of 8 explanatory variables used, and the censoring rate is about 4%. For details on these data, see Klein and Moeschberger (1997). As a result of real data analy-

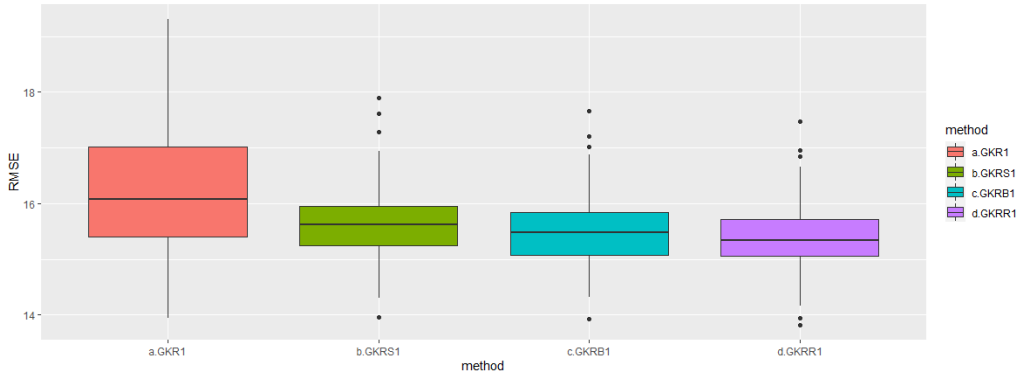


Figure 6.21: Real data analysis : Bfeed data

sis, it was confirmed that the mean value and variance for the test RMSE were smaller when an ensemble method such as bagging or random forests was applied. Of course, since the censoring ratio is about 4%, which is not that large, it can be seen that the performance improvement of the method applying the ensemble method has occurred a lot compared to the data with a large censoring ratio. However, as a result, it is judged that it is sufficient to prove that the kernel ridge censored regression analysis using ensemble method has superior predictive power compared to other methodologies.

6) Summary of real data analysis results

The results of the test RMSE calculated from the real data analysis performed so far can be summarized as shown in Table 6.1.

data		GKR1	GKRS1	GKRB1	GKRR1	Censoring rate	Number of explanatory variables	Number of observations
UIS	**mean**	186.132	187.691	187.857	**154.319**	0.2	8	398
	sd	26.033	25.899	25.836	**28.474**			
PBC	**mean**	4300.305	4300.305	4300.305	**4236.192**	0.5	17	276
	sd	1135.303	1135.303	1135.303	**1145.583**			
Cancer	**mean**	509.320	509.327	509.331	**465.783**	0.3	7	167
	sd	72.125	72.137	72.138	**74.089**			
Retinopathy	**mean**	29.230	29.117	28.859	**28.145**	0.6	6	394
	sd	8.663	8.757	8.704	**8.788**			
Bfeed	**mean**	16.192	15.639	15.502	**15.366**	0.04	8	927
	sd	1.069	0.696	0.677	**0.664**			

Table 6.1: Summary of real data analysis results

As a result, it is clear that there is a difference in the degree of performance improvement because the rate of censoring and the number of explanatory variables used are different for each data used in real data analysis. However, in the case of an overall analysis, it was confirmed that the predictive power of the time it takes for the event of interest to occur is better when the random forests method is applied. Accordingly, through this real data analysis, it was possible to prove that the kernel ridge censored regression analysis using ensemble method has superior predictive power overall when compared to other methodologies.

This concludes the explanation of the research conducted in relation to kernel ridge censored regression analysis. In Chapter 7 that follows, I will explain research conducted on ways to improve the time-dependent Area Under Curve (AUC) that can be obtained through analysis of survival data.

7 Time-dependent AUC

7.1 ROC curve

ROC (Receiver Operating Characteristic) curve is a graph drawn by measuring the performance of the model for various thresholds, and is mainly used when evaluating the performance of a classification model, such as Logistic regression model, Tree based model, and Support Vector Machine(SVM) etc. In order to understand the ROC curve, it is first to understand the confusion matrix, which summarizes the results of classification and shows it in the form of a table, which is shown in Table 7.1. Through this confusion matrix, various indicators

		Predicted values	
	Total population $= P + N$	Positive(PP)	Negative(PN)
Actual values	Positive(P)	True positive (TP)	False negative (FN)
	Negative(N)	False positive (FP)	True negative (TN)

Table 7.1: Confusion matrix

to evaluate the performance of the classification model can be calculated, and representatively, sensitivity: true positive rate (TPR), specificity: true negative rate (TNR), accuracy(ACC), precision(PRE), and misclassification rate(MCR). For more details on these, see Fawcett, T. (2006). In here, sensitivity and specificity show an inversely proportional relationship with each other. That is, when sensitivity increases, specificity decreases, and when sensitivity decreases, specificity increases. Therefore, it is virtually impossible to increase the sensitivity and specificity at the same time, and it can be said that the higher the sensitivity and specificity value, the better the performance of the classification model used.

$$TPR = \frac{TP}{TP + FN} \tag{1.1}$$

$$TNR = \frac{TN}{TN + FP} \tag{1.2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.3}$$

$$PRE = \frac{TP}{TP + FP} \tag{1.4}$$

$$MCE = \frac{FP + FN}{TP + TN + FP + FN} \tag{1.5}$$

The ROC curve sets the x axis to the false positive rate, that is, $1 - TNR$, and the y axis to the true positive rate. It can be completed by calculating the false positive rate and true positive rate for all thresholds, making a point on the coordinates and connecting it with a smooth curve. Here, the thresholds are reference points set for classifying the predicted values of the response variables in the classification model. If binary classification is to be performed, the thresholds will be random probabilities that serve as criteria for classification, and their values range from 0 to 1. The area under the ROC curve is called AUC (Area Under Curve), and it

can be said that the larger the area of this AUC, the better the performance of the classification model used. AUC has a value between 0 and 1, but the minimum value of AUC is usually observed around 0.5. In other words, the closer the AUC value is to 1, the better the classification model's performance is. The ROC curve can be drawn as shown in Figure 7.1, and it can be judged that a classification model with better performance is used as the area of the yellow area, is equal to, AUC, is closer to 1.

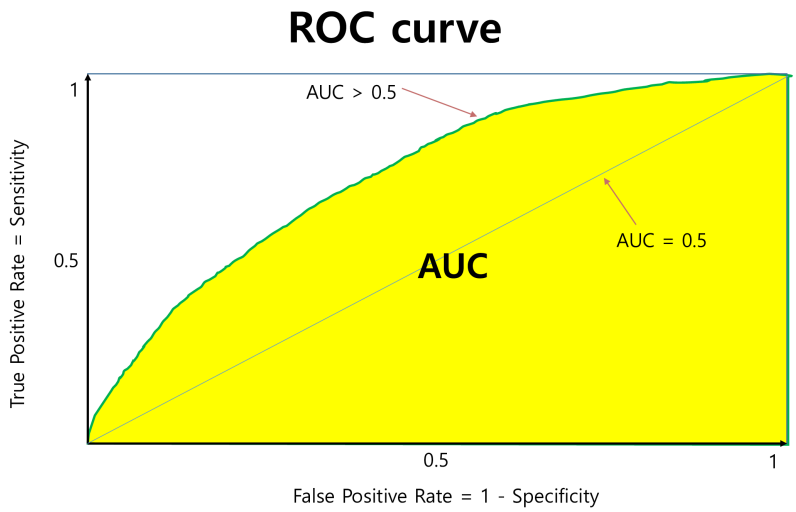


Figure 7.1: ROC curve

7.2 Time-dependent AUC

The ROC curve introduced in Section 7.1 is drawn in only one form from one model that fits specific data. Therefore, the AUC calculated for this ROC curve is also determined with only one value. However, the survival data introduced in Section 2.1 includes the patient's survival time, and a survival model that fits these data depends on the survival time. Therefore, the ROC curve by this survival model is drawn according to the time point of each survival time, and the shape changes depending on the time point of each survival time. Therefore, the AUC obtained through this also shows a different value depending on the time point of each survival time, and this AUC is called time-dependent AUC. For further details on this, see Kamarudin et al. (2017) and Cho, J. (2022).

Refer to Kamarudin et al. (2017), the time-dependent AUC for the survival model is defined as follows. First, let's define some variables to understand the time-dependent AUC. First, let T_i be the time of disease onset for individual values for each $i(= 1, \ldots, n)$, and X_i be the marker value. Here, the marker value is a value usually used to measure the risk status of a specific patient in medical research. If a research on diabetes is conducted, the patient's blood glucose level can be used as the marker value. In general, it is common to use the risk score calculated through the regression model or classification model as a marker value. In this study, the risk score calculated through the Cox regression model was used as the marker value. And let C_i be the censoring time, $Z_i = \min(T_i, C_i)$ be the observed event time, and $\delta_i = I(T_i \leq C_i)$ be defined as a censoring indicator.

Finally, if $D_i(t)$ is disease status at time t, which has a value of 1 if disease occurs, and 0 otherwise, time-dependent sensitivity $Sen(c,t)$ and time-dependent specificity $Spe(c,t)$, which are calculated using threshold c and time t, and time-dependent AUC $AUC(t)$ which is calculated through the time-dependent ROC curve $ROC(t)$ from time-dependent sensitivity and specificity can be defined as follows.

$$Sen(c,t) = P(X_i > c | D_i(t) = 1) \tag{2.6}$$

$$Spe(c,t) = P(X_i \leq c | D_i(t) = 0) \tag{2.7}$$

$$AUC(t) = \int_{-\infty}^{\infty} Sen(c,t) d[1 - Spe(c,t)] \tag{2.8}$$

In above equation, $1 - Spe(c,t) = \frac{\partial[1-Spe(c,t)]}{\partial c} dc$.

Heatherty, P.J. and Zheng, Y. (2005) proposed three methods for defining time-dependent sensitivity and specificity and the time-dependent AUC calculated through them. Three definition methods are introduced as follows.

1) Cumulative sensitivity and dynamic specificity (C/D)

$Sen^C(c,t) = P(X_i > c | T_i \leq t)$

$Spe^D(c,t) = P(X_i \leq c | T_i > t)$

$AUC^{C,D}(t) = P(X_i > X_j | T_i \leq t, T_j > t), i \neq j$

2) Incident sensitivity and dynamic specificity (I/D)

$Sen^I(c,t) = P(X_i > c | T_i = t)$

$Spe^D(c,t) = P(X_i \leq c | T_i > t)$

$AUC^{I,D}(t) = P(X_i > X_j | T_i = t, T_j > t), i \neq j$

3) Incident sensitivity and static specificity (I/S)

$Sen^I(c,t) = P(X_i > c | T_i = t)$

$Spe^S(c,t^*) = P(X_i \leq c | T_i > t^*)$

$AUC^{I,S}(t,t^*) = P(X_i > X_j | T_i = t, T_j > t^*), i \neq j$

In definition I/S, t^* is pre-specified end point. Normally it is considered a long enough time to observe the event. Plus, definition C/D is appropriate to use when you are interested in finding a specific time to divide an individual patient into diseased and non-diseased cases, and definition I/D is appropriate to use when you want to divide an individual into those who have disease and those who do not using predetermined specific time. In this study, time-dependent AUC based on definition C/D was applied, and the following Inverse Probability of Censoring Weighting (IPCW) estimation method was used to estimate time-dependent sensitivity, specificity and time-dependent AUC.

$$\hat{Sen}(c,t) = \frac{\Sigma_{i=1}^n I(X_i > c, Z_i \leq t)\{\delta_i / n\hat{S}_c(Z_i)\}}{\Sigma_{i=1}^n I(Z_i \leq t)\{\delta_i / n\hat{S}_c(Z_i)\}} \tag{2.9}$$

$$\hat{Spe}(c,t) = \frac{\Sigma_{i=1}^n I(X_i \leq c, Z_i > t)}{\Sigma_{i=1}^n I(Z_i > t)} \tag{2.10}$$

$$\hat{AUC}(t) = \frac{\Sigma_{i=1}^n \Sigma_{j=1}^n \{\delta_i / n\hat{S}_c(Z_i)\} I(Z_i \leq t, Z_j > t) I(X_i > X_j)}{\Sigma_{i=1}^n \{\delta_i / n\hat{S}_c(Z_i)\} I(Z_i \leq t) \Sigma_{j=1}^n I(Z_j > t)} \tag{2.11}$$

Here, $\hat{S}_c(Z_i)$ is the Kaplan-Meier estimator of the survival function of the censoring time C_i at the ith observed event time Z_i.

8 Cox regression model

8.1 Cox proportional hazard model

The Cox proportional hazard model is a semi-parametric method frequently used to modeling the hazard function in survival analysis using survival data. As explained in Section 2.2, in order to use this model, the assumption that the hazard between the groups to be compared is uniformly proportional during the follow-up period must be satisfied. For more information on this, see Kleinbaum, D.G. and Klein, M. (2010) and Kim, J. (2016).

Now, I will explain how to estimate the regression coefficient vector β in the Cox proportional hazard model. However, here, the process is described under the assumption that there is no ties in the survival data. First, for the survival data (t_i, δ_i, x_i) for each $i(= 1, \ldots, n)$, assume that $t_i = \min(y_i, c_i)$ is the observed survival time, and $\delta_i = I(y_i \leq c_i)$ is the censoring indicator (In here, y_i is the real survival time, c_i is the censoring time). If $x_i = (x_{i1}, \ldots, x_{ip})^T$ is a time-independent covariate vector and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a regression coefficient vector, the Cox proportional hazard model results in the following form.

$$h(t|x_i) = h_0(t) \exp(x_i^T \beta) = h_0(t) \exp(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p) \tag{1.1}$$

Calculating the hazard ratio for the above model is as follows.

$$HR = \frac{h(t|x_i)}{h(t|x_j)} = \exp[\Sigma_{b=1}^{p}\beta_b(x_{ib} - x_{jb})] \tag{1.2}$$

In other words, the hazard ratio for the Cox proportion hazard model does not depend on the survival time t. Therefore, this model is used only when the proportional hazard assumption is satisfied. In the above model equation, the regression coefficient vector β is estimated by maximizing the partial likelihood function using the Newton-Raphson algorithm. If we assume that there are no ties in the survival data, the partial likelihood function can be written as follows.

$$PL(\beta) = \prod_{i=1}^{n}[\frac{h_0(t_i)\exp(x_i^T\beta)}{\Sigma_{l\in R(t_i)}h_0(t_i)\exp(x_l^T\beta)}]^{\delta_i} = \prod_{i=1}^{n}[\frac{Z_i\exp(x_i^T\beta)}{\Sigma_{l\in R_i}Z_l\exp(x_l^T\beta)}]^{\delta_i} \tag{1.3}$$

Here, $Z_i = Z_i(t)$ has a value of 1 if the ith object belongs to the risk set at the time t, otherwise 0. And $R_j = R(t_j)$ is a risk group, assuming that there are no ties, means a group of living individuals who have not experienced an event until just before time t_j.

The estimator $\hat{\beta}$ for the regression coefficient vector is obtained through the process of maximizing the log partial likelihood function $l(\beta) = \log[PL(\beta)]$. In this process, score function vector $U(\beta)$ and information matrix $I(\beta)$ are used as follows.

$$U(\beta) = [U(\beta_k)]_{p\times 1} = [\frac{\partial l(\beta)}{\partial\beta_1}, \ldots, \frac{\partial l(\beta)}{\partial\beta_p}]^T, (k = 1, \ldots, p) \tag{1.4}$$

$$I(\beta) = [I_{gh}(\beta)]_{p\times p} = -\begin{bmatrix} \frac{\partial^2 l(\beta)}{\partial\beta_1^2} & \frac{\partial^2 l(\beta)}{\partial\beta_1\partial\beta_2} & \cdots & \frac{\partial^2 l(\beta)}{\partial\beta_1\partial\beta_p} \\ \vdots & \frac{\partial^2 l(\beta)}{\partial\beta_2^2} & \cdots & \frac{\partial^2 l(\beta)}{\partial\beta_2\partial\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial\beta_p\partial\beta_1} & \frac{\partial^2 l(\beta)}{\partial\beta_p\partial\beta_2} & \cdots & \frac{\partial^2 l(\beta)}{\partial\beta_p^2} \end{bmatrix}, (g, h = 1, \ldots, p) \tag{1.5}$$

Here, $U(\beta_k)$ and $I_{gh}(\beta)$ are calculated as follows.

****

$U(\beta_k) = \frac{\partial l(\beta)}{\partial \beta_k} = \Sigma_{i=1}^{n} \delta_i [x_{ik} - \frac{\Sigma_{l \in R_i} x_{lk} \exp(x_l^T \beta)}{\Sigma_{l \in R_i} \exp(x_l^T \beta)}]$

$I_{gh}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta_g \partial \beta_h}$

$= \Sigma_{i=1}^{n} \frac{\Sigma_{j \in R(t_i)} x_{jg} x_{jh} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}{\Sigma_{j \in R(t_i)} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}$

$-\Sigma_{i=1}^{n} [\frac{\Sigma_{j \in R(t_i)} x_{jh} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}{\Sigma_{j \in R(t_i)} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}][\frac{\Sigma_{j \in R(t_i)} x_{jh} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}{\Sigma_{j \in R(t_i)} \exp[\Sigma_{k=1}^{p} \beta_k x_{jk}]}]$

****

Using score function vector and information matrix in equation 1.4 and 1.5, and the appropriate pre-determined initial value $\hat{\beta}^{(0)}$ for the estimator $\hat{\beta}$ of the regression coefficient vector, Newton-Raphson algorithm $\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + I^{-1}(\hat{\beta}^{(m)}) U(\hat{\beta}^{(m)})$ is repeated to estimate β until the log partial likelihood derivative converges ($l(\hat{\beta}^{(m+1)}) \approx l(\hat{\beta}^{(m)})$). This allows us to obtain a Breslow estimator for the cumulative hazard function as follows.

$$\hat{H}(t|x) = \Sigma_{t(i) \leq t} \frac{\Sigma_{l=1}^{n} \delta_i I(t_l = t_i)}{\Sigma_{l \in R_i} \exp(x_l^T \hat{\beta})} \tag{1.6}$$

In here, $t_{(1)} < \cdots < t_{(n)}$ are order statistics of time t. And using the estimator $\hat{\beta}$ for the regression coefficient vector β, $x_i^T \hat{\beta}$, $i = 1, \ldots, n$ can be calculated. This is the estimator of the risk score obtained through the Cox proportional hazard model. This value can be used as a necessary marker value when calculating the time-dependent AUC introduced in Section 7.2. The estimator $\hat{\beta}$ of the regression coefficient vector calculated through this process has consistency that converges to the true value β as the sample size increases, asymptotic normality that follows an approximately normal distribution, and efficiency with the smallest variance among the various estimators for β. In addition, for cases where ties exist in the

survival data, methods for handling these ties have been proposed. These include the exact method, which treats ties by calculating the partial likelihood function from the set of possible risks, the Breslow approximation(1974), which is often used when the approximation is good, and the Efron approximation(1977), which is close to partial likelihood. For a detailed description of this, see Breslow, N.E. (1975) and Efron, B. (1974).

8.2 Extended Cox regression model

The Cox proportional hazard model described in Section 8.1 can be used when it is assumed that the hazard between the groups to be compared is uniformly proportional during the follow-up period. However, when looking at various survival data, the explanatory variables in the covariate vector will not always be independent of the survival time. In some cases, it is possible that there may be a time varying covariate that is associated with a change in survival time. A model that can be considered in this case is the Extended Cox regression model. For details on the Extended Cox regression model, see Zhang et al. (2018) and Therneau, T. and Grambsch, P. (2000). For example, suppose we have the following Cox regression model.

$$h(t, x_i(t)) = h_0(t) \exp[\Sigma_{b=1}^{p_1} \beta_b x_{ib} + \Sigma_{b=1}^{p_2} \gamma_b x_{ib}(t)] \tag{2.7}$$

The regression coefficient vector for the above model is $\beta^* = (\beta_1, \beta_2, \ldots, \beta_{p_1}, \gamma_1, \gamma_2, \ldots, \gamma_{p_2})^T$ and the covariate vector is $x_i(t) = (x_{i1}, x_{i2}, \ldots, x_{ip_1}, x_{i1}(t), x_{i2}(t), \ldots, x_{ip_2}(t))^T$. That is, some explanatory variables depend on the change in survival time t. The hazard ratio for this model is calculated as follows.

$$HR = \frac{h(t, x_i(t))}{h(t, x_j(t))} = \exp[\Sigma_{b=1}^{p_1} \beta_b (x_{ib} - x_{jb}) + \Sigma_{b=1}^{p_2} \gamma_b (x_{ib}(t) - x_{jb}(t))] \tag{2.8}$$

In other words, the Extended Cox regression model including time varying covariates does not satisfy the proportional hazard assumption because the hazard ratio changes with the survival time t. Therefore, this model can be applied when the assumption that the risk between groups is proportionally proportional to the survival time during the follow-up period is not satisfied.

9 Calibration

9.1 Calibration

The concept of calibration or post processing is commonly referred to in materials science and metrology, and usually refers to confirming that the measured values calculated for research are well calculated in accordance with known calibration standards in comparing measurement methods. When this concept was approached for survival analysis, Pepe, M. and Janes, H. (2013) defines that it is well-calibrated when the number of events within any time interval measured by the fitted model through survival data is similar to the actual observed number. For details on calibration in survival analysis, see Goldstein et al. (2020).

Pepe, M. and Janes, H. (2013) define calibration in survival analysis as follows. For easier explanation, only the case of a binary outcome will be explained. For observation (t, δ, x) with covariates x, if y is the real survival time and c is the censoring time, then $t = \min(y, c)$ is the observed survival time, $\delta = I(y < c)$ becomes a censoring indicator. Here we assume censoring-at-random, that is, $y \perp c|x$. In other words, for given covariates x, we assume that y and c are independent each other. In addition, let P be the joint distribution of (y, x), F be the conditional cumulative distribution function(CDF) of $y|x$, and let d be the binary

outcome distributed conditional on covariates x. And suppose d and x have a joint distribution $P(d, x)$. If θ is the parameter estimated through model fitting and $risk_\theta(x)$ is defined as the modeled probability $P_\theta(d = 1|x)$, then calibration is a case that satisfies the following conditions.

$$P(d = 1|risk_\theta(x) = r) \approx r \tag{1.1}$$

In other words, if the event frequency is r for subjects with modeled risk r, we say that it is well-calibrated. Goldstein et al. (2020) expanded this concept and established the concept of calibration for the survival problem as follows. If F_θ is the model CDF for the joint distribution $P(y, x)$, the risk for survival analysis is defined by $risk_\theta(y, x) = F_\theta(y|x)$. For this, for all sub-intervals $V = [a, b]$ included in the interval $[0, 1]$, the calibration condition for the survival problem is defined as follows.

$$P(risk_\theta(y, x) \in V) = E_{P(y,x)}[I(F_\theta(y|x) \in V)] = |V| \tag{1.2}$$

That is, when $F_\theta = F$ is satisfied, that is, if the built model shows perfect predictive power, the probability that $F_\theta(y|x)$ takes a value within the interval V is exactly $|V|$. The reason is that if the perfect calibration condition is satisfied, the model CDF F_θ follows the Uniform distribution $U(0, 1)$ according to the principle of inverse transform sampling.

9.2 Concordance Index

Concordance Index (C-Index) is a type of index that evaluates the accuracy of the model built in survival analysis. It evaluates whether the model predicts the sequence of events of interest by relatively evaluating the survival times of several subjects. For details on C-Index, see Steck et al. (2007). Erik, D. (2018) defines C-Index as follows.

$$C(\eta) = \frac{1}{D}\Sigma_{i:\delta_i=1}\Sigma_{j:t_j>t_i}I(\eta_i > \eta_j) = \frac{1}{D}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\delta_i Y_j(t_i)I(\eta_i > \eta_j) \tag{2.3}$$

In the above equation, $Y_j(t_i) = I(t_j > t_i)$ and η_i is the risk score of patient i. and D is the cardinality of the set for all possible pairwise comparisons. The larger the C-Index value, the higher the accuracy of the constructed model. Since this C-Index is composed of an indicator function, its form is not continuous. Therefore, it is impossible to differentiate. To complement this point, Steck et al. (2007) proposed to create a differentiable alternative index that is always less than or equal to the C-Index by using the log-sigmoid lower bound. First, the sigmoid function has the following form.

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{2.4}$$

For this, based on all real number x, set two functions $f(x) = I(x > 0)$ and $g(x) = 1 + [\log \sigma(x)/\log 2]$. If you draw a graph for them, it can be drawn as shown in Figure 9.1. In conclusion, it can be seen that $I(x > 0) \geq 1 + [\log \sigma(x)/\log 2]$ holds for all real numbers x. Based on this fact, the concave function $\tilde{C}(\eta)$ in a differentiable form that can be used instead of C-Index $C(\eta)$ when evaluating

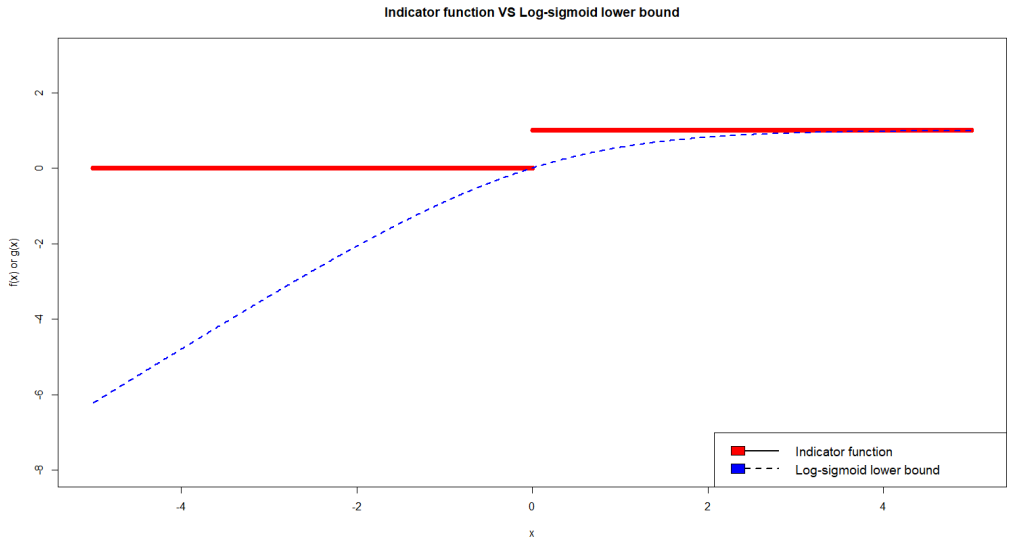


Figure 9.1: Indicator function VS Log-sigmoid lower bound

the model can be written as follows. Through the process of maximizing this concave function, it is possible to determine the estimator and model that show the optimal performance. In here, $C(\eta) \geq \tilde{C}(\eta)$.

$$\tilde{C}(\eta) = \frac{1}{D}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\delta_i Y_j(t_i)[1 + \log(\sigma(\eta_i - \eta_j))/\log 2] \tag{2.5}$$

Differentiating this concave function $\tilde{C}(\eta)$ with respect to the ith risk score is as follows.

$$\frac{\partial\tilde{C}(\eta)}{\partial\eta_i} = \delta_i\Sigma_{k\neq i}Y_k(t_i)[1 - \sigma(\eta_i - \eta_k)] - \Sigma_{k\neq i}\delta_k Y_i(t_k)[1 - \sigma(\eta_k - \eta_i)] \tag{2.6}$$

If the ith risk score η_i is set to $\eta_i = x_i^T\beta$, which is a form that can be calculated through the Cox regression model described in Chapter 8, $\tilde{C}(\eta)$ can be differentiated with respect to the regression coefficient vector β as follows.

****

$\frac{\partial \tilde{C}(\beta)}{\partial \beta} = (\frac{\partial \eta}{\partial \beta^T})^T \frac{\partial \tilde{C}(\eta)}{\partial \eta} = X^T \frac{\partial \tilde{C}(\eta)}{\partial \eta}$

$= \Sigma_{i=1}^{n} x_i [\delta_i \Sigma_{k:Y_k(t_i)=1} [1 - \sigma(\eta_i - \eta_k)] - \Sigma_{j:Y_i(t_j)=1} \delta_j [1 - \sigma(\eta_j - \eta_i)]]$

****

Here, if we solve equation $\frac{\partial \tilde{C}(\beta)}{\partial \beta} = 0$, the regression coefficient vector estimator $\hat{\beta}$ that maximizes the concave function $\tilde{C}(\beta)$ can be obtained.

10

10.1

11

11.1

12

12.1

13

13.1

14

14.1

15

15.1

References

Buckley, J. and James, I. (1979). Linear regression with censored data, *Biometrika,* **66**, 429-436.

Koul, H., Susarla, V., Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *Annals of Statistics.* **9**, 1276–1288.

Beran, R. (1981). *Non-parametric regression with randomly censored survival data.* Technical Report, Univ. California, Berkeley.

Suarez, R.P., Abad, R.C., and Fernandez, J.M.V. (2021). *Bootstrap Selector for the Smoothing Parameter of Beran's Estimator.* Engineering Proceedings.

Geerdens, C., Acar, E.F. and Janssen, P. (2018). Conditional copula models for right-censored clustered event time data. *Biostatistics.* **19** (2), 247-262.

Leurgans, S. (1987). Linear models, random censoring and synthetic data, *Biometrika,* **74**, 301–309.

Breiman, L. (1996). *Out-of-bag estimation,* Technical report, Department of Statistics, University of California at Berkeley, CA, USA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* **58**, 267-288

Sposto, R. (2002). Cure model analysis in cancer: an application to data from the Children's Cancer Group, *Statistics in medicine.* Volume **21**, Issue **2**, 293-312.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics.* **1**, 302-332.

Kleinbaum, D.G. and Klein, M. (2010). *Survival Analysis,* Springer.

Breslow, N.E. (1975). Analysis of Survival Data under the Proportional Hazards Model, *International Statistical Review / Revue Internationale de Statistique.* **43** (1): 45–57.

Efron, B. (1974). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association.* **72** (359): 557–565.

Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The Elements of Statistical Learning, 2nd Edition,* Springer.

Gail, M., Krickeberg, K., Samet, J.M., Tsiatis, A. and Wang, W. (2012). *Survival Analysis, A Self-Learning Text, 3rd Edition,* Springer.

Zhou, Z.H. (2012). *Ensemble Methods: Foundations and Algorithms,* CRC Press, Boca Raton, FL.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in R,* Springer.

Nguyen, V. (2015). Mahalanobis kernel-based support vector data description for detection of large shifts in mean vector, *Electronic Theses and Dissertations.*

1160.

Huh, M. (2015). Kernel-trick regression and classification. *Communications for Statistical Applications and Methods.* **22**, 201-207.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels,* MIT Press.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, 2nd Edition,* Chapman Hall.

Freund, Y., Schapire, R. and Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence.* **14** (5), 771-780.

Lee, S., Han, S. and Hwang, S. (2016). Ensemble approach for improving prediction in kernel regression and classification. *Communications for Statistical Applications and Methods.* **23**, 355-362.

Minh, H.Q., Niyogi, P. and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. *International Conference on Computational Learning Theory, COLT 2006: Learning Theory* pp 154-168.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006). Support vector machines in R, *Journal of statistical software.*

Souza, C.R. (2010). Kernel functions for machine learning applications, *Creative Commons Attribution-Noncommercial-ShareAlike 3.0, crsouza.com.*

Sabin, C. and Petrie, A. (2019). *Medical statistics at a glance,* John Wiley Sons, Ltd.

Chen, D.G.D., Peace, K.E. and Zhang, P. (2017). *Clinical trial data analysis using R and SAS,* Chapman and Hall.

Hosmer, D.W., Lemeshow, S., and May, S. (2008). *Applied survival analysis: regression modeling of time-to-event data,* Wiley-Interscience, New Jersey.

Therneau, T. and Grambsch, P. (2000), *Modeling Survival Data: Extending the Cox Model,* Springer-Verlag, New York.

Klein and Moeschberger (1997), *Survival Analysis Techniques for Censored and truncated data,* Springer. National Longitudinal Survey of Youth Handbook The Ohio State University, 1995.

Goldstein, M., Han, X., Puli, A., Perotte, A. and Ranganath, R. (2020), *X-CAL: Explicit Calibration for Survival Analysis,* Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

Pepe, M. and Janes, H. (2013), Methods for evaluating prediction performance of biomarkers and tests, *Risk assessment and evaluation of predictions,* pages 107–142. Springer, 2013.

David, M.S. and Lisa, J.S. (2018), *Testing Calibration of Cox Survival Models at Extremes of Event Risk,* Frontiers in genetics, 2018-frontiersin.org

Kamarudin, A.N., Cox, T. and Kolamunnage-Dona, R. (2017), *Time-dependent ROC curve analysis in medical research: current methods and applications,* BMC Medical Research Methodology (2017)

Heagerty, P.J. and Zheng, Y. (2005), Survival model predictive accuracy and ROC

curves, *Biometrics.* 2005 ; **61** (1): 92–105.

Yanagisawa, H., Iwamori, T., Koseki, A., Kudo, M., Ghalwash, M. and Chakraborty, P. (2021), *Simpler Calibration for Survival Analysis,* ICLR 2022 Conference.

Austin, P.C. (2012), Generating survival times to simulate Cox proportional hazards models with time-varying covariates, *Statistics in medicine, 2012–Wiley Online Library.* **31**, 3946-3958.

Steck, H., Krishnapuram, B., Raykar, V.C., Dehing-Oberije, C., Lambin, P. (2007), *On Ranking in Survival Analysis: Bounds on the Concordance Index,* Advances in Neural Information Processing Systems 20 (NIPS 2007).

Erik, D. (2018), *A convex approximation of the concordance index (C-index),*, Blog of Erik Drysdale, last modified August, 22nd. 2018, accessed July, 5th. 2022, https://www.erikdrysdale.com/survConcordance/

Fawcett, T. (2006), An Introduction to ROC Analysis, *Pattern Recognition Letters.* **27** (8): 861–874.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. and Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of translational medicine,* **6** (7).

Kim, J. (2016). *The basic survival analysis using R,* FREEACADEMY INC.

Han, S. (2016). *A study on kernel ridge regression using ensemble method* (Master's thesis), The Graduate school of Hankuk University of Foreign Studies

Hwang, S. (2017). *A study on efficiency of kernel ridge logistic regression classification using ensemble method* (Master's thesis), The Graduate school of Hankuk University of Foreign Studies

Kim, J. (2018). *A comparison study for regression coefficient estimation in robust LASSO regression* (Master's thesis), The Graduate school of Hankuk University of Foreign Studies

Kim, J. (2018). *Variance reduction via guided Non-Parametric regression in censored data* (Master's thesis), The Graduate school of Hankuk University of Foreign Studies

Lee, S. (2018). *Variable selection in censored regression models* (Master's thesis), The Graduate school of Hankuk University of Foreign Studies

Lee, J. (2020). *A study on recent boosting methods* (Master's thesis), The Graduate school of Konkuk University

Jeon, B. (2022). *Variance reduction via guided non-parametric regression in censored data with dependent censoring* (Master's thesis), The Graduate school of Jeonbuk National University

Cho, J. (2022). *A Comparison of Determining Optimal Cutpoints in Continuous Biomarkers Utilizing a Time-dependent ROC Curve: A simulation study* (Master's thesis), The Graduate school of Jeonbuk National University

Kim, E. (2022). *Comparing weighting methods in propensity score analysis for multiple treatments* (Master's thesis), The Graduate school of Jeonbuk National

University

Jeong, D. (2022). *Comparative Study on Prediction Performance in Classification of Imbalanced Data: Simulation based approach* (Master's thesis), The Graduate school of Jeonbuk National University

국문초록

중도절단이 포함된 생존자료 분석 시 기계학습 방법을 통한 성능향상에 관한 연구

황 성 윤

전북대학교 대학원 통계학과

본 논문은 총 2가지의 연구에 대한 내용을 담고 있으며 모두 생존자료를 분석하는 것과 관련이 있다.

첫 번째 연구는 중도절단(censoring)이 포함된 데이터에 대하여 회귀분석을 실시하는 경우 예측력(predictive power)을 향상시킬 수 있는 방법에 관한 것이다. 중도절단은 보통 의학 분야에서 자주 등장하는 환자의 생존시간(survival time)과 관련한 생존자료(survival data)에서 환자가 연구대상인 질병 이외의 요인에 의해 사망하게 되는 등의 내부적 또는 외부적인 원인에 의하여 발생하게 된다. 생존자료를 분석하는 가장 큰 목적은 어떠한 요인이 환자의 생존시간에 유의미한 영향력을 미치게 되는지 확인하고 이를 통해 환자의 생존시간을 예측하는 것이다. 이러한 중도절단이 포함된 생존자료의 경우는 추정의 대상이 되는 생존시간이 부분적으로만 관측되기 때문에 이를 대체하기 위한 인조변수(synthetic response)를 만들어서 자료를 분석할 수 있다. 하지만, 이러한 인조변수는 설명변수(explanatory variable)가 주어졌을 경우의 조건부분산(conditional variance)이 원래 생존시간의 조건부분산보다 커지는 경향이

있고 생존시간이 증가할수록 증가하는 폭도 커지는 특성이 있다. 이 때문에 추정량에 대한 안정성이 떨어져서 문제가 될 수 있다. 이러한 문제점을 보완하기 위해 본 연구에서는 인조변수에 대한 회귀모형을 구축할 경우 변환함수를 따로 지정할 필요 없이 복잡한 비선형 데이터에 대해 적절한 사상함수를 사용해 설명변수 공간에 있는 데이터를 고차원의 특성 공간으로 이동시키는 커널트릭 기법(kernel trick method)과 다중공선성(multicollinearity)의 문제가 있을 때 적용 가능한 능형 회귀분석(ridge regression) 방법을 적용한다. 여기에 추가로 배깅(bagging) 및 랜덤포레스트(random forest)와 같은 앙상블 기법(ensemble method)을 적용하여 추정량의 분산을 줄임으로써 생존시간에 대한 예측력을 향상시키는 방법에 관하여 제안하고자 한다. 컴퓨터 모의실험을 통하여 다양한 상황을 가정하고 중도절단이 포함된 데이터에서 설명변수에 대한 예측력을 비교, 분석하였다. 이를 통해 본 연구에서 제안하고자 하는 방법이 일반적인 방법과 비교했을 때 전체적으로 우수한 예측력을 보임을 확인할 수 있었다.

두 번째 연구는 생존자료를 분석 시 산출할 수 있는 time-dependent AUC를 전체적으로 향상시킬 수 있는 방법에 관한 것이다. 계속 설명...

주요용어 : 생존자료, 인조변수, 능형 회귀분석, 기계학습, 커널트릭 기법, 앙상블 기법, time-dependent AUC, Cox 비례위험모형

Acknowledgement

...

July, 2024

Seong-Yun Hwang