Cox Proportional Hazards Regression for Interval Censored data with an Application to College Entrance and Parental Job Loss

Hee Jin Kim
Chungnam National University
SungHun Kim
Chungnam National University
Eunjee Lee (eunjee.cnu@gmail.com)
Chungnam National University

Research Article

Keywords: Parental job loss, college entrance, survival analysis, Cox proportional hazards rate, interval-censoring, multiple imputation

Posted Date: May 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1611050/v1

Cox Proportional Hazards Regression for Interval Censored data with an Application to College Entrance and Parental Job Loss

Hee Jin Kim[1], SungHun Kim[1] and Eunjee Lee[1*]

[1*]Department of Information and Statistics, Chungnam National University, 99, Daehak-ro, Yuseong-gu, 34134, Daejeon, Republic of Korea.

*Corresponding author(s). E-mail(s): eunjee.cnu@gmail.com;
Contributing authors: black7cat@naver.com;
uiop8533@gmail.com;

Abstract

This paper conducted a survival analysis by fitting a Cox proportional hazards model to the Korea Labor panel data in order to analyze the effect of parental job loss on children's admission to a college and university. Since the Korea Labor panel data is interval- and right-censored, we compared three imputation methods: simple omission, imputation as the average of the left and the right values of the interval, and multiple imputations proposed by Pan (2000). Their integrated area under the curve (iAUC) and mean square error (MSE) were compared in order to assess the predictive and estimation performances. It was found that, within the simulation, the multiple imputation method had a lower MSE than the other two methods. However, no difference was observed in the iAUC values. In the group where each householder had at least a college degree, the parental job loss variable was significantly related to college or university admission of the first child in a household regardless of the interval-censoring imputation method.

Keywords: Parental job loss, college entrance, survival analysis, Cox proportional hazards rate, interval-censoring, multiple imputation

1 Introduction

In South Korea during the 1980s, college graduates were guaranteed quality jobs because college entrance rates were at an all time low. Therefore, because university admission was regarded as a measure of success, the search for factors associated with admission continued for a long time. Factors related to adolescence-level college admission can be largely divided into economic and environmental factors. Parents' income can affect household economic level as well as the child's educational achievement. [1] shows that household financial stability ensures material support for raising children and further affects adolescent self-esteem. The Canadian Survey of Labour and Income Dynamics (SLID) found that parents' unemployment and decrease in income due to dismissal and company bankruptcy had a negative impact on adolescent students who graduated from high school with the intention of going to college or university[2]. [3] analyzed differences with regard to college admission among adolescents based on differences between parents' unemployment period. The subjects were all households where parents experienced unemployment. The experimental group—the subject of research interest—experienced parental unemployment when the children were aged 15-17 years old, and the control group included individuals who experienced parental unemployment when they were aged 21-23 years. The analysis included the experimental group and the control group in the linear regression model, and it was therefore confirmed that the experimental group's college entrance rate was 10% lower than that of the control group. [4] analyzed the impact of parents' unemployment on their children's educational achievement. Using data for a period of three years (1998-2000)from the Korean Labor Income Panel Study (KLIPS) and data regarding children who experienced parental unemployment at the ages of 16 to 18 years, [4] confirmed that the probability of achieving college admission at the age of 19 was low based on their logistic regression analysis. In particular, it revealed that the effects of household poverty and economic loss were similar in magnitude.

Compared to past years, the job market landscape has changed immensely due to the sharp increase in college entrances and the use of "blind" hiring practices in public sectors in South Korea. "Blind" is a new hiring method that does not require a photo, school background, or personal information unrelated to the duties involved in the position. The proportion of four-year college graduates, which accounted for 11.4% of the total unemployed population in 2000, increased steadily to 20.5% in 2008 and 32% in 2016. Although college graduation does not guarantee employment these days, there is still a wage gap depending on education level. According to OECD Korean statistics, in 2018, when the wages for high school graduates reached 100 as a relative wage for education based on the age range of 25 to 64 years, it was 114.2% for college graduates, 143.1% for college graduates, and 189.6% for graduate graduates [5]. Therefore, current research should aim to develop a policy targeting adolescents who cannot expect parental help.

Meanwhile, the university entrance rate in South Korea in 2019 was 69.8%, higher than the OECD average of 44.9%[6]. In the current situation, where the majority of high school graduates go on to college, there is a need to change research direction by analyzing the factors affecting delays in college admissions. Therefore, this work used data from the KLIPS to analyze the relationship between parents' economic factors and the timing of their children's college admissions. Since the time taken to gain admission and enroll in college can be considered the survival time, this study used the Cox proportional hazard model, a representative semi-parametric survival analysis model.

Chapter 2 explains the background of this study's use of the Cox proportional hazard model and the characteristics of interval censoring in the data used in this study. Chapter 3 introduces several methodologies for handling interval censoring in the Cox proportional hazard model and describes the multiple imputation approach, which is the focus of this paper. Chapter 4 describes the simulation for comparing the performance of the methodology for processing interval censoring and the results. In Chapters 5 and 6, the factors affecting the child's university entrance are analyzed using data from the Korea Labor Panel Survey for the last 20 years, and its significance is examined.

2 Theoretical Background

2.1 Korean Labor Income Panel Study (KLIPS)

This study's utilized data were taken from the Korean Labor & Income Panel Study (KLIPS), and the survey was administered to households living in non-farm areas and their household members. The members of the panel sample were household members from a sample of 5,000 households. The KLIPS, a longitudinal survey, follows up on the subjects once a year to gain data about economic activities, labor market movement, income activities and consumption, education and vocational training, and social life.

As the Korea Labor Panel Data form the oldest data in the panel survey, many years of data regarding education level, economic activity status, and family composition of households and individuals have been accumulated. The youth panel survey also collects data through a long-term follow-up survey on economic activity information from the survey targets. Nevertheless, this survey is limited in that it does not provide enough information about the parents' educational level and economic activity status because it targets young people.

2.2 Characteristics of Time to Admission as Survival Data

Survival Analysis considers the time elapsing until the interest event occurs as a response variable. While the medical field has widely used survival analysis, the social science field has rarely conducted survival analysis for repeatedly measured data. In particular, the panel data of the Korea Labor Institue,

which is collected annually from a survey administered to the same target audience, is appropriate for survival analysis methodology in that the time to enter university can be utilized as a response variable. However, if a subject enters university during a year when the subject's response is missing, the exact timing of the college entrance will be unknown, and interval censoring will occur. Furthermore, right-censoring occurs when no further investigation can be made due to the moving of, immigration of, or loss of contact with the household to be investigated.

3 Cox proportional hazards regression with interval censoring

3.1 Cox proportional hazards regression

Survival models explore a relationship between a hazard function and a set of covariates. The Cox proportional hazards model assumes that the effect of a unit increase in a covariate is multiplicative to the hazard rate with a proportional hazard assumption. The proportional hazard model has a non-parametric form in that it does not assume any distribution for the survival time or specification for the baseline hazard function. Furthermore, because it assumes only the model for the regression coefficient β_k and uses a parametric method to conduct estimations β_k, it is a semi-parametric model.

We will assume that T is a non-negative continuous random variable representing the survival time or the time to a specific event (e.g., time to admission). The hazard function at time t is defined as

$$h(t) = \lim_{dt \to 0} \frac{P\{t \leq T \leq t + dt | T \geq t\}}{dt}.$$

The Cox regression specifies the hazard function of i-th subject with covariates $\boldsymbol{x}_i$ in the following manner.

$$h_i(t) = h(t|\boldsymbol{x}_i) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k x_{ik}\right)$$

The covariate vector is given by $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, and the baseline hazard function at time t is denoted by $h_0(t)$. The regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ are estimated by maximizing the partial likelihood as $\hat{\boldsymbol{\beta}}$, where the partial likelihood is given by

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right]^{\delta_i}. \tag{1}$$

δ_i is an indicator variable for the censoring of the i-th subject. $R(t_i)$ is a risk set that is exposed at any risk at time t_i, which includes subjects that have not experienced the event before t_i and are not censored. A null hypothesis for $\boldsymbol{\beta}$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, is tested using a Wald test, a likelihood ratio test, or a Score test.

3.2 Interval censoring

The data with censoring are denoted by $D = \{(A_i, \boldsymbol{x}_i), i = 1, \cdots, n\}$, where A_i is $(L_i, R_i]$, and $\boldsymbol{x}_i$ is a p-dimensional covariate vector. If the survival time T_i is left-censored, $L_i = 0$; if the survival time is not censored, $L_i = R_i$. If it is right-censored, $R_i = \infty$. Since the partial likelihood function in the equation 1 can be calculated for the data with right-censoring but not for the data with left- or interval-censoring, an additional step is necessary for the interval-censored cases in the KLIPS data. We considered three approaches for dealing with the interval censored cases: omitting the interval-censored cases, midpoint imputation, and multiple imputation. Midpoint imputation refers to imputing the interval-censored time to event by using the midpoint of the interval $L_i = R_i$ as $(L_i + R_i)/2$. Since the midpoint imputation may cause biased analysis result, the following methods are suggested. As a nonparametric method, [7] suggested nonparametric survival functional estimation methods satisfying self-consistency. [8] proposed iterative convex minorant (ICM) algorithms to improve convergence speed of Turnbull's method. [9] combined an EM algorithm and an ICM algorithm as an EM-ICM algorithm. [10] applied a Newton-Raphson algorithm to the Cox regression by adding covariates to the model, in a nonparametric way. [11] proposed the use of multiple imputation of [12] for interval-censored data and the employment of Cox regression for the imputed data. While the midpoint imputation can be classified through simple imputation, multiple imputation is one of the probability-based imputation methods.

3.3 Multiple Imputation for Cox regression

[11] proposed the use of multiple imputation of [12] for interval-censored data and the employment of Cox regression for the imputed data. In this paper, we used the MIICD package[13] in the R program to implement the multiple imputation method. For the imputation of the interval-censored data, we considered the use of poor man's data augmentation (PMDA) or asymptotic normal data augmentation (ANDA). When there are few missing values, the PMDA methodology underestimates the actual variability (Wei and Tanner, 1991), and ANDA is recommended for the imputation algorithm. In addition, when the number of truncated data is small, the regression coefficient converges to 0, so it is recommended to use "Link estimate" instead of a Breslow method to estimate the baseline survival function [11].

This method uses an iterative algorithm and generates multiple imputed data sets. The subscript (k) and the superscript (i) represent the k-th imputed data set and the i-th iteration, respectively. Let's say $(T_{(k)}, \delta_{(k)}, Z)$ is m censoring values for $k = 1, \ldots, m$. T is the observed event time, δ is whether or not it is censored, and Z is the set of covariates. When $\{L_j < T_j < R_j\}$, $T_{(k),j} = T_j$ and $\delta_{(k),j} = 1$. In the case of $R_j = \infty$, $T_{(k),j} = L_j$, and $\delta_{(k),j} = 0$. The multiple imputation method proposed by Pan converts the interval-censored data to the right-censored data using PMDA or ANDA method, and then calculates it

through the partial likelihood ratio. The detailed algorithm is as follows. Without loss of generality, only one explanatory variable x_j and the corresponding regression coefficient β are considered.

1. In the i-th iteration, the estimates for the regression coefficient and the baseline survival function are denoted by $\hat{\beta}^{(i)}$ and $\hat{S}_0^{(i)}$. Note that the starting value is $\hat{\beta}^{(0)}= 0$. After assuming a uniform distribution for L_j and R_j in the m sets, the failure time X_j is randomly generated and designated as an imputed value. This is expressed as $T_{(k),j} = X_j$, $\delta_{(k),j}= 1$. The baseline survival probability $\hat{S}_{0,(k)}^{(0)}$ is the a Breslow estimate of the baseline survival probability for the kth replaced data set.
2. Generate m sets of imputed data $\{X_{(1)}, \delta_{(1)}, x\}$, $\cdots$, $\{X_{(m)}, \delta_{(m)}, x\}$, which are possibly right-censored as follows. For each observation L_j, R_j, x_j, $j = 1, \cdots, n$, m created as right-censored data by replacing interval censoring are empirically appropriate, and $\hat{S}_0$ in the second step is discontinuously assumed as follows: Interval censoring (L_j, R_j) and ith base survival function $[\hat{S}_0^{(i)}]^{\exp\left(Z_j\hat{\beta}^{(i)}\right)}$ is $\{p_1, \ldots, p_{k_j}\}$ following the probability mass function at $\{t_1, \ldots, t_{k_j}\}$. Here, the failure time X_j is randomly proportional to the probability at $\{t_1, \ldots, t_{k_j}\}$ with the probability mass function value $\{p_1, \ldots, p_{k_j}\}$.
3. Since all the interval-censored values are imputed, the Cox proportional hazard model can be employed. Through this, the regression coefficient estimate can be considered as being $\hat{\beta}_{(k)}^{(i)}$, and the covariance estimate can be considered as being $\widehat{\Sigma}_{(k)}^{(i)}$.
4. $\left(T_{(k)}, \delta_{(k)}, Z\right)$ denotes the kth right-censored data of m sets obtained through the imputation of the interval censored data. $\hat{\beta}_{(k)}^{(i)}$ denotes the regression coefficients obtained by fitting a Cox proportional hazard model. The Breslow estimate $\hat{S}_{0,(k)}^{(i)}$ for the basis survival function is calculated based on $\left(T_{(k)}, \delta_{(k)}, Z\right)$ and $\hat{\beta}_{(k)}^{(i)}$.
5. In the ith iteration, $\hat{\beta}_{(k)}^{(i)}$ of m sets is summed and divided by m, which is denoted by $\beta^{(i+1)}$. In this way, the basis survival function is also obtained. The covariance is the sum of the intra-group and inter-group imputation variances. This can be expressed as an equation as follows.

$$\hat{\beta}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_{(k)}^{(i)}, \quad \hat{S}_0^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{S}_{0,(k)}^{(i)}$$

$$\widehat{\Sigma}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\Sigma}_{(k)}^{(i)} + \left(1+\frac{1}{m}\right)\frac{\sum_{k=1}^{m}\left(\hat{\beta}_{(k)}^{(i)} - \hat{\beta}^{(i+1)}\right)}{m-1}$$

Finally, it repeats from the first until the $\hat{\beta}^{(i)}$ converges.

The ANDA method includes a variation in the fifth step of the PMDA above. In the fifth step, the normal distribution is assumed with a mean vector of the regression coefficients, and a covariance matrix of the covariances are obtained from the kth set; furthermore, the estimated value of the regression coefficients is obtained.

$$\hat{g}^{(i+1)}(\beta) = \frac{1}{m}\sum_{k=1}^{m} N\left(\hat{\beta}_{(k)}^{(i)}, \widehat{\Sigma}_{(k)}^{(i)}\right)$$

4 Simulation study

In order to find an appropriate imputation method for the KLIPS data with interval censoring, the imputation performance of the imputation methods for interval-censored data, simple omission of the interval-censored data, the midpoint imputation, and the multiple imputation were compared. The mean squared error (MSE) was calculated to evaluate the performance of the three imputation methods. The MSE is a method for measuring the accuracy of the estimated regression coefficient value.

$$MSE = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2$$

4.1 Simulation settings

In order to generate the simulation data, we mimicked the censoring rate of the KLIPS data, while we considered several scenarios for the interval-censoring rate from the low values to the high values. The right-censoring rate was fixed at 20%, and the interval-censoring rate was fixed as 15%, 30%, or 45%. The simulation data were sampled with replacement from the KLIPS data of 376 subjects whose parental education level reached college or higher. We considered 100, 300, and 1000 sample sizes in order to compare the imputation methods for different cases.

The baseline hazard function was generated from the exponential distribution and the Weibull distribution. A nonparametric method was also employed to mimic a case where the data were generated from Cox regression; this is known as the flexible-hazard method. For the non-censored cases, uniform distribution is used to generate the left and right bound times L and R. If the left and right bound times are not the same, the case is regarded as interval-censored. The survival time was generated assuming that the true regression coefficient was $\boldsymbol{\beta} = (0.058, 0.045, -0.876, -0.052)$, which can be obtained by repeating 100 times for the MIICD package to which the multiple imputation method was applied.

Table 1 MSE values are presented when the exponential distribution is assumed

right-censoring	20%								
Sample size	100			300			1000		
interval-censoring	15%	30%	45%	15%	30%	45%	15%	30%	45%
omission	0.444	0.432	0.473	0.195	0.216	0.190	0.088	0.095	0.105
midpoint imputation	0.337	0.309	0.317	0.184	0.191	0.157	0.080	0.080	0.080
multiple imputation	0.333	0.305	0.305	0.188	0.187	0.154	0.080	0.080	0.080

4.2 Simulation results

The MSE values are summarized in Table 1 where the exponential distribution is assumed. When the baseline hazard function follows an exponential distribution, the MSE tends to decrease as the sample size increases, regardless of the interval-censoring rate. Furthermore, for a given sample size and the interval-censoring rate, the MSE of the multiple imputation method is slightly lower than that of the midpoint imputation and omission method for the sample sizes of 100 and 300. In the case of sample sizes of 1000, there is no difference between the MSE of the multiple imputation and the midpoint imputation, and the MSEs of the midpoint imputation and the multiple imputation MSE are lower than that of the omission.

Table 2 MSE when the Weibull distribution is assumed

right-censoring	20%								
Sample size	100			300			1000		
interval-censoring	15%	30%	45%	15%	30%	45%	15%	30%	45%
omission	0.323	0.341	0.359	0.223	0.234	0.239	0.075	0.081	0.090
midpoint imputation	0.293	0.283	0.278	0.221	0.225	0.218	0.071	0.071	0.070
multiple imputation	0.295	0.283	0.276	0.218	0.222	0.216	0.071	0.071	0.071

The simulation results are summarized in Table 2 where the right-censoring rate is set as 20%, and Weibull distribution is assumed for the baseline hazard function. As in the case of the exponential distribution, the MSE value tends to decrease as the sample increases, regardless of the interval-censoring rate. Furthermore, for a given sample size and the interval-censoring rate, the MSE of the multiple imputation is lower than that of the midpoint imputation when the sample size is 100 and 300. The MSE of the midpoint imputation is lower than that of the multiple imputation and omission for some cases. For a sample size of 1000, there is little difference between the MSE of the multiple imputation and the midpoint imputation, and in some cases, the MSE of the midpoint imputation is lower than that of the multiple imputation. The iAUC was low regardless of the right-censoring rate and the interval-censoring rate, and iAUC decreased slightly as the sample size increased.

The simulation results are summarized in Table 3 where the right-censoring rate is fixed at 20%, and Weibull distribution is assumed for the baseline hazard function. As in the case of assuming Weibull distribution, the MSE value tends to decrease as the sample is enlarged, regardless of the interval-censoring rate. Furthermore, when the sample size and the interval-censoring rates are the same, the MSE of the midpoint imputation is lower than that of the multiple imputation if the sample size reaches 100 and 300. For a sample size of 1000, there is little difference between the MSE of the multiple imputation and the midpoint imputation, and in some cases, the multiple imputation MSE is lower than that of the midpoint imputation.

Table 3 MSE when the flexible-hazard method is assumed

right-censoring	20%								
Sample size	100			300			1000		
interval-censoring	15%	30%	45%	15%	30%	45%	15%	30%	45%
omission	0.740	0.775	0.952	0.249	0.313	0.406	0.061	0.075	0.090
midpoint imputation	0.694	0.710	0.788	0.233	0.212	0.220	0.056	0.055	0.057
multiple imputation	0.689	0.683	0.726	0.231	0.212	0.216	0.056	0.055	0.055

The simulation results are summarized in Table 3 where the right-censoring rate is fixed at 20%, and the flexible-hazard method is used for determining the baseline hazard probability. As in the case of assuming the flexible-hazard method, the MSE value tends to decrease as the sample is enlarged, regardless of the interval-censoring rate. Furthermore, when the sample size and the interval-censoring rates are the same, the MSE of the multiple imputation is lower than that of the midpoint imputation if the sample size is 100, 300, and 1000. In some cases, the midpoint imputation MSE is lower than that of the multiple imputation.

The multiple imputation method showed a lower MSE than the midpoint imputation and omission when the sample size was 100, and the right-censoring rate was fixed at 20%. However, sample sizes of 300 and 1000 showed similar MSEs to the midpoint imputation, regardless of the interval-censoring rate. Midpoint imputation is affected by the sample size, so the MSE decreases as the sample gets larger; however, the multiple imputation method shows a low residual regardless of the sample size, so it is a good method for imputing the interval-censoring with a robust model. Therefore, we will proceed to use the Cox proportional hazards model with a multiple imputation method, which shows a better model estimation accuracy.

5 Data analysis

5.1 Data

The data used for the analysis were based on 20 years of panel data, which were collected from surveys conducted between 1998 and 2017 as part of the Korea Labor Panel Survey (KLIPS). We analyzed the effect of parental income loss on children's college entrance. Since the Korean Labor Panel Survey is conducted in non-rural areas, the survey data were limited in scope for application of analysis results to rural areas. Of the 989 subjects, 58 (5.9%) were right-censored, and 79 (7.9%) were interval-censored.

Table 4 Variables and their descriptive statistics

		Frequency	Proportion (%)
Education level of householder	middle school graduation or less(1)	153	15.5%
	high school graduation or less(2)	458	46.3%
	college graduation or less(3)	378	38.2%
Sex of the first child	Male(0)	500	50.6%
	Female(1)	489	49.4%
Poverty	No(0)	926	93.6%
	Yes(1)	63	6.4%
Whether parents are unemployed	No(0)	963	97.4%
	Yes(1)	26	2.6%
Double income	No(0)	109	11.0%
	Yes(1)	880	89.0%
The number of household members	2	19	1.9%
	3	130	13.1%
	4	635	64.2%
	5	173	17.5%
	6	32	3.2%
Censoring	right-censoring	58	5.9%
	interval-censoring	79	7.9%
	No censoring	852	86.2%
Total		989	100%

Among the variables of KLIPS, the householder's education level, gender of the first child, gender of the householder, poverty, employment status of the first child's parents, double-income status of the first child's parents, and the number of household members were selected by as covariates based on the study by [4]. The descriptive statistics of the chosen variables are summarized in Table 4. However, the Fisher Exact test results showed that the correlation

Table 5 Comparison of model estimation for the sample whose householder graduated high school or less.

	high school graduation or less											
	omission n= 576, number of events= 524				midpoint imputation n= 610, number of events= 558				multiple imputation n= 610, number of events= 558			
	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value
Sex	0.108	1.114	0.108	0.316	0.074	1.077	0.086	0.349	0.055	1.056	0.113	0.628
Double income	0.285	1.330	0.186	0.125	0.138	1.148	0.148	0.370	0.068	1.071	0.195	0.727
Whether parents are unemployed	-0.787	0.455	0.384	0.040*	0.186	1.204	0.254	0.330	-0.877	0.416	0.391	0.025*
The number of household members	-0.091	0.913	0.085	0.283	-0.043	0.958	0.063	0.409	-0.059	0.943	0.088	0.502

between household poverty and parental unemployment experience was significantly high (p-value=0.025); furthermore, the poverty variable was excluded from the real data analysis. Assuming that the effect of parental unemployment may vary depending on the household head's academic background, we divided the sample into two subsets: a sample where the household head's education level included achievement of a high school diploma or a lower qualification and a sample where the household head's education level included achievement of a college degree or a higher qualification.

5.2 Analysis results

In table 5, model estimation results were summarized for the sample where household heads' education levels included qualifications under or equal to achievement of a high school diploma. The first child's probability of being admitted to a college under circumstances of parental unemployment was 58.4% lower than that in the other cases. The variables of parental unemployment were significant in the case of omission and multiple imputation at a significance level of 5%. The midpoint imputation produced an opposite result in terms of the direction of the effect of the parental unemployment variable. Table 6 shows model estimation results for the sample where household heads' education levels included qualifications that were below or equal to achievement of a high school diploma. The first child's probability of being admitted to a college under circumstances of parental unemployment was 57.5% lower than that in the other cases. The variables of parental unemployment were significant in all the imputation methods at a significance level of 5%. When the interval-censored data were omitted, in the case of double-income households, the probability of being admitted to a college was 46% higher than that in the other cases. This shows that use of an inappropriate imputation method for

Table 6 Comparison of significance of regression coefficients for the households whose householder graduated college or more.

	college graduation or more											
	omission n= 333, number of events= 315				midpoint imputation n= 376, number of events= 357				multiple imputation n= 376, number of events= 357			
	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	p-value
Sex	0.114	1.120	0.108	0.292	0.108	1.114	0.108	0.316	0.070	1.073	0.112	0.530
Double income	0.381	1.464	0.188	0.043*	0.285	1.330	0.186	0.125	0.031	1.032	0.213	0.883
Whether parents are unemployed	-0.753	0.471	0.383	0.049*	-0.787	0.455	0.384	0.040*	-0.857	0.425	0.388	0.027*
The number of household members	-0.119	0.888	0.888	0.176	-0.091	0.913	0.085	0.283	-0.040	0.961	0.088	0.649

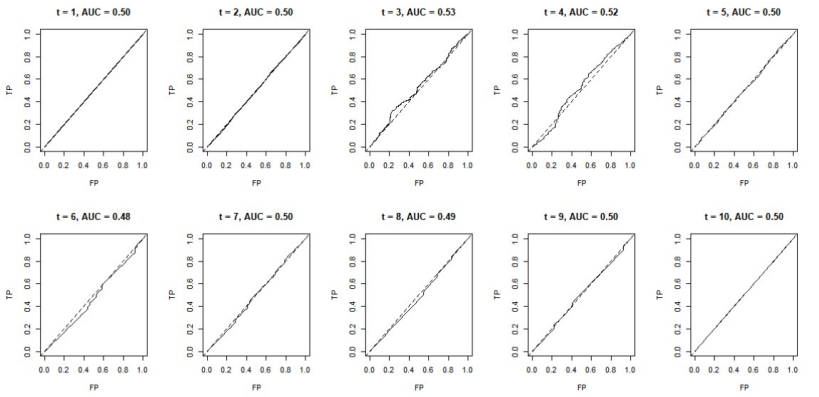
the interval-censoring (such as simple omission or midpoint imputation) could distort the data analysis results.

5.3 Comparison of predictive performance according to the imputation method

We used the time dependent receiver operating characteristic (ROC) curve to evaluate the predictive power of survival data instead of a simple ROC curve, which is used for evaluating the predictive power of binomial response variables. The area AUC (t) under the time-dependent ROC curve can be calculated at each time point t. The integrated AUC (iAUC) was used to compare the prediction performance of statistical methods for interval censoring. The closer iAUC is to 1, the better the model is; the closer it is to 0.5, the less accurate the model is. The iAUC of the sample where household heads had educational qualifications that were below or equal to the achievement of a high school diploma was estimated as 0.51 based on the 5-folds cross-validation. The iAUC of the sample where household heads had educational qualifications that were equal to or higher than the achievement of a college degree was estimated as 0.53 based on the 5-folds cross-validation.

This result implies that the predictive performance of the Cox regression model was very poor. Thus, this estimated model is limited to only interpretation not prediction.

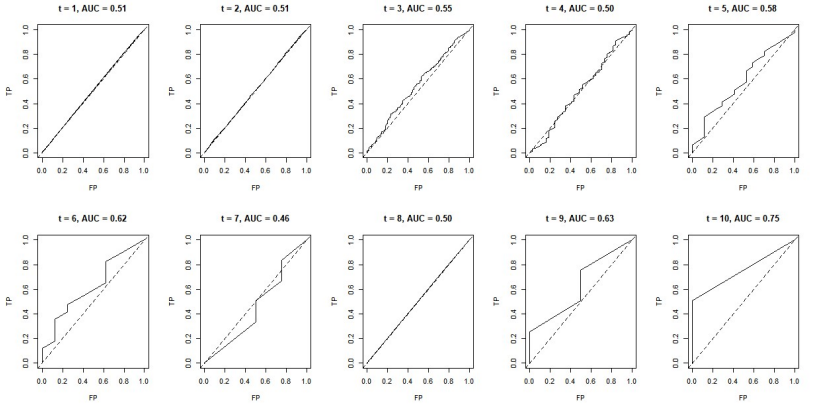
Fig. 1 AUC curves at a curtain time point t when interval censored data are imputed by the multiple imputation method



6 Conclusion

Using 20 years' worth of data (1998 to 2017) from the Korean Labor Panel Data (KLIPS), we analyzed how much college admissions could be affected by parental unemployment status when a first child in the household was

Fig. 2 AUC curves at a curtain time point t when interval censored data are imputed by the multiple imputation method



aged 18 years and preparing for college. Assume that the child is admitted to college in 2018, but the household answered the survey in 2019. In this case, the panel data produced an interval censoring, since the researcher is unaware of the exact time of admission due to lack of response. We considered three imputation methods for the interval censoring: simple omission, midpoint imputation, and the multiple imputation proposed by Pan. In order to choose an appropriate imputation method for this data, we ran extensive simulation studies. Mean Squared Errors (MSEs) were compared in order to evaluate the performance of the imputation methods. For the simulation study, 100, 300, and 1000 samples were re-sampled with replacement from the real data of 376 subjects whose parental education qualifications included college graduation or a higher qualification. The right-censoring rate was set as 0.2, and the interval censoring rate varied from 0.15 to 0.45. Overall, the model estimation accuracy of the multiple imputation method was found to be higher than that of other imputation methods. The estimation accuracy midpoint imputation was affected by the sample size, so the MSE decreased as the sample grew larger. On the other hand, the multiple imputation method showed a low residual regardless of the sample size. Therefore, we can conclude that multiple imputation is a good method for ensuring robust model estimation.

The real data analysis showed that the effect of the variable "whether or not the parents are unemployed" on the time taken to be admitted to a college was significant only when the householder's academic background was higher than and equal to college graduation. When the interval-censored data were removed, the double income and parents' unemployment variables became significant. First children of double-income parental households had a 46% higher probability of entering a college than others. On the contrary, when the first children experienced their parents' unemployment at the age of 18, the probability of college admission was reduced nearly by 53% compared to cases where they did not. Therefore, our study suggests that college entrance

is affected by parental financial status—in particular for households where the householder's academic background is higher than and equal to college graduation. Therefore, this study suggests that a policy targeting adolescents who cannot expect parental help be developed.

Author Contributions. Conceptualization, H.K., E.L.; methodology, H.K., E.L.; software, H.K.; formal analysis, H.K., E.L.; data curation, H.K.; writing—original draft preparation, H.K., E.L., and S.K.; writing—review and editing, H.K., E.L., and S.K.; supervision, E.L.. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding. This study was funded by 2019 research fund of Chungnam National University

Informed Consent Statement. Informed consent was obtained from all subjects involved in the study

Data Availability Statement. The data is available in the following website: https://www.kli.re.kr/klips$_e ng/index.do$.

Conflicts of Interest. The authors declare no conflict of interest.

References

[1] Conger, R.D., Conger, K.J., Elder Jr, G.H., Lorenz, F.O., Simons, R.L., Whitbeck, L.B.: A family process model of economic hardship and adjustment of early adolescent boys. Child development **63**(3), 526–541 (1992)

[2] Coelli, M.B.: Parental job loss and the education enrollment of youth. Labour Economics **18**(1), 25–35 (2011)

[3] Pan, W., Ost, B.: The impact of parental layoff on higher education investment. Economics of Education Review **42**, 53–63 (2014)

[4] 구인회: 경제적 상실과 소득수준이 청소년의 교육성취에 미치는 영향. 한국사회복지학 **53**, 7–30 (2003)

[5] OECD: Education at a Glance 2020, p. 476 (2020). https://www.oecd-ilibrary.org/content/publication/69096873-en

[6] OECD: Population with tertiary education (2020)

[7] Turnbull, B.W.: The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological) **38**(3), 290–295 (1976)

[8] Groeneboom, P., Wellner, J.A.: Information Bounds and Nonparametric Maximum Likelihood Estimation vol. 19. Springer, ??? (1992)

[9] Wellner, J.A., Zhan, Y.: A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. Journal of the American Statistical Association **92**(439), 945–959 (1997)

[10] Finkelstein, D.M.: A proportional hazards model for interval-censored failure time data. Biometrics, 845–854 (1986)

[11] Pan, W.: A multiple imputation approach to cox regression with interval-censored data. Biometrics **56**(1), 199–203 (2000)

[12] Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. Wiley, ??? (1987). https://books.google.co.kr/books?id=0KruAAAAMAAJ

[13] Delord, M., Génin, E.: Multiple imputation for competing risks regression with interval-censored data. algorithms **11**(18), 22 (2015)