Communications for Statistical Applications and Methods
2021, Vol. 28, No. 5, 447–462

https://doi.org/10.29220/CSAM.2021.28.5.447
Print ISSN 2287-7843 / Online ISSN 2383-4757

Household, personal, and financial determinants of surrender in Korean health insurance

Hyunoo Shim[a], Jung Yeun Min[b], Yang Ho Choi[1,a]

[a]Department of Actuarial Science, Hanyang University, Korea;
[b]Department of Finance and Insurance, Hanyang University, Korea

Abstract

In insurance, the surrender rate is an important variable that threatens the sustainability of insurers and determines the profitability of the contract. Unlike other actuarial assumptions that determine the cash flow of an insurance contract, however, it is characterized by endogenous variables such as people's economic, social, and subjective decisions. Therefore, a microscopic approach is required to identify and analyze the factors that determine the lapse rate. Specifically, micro-level characteristics including the individual, demographic, microeconomic, and household characteristics of policyholders are necessary for the analysis. In this study, we select panel survey data of Korean Retirement Income Study (KReIS) with many diverse dimensions to determine which variables have a decisive effect on the lapse and apply the lasso regularized regression model to analyze it empirically. As the data contain many missing values, they are imputed using the random forest method. Among the household variables, we find that the non-existence of old dependents, the existence of young dependents, and employed family members increase the surrender rate. Among the individual variables, divorce, non-urban residential areas, apartment type of housing, non-ownership of homes, and bad relationship with siblings increase the lapse rate. Finally, among the financial variables, low income, low expenditure, the existence of children that incur child care expenditure, not expecting to bequest from spouse, not holding public health insurance, and expecting to benefit from a retirement pension increase the lapse rate. Some of these findings are consistent with those in the literature.

Keywords: surrender, lapse rate, determinants, lasso regularized regression, panel survey data, random forest missing value imputation

1. Introduction

Insurance is a means of financially transferring future risks to the insurer. Life or health insurance is a security in return for the contractor's payment of insurance premiums, such as for death, illness, or injury. Specifically, it is a contract that promises to receive benefit from an insurer in the event of a risk event, in return for the contractor's payment of insurance premiums. Such a contract is effective from the commencement (subscription) of the contract to its termination (maturity, death, or lapse). Unlike manufacturing goods or service contracts, typical life insurance contracts are characterized by long-term features, where long-term means the long period from contract inception to termination.

The initiation of an insurance policy is caused by demand for insurance. Traditionally, many studies have examined the factors that affect insurance demand. Empirical studies of the driving factors of insurance demand include Anderson and Nevin (1975); Berekson (1972); Burnett and Palmer (1984);

[1] Corresponding author: Department of Actuarial Science, Hanyang University-ERICA Campus, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, Gyeonggi-do 15588, Republic of Korea. E-mail: ychoi@hanyang.ac.kr

Published 30 September 2021 / journal homepage: http://csam.or.kr

Duker (1969); Ferber and Lee (1980); Fitzgerald (1989); Gandolfi and Miners (1996); Hammond,*et al.* (1967). On the contrary, research on risk aversion and insurance demand is evolving in empirical (Greene, 1963) and theoretical studies (Chesney and Loubergé, 1986; Cook and Graham, 1977; Hanoch and Levy, 1969). There is also research on the effect of life settlement on insurance demand (Hong, 2020).

The termination of a contract is broadly divided into three types, when the contract reaches its maturity specified in the policy, when the policyholder dies before its maturity date, and when the policyholder surrenders the policy or the contract lapses before the maturity date. The notions of surrender and lapse are similar in that they lead to termination before the contract period ends; however, technically speaking, surrender is voluntary and lapse is involuntary. Therefore, in this study, we use the two terms interchangeably unless confusion would arise. Unlike maturity, surrender is driven by the actions of the policyholder; thus, it is uncertain.

Studies of insurance surrender (and lapse), while arriving later than the above studies of insurance demand, have recently been active (Eling and Kochanski, 2013). It is necessary to understand and accurately predict surrendering insurance for a number of reasons. First, an unpredictable amount of lapse in group policies poses several risks to the insurer that has underwritten the policy, such as excess loss risk and liquidity risk. Unpredictable lapse results in a discrepancy in the timing between the income of insurance premiums and expenditure of insurance claims (i.e. cash flow uncertainty). Accordingly, it leads to risk management difficulties for insurers. Consequently, an insurer must manage the lapse of contracts after they are underwritten. Second, the surrender rate affects the fair value measurement of insurance contracts, and thus it is also an integral factor in assessing insurance liabilities in accordance with insurance supervision regulations. Although the insurer may not be proactive in managing lapses of contracts, its risk is subject to mandatory assessment and surveillance under IFRS17, an insurance supervisory regulation. Among previous studies of insurance contracts focused on the effect of lapse rates, Albizzati and Geman (1994); Bacinello (2003a,b); Grosen and Jørgensen (2000) assessed the value of life insurance contracts with surrender options embedded, Bacinello (2005) investigated unit-linked contracts and Linnemann (2003) calculated the value of participating contracts numerically. In summary, the accurate prediction of lapse behavior plays an essential role in the stable management of an insurance business.

A statistical distribution or model of the surrender rate is a necessary actuarial assumption for the fair value evaluation and risk assessment of insurance contracts. Predicting this rate has long been an important goal in actuarial science. Methods for predicting a policyholder's surrender behavior include a financial engineering model based on the rational expectations hypothesis (Giovanni, 2010), a stochastic process model based on optimal intervention (Steffensen, 2002), an income shock model that includes the emergency fund hypothesis (Outreville, 1990), policy replacement models, macroeconomic models based on the interest rate hypothesis (Russell *et al.*, 2013; Schott, 1971), a statistical model with economic variables (Kim, 2005), a generalized linear model (Bajaj, 2017), a structural model (Bauer *et al.*, 2017) a behavioral economic model (Shefrin, 2002), and a copula-based dependence model (Neves *et al.*, 2014). Since the longitudinal data in this study are short, it is difficult to test the interest rate hypothesis, which proposes that the policyholder's surrender decisions are affected by the time series of macroeconomic interest rates. Furthermore, it is difficult to test Linton's emergency fund hypothesis (Linton, 1932) that the policyholder's surrender is a function of economic pressure, since there is no significant economic pressure (e.g. economic crisis) during the data observation period.

Instead, in this study, we aim to find the surrender determinants of Korean health insurance overlooked by previous studies by analyzing the effect of micro-level variables such as household and

individual variables. The data set in this study is based on panel data from Korean Retirement and Income Study (KReIS). Owing to the low surrender rate, the number of surrendered cases in the samples is small and, even worse, there are many missing values in the panel data. Hence, we apply an algorithm to replace the missing values with imputed values to prevent the loss of information when we remove rare surrender cases. As the number of variables is high, we analyze the lapse rate using a regression model with lasso penalization applied to extract the important factors and eliminate the relatively less important variables.

The remainder of this paper is organized as follows. Section 2 discusses the ordinary logistic regression model and lasso penalized regression model of surrender. Section 3 explains the data used in this study including the data preprocessing stage. Specifically, Section 3.1 describes the descriptive statistics of the panel survey data and Section 3.2 describes the treatment applied to the missing values frequently observed in this type of data. Section 4.1 presents the analysis results and Section 4.2 compares them with those of preceding work in the literature. Finally, Section 5 summarizes the results and findings and discusses future research directions.

2. Regression with lasso regularization

This work is based on a regression model with a large number of variables. Thus, it is important to choose which variables are significant, whether before or during model construction. Bajaj (2017) analyzed the determinants that affect the decision to surrender by applying variable selection methods such as forward and backward variable selection in a generalized linear model. However, the samples we use are a type of panel data with many variables, so there is a limitation to using the variable selection method. First, forward or backward variable selection may yield R-squared values that are overbiased. Second, Altman and Anderson (1989) argued that this may generate incorrectly narrow confidence intervals for the predicted values. Third, Tibshirani (1996) showed that this could result in biased regression coefficients, where the other regression coefficients remain high. In addition, using such a method does not properly select collinear explanatory variables. Derksen and Keselman (1992) revealed that data with large samples do not resolve the above issues.

The method of best subset model selection, on the other hand, seems ideal, but it has an interpretation problem when a model with more variables is not always a subset of a model with fewer variables (Judd *et al.*, 2008). It also has a drawback that it is not applicable to data with a large number of candidate variables, such as the panel data used in the current study. To overcome such limitations, we construct an elastic-net regularized regression model that allows us to perform regression analysis while simultaneously selecting the variables.

First, we describe our ordinary logistic regression (OLR). The random response variable for surrender, W, can take binary values, $W = \{0, 1\}$, where 0, 1 indicates no surrender and surrender, respectively. First, we collect observations for this random variable. We define an indicator variable of y_i such that $y_i = 1$ when the observed value of W for the i^{th} person, w_i is equal to 1: $y_i = I(w_i = 1)$.

Then, we construct a model that predicts the probability of surrender using explanatory variables X as follows,

$$\Pr(W = 1|X = x) = \frac{e^{\beta_0+\beta^T x}}{1 + e^{\beta_0+\beta^T x}}, \tag{2.1}$$

where X indicates a column vector of the explanatory variables, x is an observed value of the random variables X, β_0 is an intercept, and β^T is a row vector of the coefficients of X. Because lapse decisions

are made at discrete times, this probability of surrender can be approximated as follows,

$$w(x;t) \equiv \Pr(W = 1|X = x) \simeq \frac{l(x;t) - l(x;t-1)}{l(x;t-1)}, \tag{2.2}$$

where $w(x;t)$ is the surrender rate with $X = x$ at time t and $l(x;t)$ is the number of policyholders with $X = x$ at time t. If we perform a logistic transformation to equation 2.1, we can obtain the following form,

$$\log \frac{\Pr(W = 1|X = x)}{1 - \Pr(W = 1|X = x)} = \log \frac{\Pr(W = 1|X = x)}{\Pr(W = 0|X = x)} = \beta_0 + \beta^T x. \tag{2.3}$$

This is the OLR for surrender. The objective function for the OLR is as follows,

$$\min_{(\beta_0,\ \beta^T)\in\mathbb{R}^{p+1}} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i\right)^2, \tag{2.4}$$

for the observations of the explanatory variables, $x_i \in \mathbb{R}^p$, and the observed responses, $y_i \in \mathbb{R}$, for $i = 1, \ldots, N$, where p is the number of variables and N is the number of observations. Penalized regression adds a penalty to the above objective function, which constrains the increase in the number of variables to reduce errors. The objective function of the penalized logistic regression (PLR) is,

$$\min_{(\beta_0,\ \beta^T)\in\mathbb{R}^{p+1}} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i\right)^2 + \lambda \left[(1-\alpha)\frac{\left\|\beta\right\|_2^2}{2} + \alpha\|\beta\|_1\right]. \tag{2.5}$$

In equation 2.5, $0 \le \alpha \le 1$. If the fitted model has $\alpha = 0$, it reduces to ridge logistic regression; if it has $\alpha = 1$, it becomes lasso logistic regression (LLR). The λ in the above objection function is a control parameter that adjusts the regularization penalty. In the λ sequence, it is common to select either $\lambda_{\min}$ to minimize binomial deviance or λ_{1se} that falls by one standard error from $\lambda_{\min}$. Krstajic *et al.* (2014) maintained that 'the main point of the 1 SE rule, with which we agree, is to choose the simplest model whose accuracy is comparable with the best model'. In this study, we choose the α values as follows. First, we select α by increasing α in 0.1 intervals between zero and one and fit the model for each α value using the following 10-fold cross-validation. We split the data set into 10 different subsets, using nine subsets as the training data and the last subset as the testing data. Then we perform the cross-validation to fit the model. Next, we measure the binary deviance of the fitted model. We run the above procedure for each α and select a pair of α and $\lambda_{1\text{se}}$, that minimizes the binomial deviations. For the 10-fold cross-validation, the fold of samples is randomly selected at the beginning but fixed in the later cross-validation.

3. Data and data preprocessing

3.1. Data

This study used KReIS panel survey data. The KReIS survey was conducted every two years from 2005 to 2017, with respondents comprising household members 50 years and older and their spouses in Korea national pension service (2019). Questions about whether respondents held private health insurance and surrendered it have been asked since the fourth survey in 2011. The rates of response to the questions, however, are significantly lower. Moreover, since the fifth survey in 2013, a significant

number of new households have been introduced into the sample, resulting in a different configuration of samples. Therefore, this study uses data from the fifth survey in 2013 to the seventh survey in 2017.

The survey asks respondents about their insurance and pension policies. Here, pensions are divided into public and private pensions, and insurance is divided into national and private health insurance (if there is no confusion, 'private health insurance' is called simply 'insurance' in the following). However, there are no questions about surrendering pensions and national health insurance, partly because national health insurance in Korea is a compulsory public insurance that cannot be withdrawn. Hence, only those who signed up for private health insurance are included in our investigation.

To analyze the driving factors of surrender, it is necessary to extract the policyholders of the previous survey from all respondents and separate withdrawn policies from in-force policies of the current survey. This requires the consideration of three points. First, the response to the surrender question differs depending on whether the respondent held an insurance policy at the time of the survey. In other words, only those respondents without insurance responded to this question (i.e. no person who held insurance at the time of the survey answered the surrender question). Second, in this study, we assume that the decision to surrender is made at the time of the survey, since there is no additional information. The surrender question of the survey only asks whether the respondent decided to surrender in the past; hence when the surrender decision was made is unknown. Therefore, it is not known whether the respondent surrendered between the previous survey and the current survey or before that. Nevertheless, temporal proximity between the explanatory variables and surrender decision is needed to explain the causality of the surrender process. Third, surrendering and other related states of policyholders are measured at discrete times as the surveys are collected even though these events occur continuously over time in practice.

Overall, keeping these three points in mind, we distinguish between in-force policies and withdrawn policies in the following ways. We regard as a surrendered policy as one withdrawn at the time of the survey (A) that was taken out by policyholders who answered that they had purchased or owned the insurance in the previous survey (B). The remainder (B–A) are considered to be those who had not surrendered their insurance in the current survey. Those who did not hold an insurance policy in the previous survey are excluded from the analysis because they were deemed not to have held an insurance contract during that period.

The panel data are divided into household and individual variables. Although the decision to withdraw differs by household member and insurance policy, the survey does not ask for individuals' surrender experience for each insurance policy. Thus, we set each household member as an observation unit. Since the survey is conducted every two year, the maximum error of the time of surrender is two years. The people at advanced ages are at high risk with regard to retirement in their life cycles. The data are characterized by the inclusion of questions related to retirement, not found in any other source.

Table 1 shows the proportion of insured people to the total number of respondents in each survey. As shown in the table, approximately $30 \sim 40\%$ of people surveyed are found to have maintained their health insurance. Table 2 shows a contingency table by gender, age group, and policy status. Table 2 shows that the survey respondents are mainly 50 to 79 years old. The proportion of health insurance policyholders does not differ by gender, whereas it decreases as age increases.

Only the insured persons in Table 1 are subject to our analysis. In each subsequent survey period, their contracts are further divided into three types: in-force policies, policies terminated by death, and surrendered policies. Since insureds are not surveyed after death, we eliminated those dead insureds from the beginning of survey period in our analysis and show the number of in-force policies and surrendered policies in Table 3. Table 3 also presents the rate at which people determine the surrender

Table 1: Number of insured people, uninsured people, total respondents, and their proportions to total

Survey	# of insured people[a]	# of uninsured people[a]	# of total respondents
5th	3,172 (37.7%)	5,239 (62.3%)	8,411 (100%)
6th	2,746 (34.4%)	5,237 (65.6%)	7,983 (100%)
7th	2,842 (37.5%)	4,730 (62.5%)	7,572 (100%)

[a] Figures within parentheses indicate the proportion to total respondents of each survey.

Table 2: Contingency table by gender, age group, and policy status

Gender	Age	# of insureds[a]	# of uninsureds[a]	Subtotal[b]
Male	0–49	15(88.2%)	2(11.8%)	17(0.2%)
	50–59	2,454(68.2%)	1,143(31.8%)	3,597(36.1%)
	60–69	1,054(38.9%)	1,657(61.1%)	2,711(27.2%)
	70–79	277(10.3%)	2,417(89.7%)	2,694(27.1%)
	80–	18(1.9%)	918(98.1%)	936(9.4%)
	Subtotal	3,818(38.4%)	6,137(61.6%)	9,955(100%)
Female	0–49	721(77.5%)	209(22.5%)	930(6.6%)
	50–59	2,677(70.0%)	1,145(30.0%)	3,822(27.3%)
	60–69	1,239(34.2%)	2,386(65.8%)	3,625(25.9%)
	70–79	2,86(7.3%)	3,631(92.7%)	3,917(28.0%)
	80–	19(1.1%)	1,698 (98.9%)	1,717(12.3%)
	Subtotal	4,942(35.3%)	9,069(64.7%)	14,011(100.0%)

[a] Figures within parentheses indicate the proportion to the subtotal count of both insureds and uninsured differed by gender and age group.
[b] Figures within parentheses indicate the proportion to the subtotal count of respondents differed by gender.

Table 3: Number of in-force policies (policyholders) and withdrawn policies (non-policyholders) in insurance

Survey time (t)	# of in-force policies at $t-1$[a,b]	# of in-force policies at t[a,c]	# of withdrawn policies at t[a,d]
6	2,936 (100%)	2,820 (96.0%)	116 (4.0%)
7	2,587 (100%)	2,452 (94.8%)	135 (5.2%)
Pooled data	5,523 (100%)	5,272 (95.5%)	251 (4.5%)

[a] Figures within parentheses indicate the proportion to the number of policyholders at $t-1$.
[b] $l(x;t-1)$
[c] $l(x;t)$
[d] $l(x;t-1)-l(x;t)$

Table 4: Descriptive statistics of missing values

Survey	# of all cases	# of complete cases	Proportion of complete cases to all cases	Proportion of missing attributes to all attributes
5th	8,411	7,523	0.89442	0.004731
6th	7,983	6,159	0.77151	0.004343
7th	7,572	6,555	0.86569	0.003630

during the transition period of $(t, t+1)$, that is, $w(x;t)$. The average surrender rate is 251/5,523=4.5%. Thus, the samples are unbalanced data in which there are fewer withdrawn policies than in-force policies.

We prepared for pooled data on insured people between the fifth and sixth surveys and between the sixth and seventh surveys. The number of samples is $N = 5,523$. The KReIS panel data include 1,322 variables in the fifth survey 1,279 in the sixth survey and 1,311 in the seventh survey. Among them, the numbers of mandatory questions, except for the unanswered questions that depended on a particular

question item, are 249 in the fifth survey 253 in the sixth survey, and 250 in the seventh survey. Among them, the number of common explanatory variables for the sixth and seventh surveys is 246, excluding the question about the surrender decision. We exclude one question (code 'p00h058') that was not sufficiently answered in the fifth survey (a variable with a missing value of 10% or more of the total number of cases). Some categorical variables have too many factor levels that have insufficient observations ($n \leq 1$). These logical factor levels are excluded from the data because it is impossible to predict them using the binary variable with insufficient data. Among the categorical variables, life satisfaction questions exist, which are answered on a five-point Likert scale.

This work uses the 'glmnet' R package for the elastic-net regularization regression. We investigate all the ranges of α by increasing the number of steps from $\alpha = 0$ to $\alpha = 1$ by 0.1. As a result, the optimal α with the lowest binomial deviance is found to be $\alpha = 1.0$. Hence, the lasso regularization method is applied, as explained below.

3.2. Missing value imputation

The existence of conditional questions in the panel data makes it almost impossible to find a case that answered all questions fully. Although the behavior for surrender may depend on special questions, our work aims to analyze the entire group rather than predict the behavior of a subgroup. Therefore, the data we use are composed of a set of explanatory variables based on the mandatory survey questions.

Missing values of panel data are caused either by mistakes by surveyors or by respondents' refusal to answer. The unofficial dataset internally managed by KReIS are divided into 'unknown', 'refused to answer', and 'not measured'. However, the official dataset only shows missing data as 'non-response'. Accordingly, all missing values are treated in the same way. As shown in Table 4, missing data represent 10.6%, 22.8%, and 13.4% of the responses to the fifth, sixth, and seventh surveys, respectively. Owing to the small number of all surrender cases (251), treating missing values is important.

The methods used to process incomplete observations (i.e. where missing values for certain variables exist) can be classified into elimination and imputation methods, etc. Elimination methods include listwise deletion and pairwise deletion, whereas imputation methods comprise explicit and implicit models. Among explicit models, there are single imputation methods such as mean imputation, median imputation, and mode imputation. Among explicit models, multiple imputation also exits, methods such as the last observation carried forward, ratio imputation, Buck's method, regression method, stochastic regression method, k-nearest neighborhood (kNN) method, and neural network method. Implicit models include hot-deck imputation and cold-deck imputation.

These methods have several limitations. Listwise deletion is simple but results in information loss. Listwise deletion and single imputation make the strong statistical assumption of missing completely at random (MCAR) among missing value imputation methods, which is difficult to apply when many variables exist, such as in this work. If the MCAR assumption does not hold, the estimator may be biased. The kNN method developed by Troyanskaya *et al.* (2001) can replace missing values for continuous data. However, our data contain categorical variables, and it is difficult to apply a regression method or a kNN method without using a modified distance measure.

Considering this, this work replaces missing values with imputed values using the random forest method, which is non-parametric machine learning imputation method. To the best of our knowledge, Stekhoven and Bühlmann (2012) was the first to suggest this algorithm. The random forest imputation method (Breiman, 2001) has the advantage of being applicable to both numeric and categorical data. It is an iterative imputation method in that it trains a model using the actual observations of a particular variable by the random forest method and then predicts and replaces the missing values. It performs

Table 5: Results of OLR and LLR models for policyholder's surrender

Type[a]	Characteristics	Variable (i)	OLR		LLR
			β_i (s.e.)[b,c]	p-value	β_i
		Intercept	–0.352 (0.322)	0.275	–2.136
FLC	Old dependents	the number of household members aged 65+			-0.021
	Young dependents	Household composition (married couple, siblings, and children)	4.005*** (1.002)	6.40×10^{-5}	2.335
		Household composition (household head and grandchildren)	1.975* (1.004)	0.049	0.579
	Unemployed family	the number of unemployed family members	–0.833** (0.274)	0.002	–0.101
LC	Marital status	Marital status (divorce)			0.026
RES	Residence	Residence (Seoul)	–1.150*** (0.257)	7.61×10^{-6}	–0.202
		Residence (province)	–0.254· (0.146)	0.082	
	Type of housing	Type of housing (apartment)			0.036
	Home-ownership	Home-ownership (self-owned)	–0.645*** (0.158)	4.54×10^{-5}	–0.265
		Home-ownership (rent with deposit)			0.120
IC	Income	Private transfer income	-2.40×10^{-4}*** (5.69×10^{-5})	2.41×10^{-5}	-1×10^{-5}
EXP	Expenditure	Other expenditure	-4.23×10^{-4}*** (4.40×10^{-5})	$< 10^{-16}$	-2×10^{-4}
	Expenditure not for dependents	Non-consumption expenditure (living subsidy excluding remittance to children)	3.99×10^{-5}* (2.01×10^{-5})	0.047	
	Expenditure for dependents	the number of children that incur childcare expenditure	1.020** (0.312)	0.001	0.024
AST	Asset	Other assets	1.23×10^{-5}*** (3.71×10^{-6})	0.001	
	Inheritance	Expected bequest/gift from spouse (yes)	–0.663** (0.233)	0.004	-0.111
		Expected bequest/gift from spouse (no spouse)			0.116
INS	Public insurance	Subscribed to national health insurance (no)	0.646· (0.375)	0.085	0.538
	Retirement annuity	Expected benefit of private retirement pension (yes)	1.039** (0.367)	0.005	0.139
SAT	Satisfaction	Relationship with siblings	–0.243** (0.082)	0.003	–0.044

[a] Family lifecycle=FLC ; Lifecycle=LC; Residence=RES; Income=Ic; Expenditure=EXP; Asset=AST; Insurance and annuity=INS; Life satisfaction=SAT. All observations are based on the survey's previous year if not specifically mentioned.
[b] '·', '*', '**', and '***' denote statistical significance at the 10%, 5%, 1%, and 0.1% levels, respectively.
[c] 's.e.': standard error.

these replacements for all the variables and repeats the process until the imputation error is reduced to a sufficient level. The resulting matrix is the final filled dataset. Research on the reliability and efficiency of this method has been conducted, for example, Shah *et al.* (2014) showed that it is unbiased and more efficient than multivariate imputation by chained equations (MICE) method.

In general, random forest is an algorithm that randomly selects part of variables from its entire set and uses them to train a model via a classification and regression tree (CART), a type of decision tree algorithms. As this method can be naturally applied to both quantitative and categorical variables, it has advantages over other methods that require further consideration, such as kNN.

We apply the one-stage random forest missing value imputation (Stekhoven and Bühlmann, 2012). We use the 'missForest' R package for the random forest imputation. Oshiro *et al.* (2012) proposed using 64-128 trees; however, in this work, the number of trees to grow in each forest is set to 50. Stekhoven and Bühlmann (2012) argued that the square root of the number of variables is best suited to the number of variables randomly sampled at each split. Thus, the numbers of variables at each split are set to 35(for the sixth survey) and 36(for the fifth and seventh surveys) in this study.

4. Empirical results

4.1. Results of ordinary logistic regression and lasso logistic regression

This section analyzes the results of the lasso-regularized regression for the surrender. While lasso regression is good at selecting variables, it has some drawbacks. If $n > p$ (not in our case) and some predictors are highly correlated, the prediction of ridge regression seems to perform better (Tibshirani, 1996). If $p > n$ (e.g. panel data), then lasso regression selects at most n variables before it reaches saturation (Zou and Hastie, 2005). Even worse, lasso regression does not permit group selection; when a group of regressors is correlated with each other, only one of them is selected in the lasso regression (Tibshirani, 1996). Some of the survey questions in our panel data are categorical and these are transformed into dummy variables, which are correlated to some degree. The second drawback is lessened for those dummy variables since we can sense if at least one of the variables is selected in the lasso regression.

Using the 10-fold cross-validation method to select the parameter λ, we find that $\lambda_{\min} = 0.00344$ and $\lambda_{1se} = 0.00795$. λ_{1se} is selected in the analysis. Table 5 presents the results of the ordinary logistic regression (OLR) and lasso logistic regression (LLR) analysis for policyholder's surrender.

The best OLR model selects 15 variables: three family lifecycle (FLC) variables, no lifecycle (LC) variables, three residence (RES) variables, one income (IC) variable, three expenditure (EXP) variables, two asset (AST) variables, two insurance and annuity (INS) variables, and one life satisfaction (SAT) variable. The best LLR model selects 17 variables: four FLC variables, one LC variable, four RES variables, one IC variable, two EXP variables, two AST variables, two INS variables, one SAT variable. The comparison of the two models shows that 12 of the 15 OLR variables overlap with the LLR variables and that the OLR model is more parsimonious than the LLR model. A policy duration is well-recognized as a kind of driver of surrender in insurance. Unfortunately, however, there is no survey question in regard to policy duration. Thus, we were not able to include it as a control variable in our analysis.

Now, we analyze the effect of each explanatory variable on the surrender, first, in the order commonly found in the two models and, second, in the order found in the LLR model. A policyholder determines whether to purchase, maintain, and terminate an insurance contract by comparing its cost and utility. Therefore, we interpret the trends of cost and utility. The first category is the FLC variables. This category refers to the physical, psychological, relational, and intellectual stages experienced by family members from childhood to old age. The surrender rate increases if there are children or grandchildren in a family, whereas it decreases if there is a person aged 65 or older in the family. The relationship between a policyholder and an elderly household member is unknown; however, the household member is thought to be the parent or spouse of the policyholder. Younger (older) dependents have a higher sensitivity to cost (utility). Boj del Val *et al.* (2020) also showed that the purchase of long-term care insurance reduces financial sustainability for the elderly based on Spanish survey data. If there are unemployed family members in the household, the surrender rate is lower, possibly to reduce future financial risks that could grow because of the current unemployment status. Second, the policyholder's divorce status among the LC variables raises the surrender rate.

The third type is the RES variables. The higher the average value of assets of households in the residential area, the lower the surrender rate. Seoul, a metropolitan city, is an economically higher residential environment than other regions and has a lower surrender rate. The surrender rate for residents of apartments in Korea is higher. Table 5 also shows a high surrender rate for monthly renters (low assets) and a low surrender rate for home-owners (high assets). As of 2015, 56.8% of citizens owned homes in Korea (Statistics-Korea, 2015), and Green and Hendershott (2001) showed

that home-ownership suppresses the mobility of workers, which leads to higher unemployment. This argument also suggests that for home-owners with insurance, the likelihood of unemployment among household members increases and that the surrender rate is lower because of the inverse correlation between the unemployed household member variable and surrender rate, as shown in the above analysis.

Fourth, high private transfer income among the IC variables lowers the surrender rate. Private transfer income usually means family assistance received by individuals, especially family members, relatives, and friends because of non-economic activities. Brown *et al.* (2012) argued that consumers purchase more insurance if they can be helped or cared for free from their family. Similar to unemployed family members, individuals in need of financial aid are interpreted as having a lower surrender rate to mitigate future financial risks. Fifth, with regard to the EXP variables, the higher consumer spending in the last year, the lower is the surrender rate. In view of insurance benefits, if a large amount of future cash outflow is expected, the consumer is considered to be prepared for that. Altogether, the higher the probability of unemployment, higher income, or higher expenditure, the more the surrender rate falls to prevent greater financial risks despite current economic difficulties. Another EXP variable is the number of children that incur childcare expenditure. If this grows, the surrender rate increases, similar to the young dependents variable (FLC).

The sixth type is the AST variable. If the policyholder is expected to receive a bequest from a spouse or parent, the surrender rate falls; by contrast, if such an inheritance is impossible because of the absence of a spouse, the surrender rate rises. This may be because the insurance cost is higher than the insurance benefit when a policyholder does not expect to receive an inheritance. The seventh category is the INS variables, which are difficult to observe in previous research. In Korea, non-holders of National Health Insurance tend to surrender more than holders. Retirement pension holders surrender more than non-holders. The law of diminishing marginal utility theory describes that the first service brings more utility than later ones. According to that, we speculate that holders of both retirement pension and health insurance feel less additional utility when choosing health insurance on top of retirement pension because of the increased likelihood that their insurance benefit will overlap. Eighth, the higher the SAT variables, the lower is the surrender rate.

Now, we finalize our regression analysis. Since the variables are observed for the policyholders mostly in their 50s or older, we need to be careful when generalizing our findings to the surrender behavior to all ages of policyholders.

4.2. Comparison with literature

Russell *et al.* (2013) empirically tested and found support for three hypotheses on lapse behavior: the interest rate hypothesis, emergency fund hypothesis, and policy replacement hypothesis. Bajaj (2017) showed that economic variables such as interest rates and financial indices affect surrender rates differently depending on the type of insurance product. Eling and Kiesenbauer (2014) focused the effect of product features and contractual properties on lapse rates.

Nevertheless, interest rates, emergency funds, and policy characteristics are not the only determinants of the surrender. Eling and Kochanski (2013) reviewed the literature on a variety of factors or variables that affect lapse behavior. Similarly, in this section, we compare our analysis results with those of prior empirical studies to assess which variables are common. Table 6 shows the significant variables commonly found in both the OLR and the LLR models and summarizes their effects on the surrender. In addition, this work investigates major research that has included the above common determinants in the analysis of the lapse. Table 6 grouped similar variables and summarizes the analysis results of those studies for comparison purpose. Only those studies relevant to this study are listed.

Table 6: Selected determinants of the lapse of health insurance: comparison with empirical and theoretical studies of lapse

Type[a]	Characteristic	Data source	Author[b]	Year	Findings[c]
FLC	Old dependents	Korea	Shim	2021	-
	Young dependents	US	Hammond	1967	+
		US	Mantis	1968	+
		Part of US	Berekson	1972	Mx
		US	Ferber	1980	-
		US	Gandolfi	1996	-
		Korea	Shim	2021	+
	Employed family	US	Showers	1994	-
		US	Ferber	1980	+
	Unemployed family	Korea	Shim	2021	-
LC	Marital status (divorce)	US	Fier	2013	+
		Korea	Shim	2021	+
	Marital status (unmarried)	US	Hammond	1967	+
		Part of US	Berekson	1972	+
		US	Ferber	1980	+
	Marital status (married without children)	US	Hammond	1967	-
		US	Ferber	1980	+
RES	Residence (urban)	US	Hammond	1967	-
		Korea	Shim	2021	-
	Residence	US	Milhaud	2018	M
	Type of housing (apartment)	US	Anderson	1975	-
		Korea	Shim	2021	+
	Home-ownership	US	Ferber	1980	-
		US	Gandolfi	1996	-
		Korea	Shim	2021	-
IC	Income	US	Hammond	1967	-
		US	Mantis	1968	-
		US	Duker	1969	-
		Part of US	Berekson	1972	-
		US	Anderson	1975	Mx
		US	Ferber	1980	Mx
		US	Burnett	1984	-
		US	Gandolfi	1996	-
		US	Russell	2013	-
		Korea	Shim	2021	-
	Transitory income	US and Canada	Outreville	1990	-
EXP	Expenditure	US	Ferber	1980	-
		Korea	Shim	2021	-
	Number of children that incur childcare expenditure	Korea	Shim	2021	+
AST	Bequest (theory)	N/A	Fitzgerald	1988	-
	Bequest (expected to bequest/gift from spouse)	Korea	Shim	2021	-
INS	Insurance (not subscribed to national health insurance)	Korea	Shim	2021	+
	Annuity (expected benefit of private retirement pension)	Korea	Shim	2021	+
SAT	Relationship with siblings	Korea	Shim	2021	-

[a] Family lifecycle=FLC; Lifecycle=LC; Residence=RES; Income=IC; Expenditure=EX; Asset=AST; Insurance and annuity=INS; Life satisfaction=SAT. All observations are based on the survey's previous year, if not specifically mentioned.
[b] The literature is listed as follows: Berekson, 1972; Gandolfi and Miners, 1996; Hammond *et al.*, 1967; Milhaud and Dutang, 2018; Outreville, 1990; Russell *et al.*, 2013; Shim (this study). Anderson and Nevin, 1975; Berekson, 1972; Burnett and Palmer, 1984; Duker, 1969; Ferber and Lee, 1980; Fitzgerald, 1989; Gandolfi and Miners, 1996; Hammond *et al.*, 1967; Mantis and Farmer, 1968; Showers and Shotick, 1994 studied the determinants of insurance demand, and we assumed that the insurance surrender behavior is opposite to insurance demand.
[c] '+'=positive effect; '-'=negative effect; 'M'=Meaningful; 'Mx'=Mixed.

This study differs from the previous research in terms of the characteristics of its variables. First, this study has individual and microeconomic variables such as income, expenditure, assets, and liabilities. On the contrary, the majority of studies in the literature have failed to address the effects of these individual variables because most are based on country-level data, state-level data, or simple survey data from specific groups. Second, the influences of unusual variables in this study are not comparable. Exceptionally comparable categories include FLC, LC, RES, and IC.

Individual and microeconomic data are limited and empirical research that finds a relationship

between surrender behavior and insurance using such individual drivers is scarce in the literature. However, studies of the relationship between demand and insurance are frequent. Accordingly, we assumed in this analysis that if insurance demand is inversely related to a lack of need and the latter drives a policyholder's surrender behavior, we can infer the policyholder's surrender behavior using the determinants of insurance demand. From here, we compare the results for each category of variables.

FLC Researchers have long studied the impact of the FLC on insurance demand. Hammond *et al.* (1967) found that household wealth, household income, the wife's educational level, and other household characteristics affect insurance purchases. In this study, we found that the more elderly household members are present in the family, the lower is the lapse rate; similarly, the more there are younger household members such as children and grandchildren, the higher is the surrender rate. Berekson (1972) revealed that the influences of the children variable are mixed. Although the comparison is not exact, Duker (1969); Ferber and Lee (1980) showed that working families spend less on the purchase of life insurance than housewife families. Showers and Shotick (1994) showed that demand for life insurance increases as the number of earners increases.

LC Fischer (1973) showed that economic lifecycle patterns related to labor income, consumption, and savings affect insurance purchases. We found that divorce enhances the surrender rate, and Fier and Liebenberg (2013) showed that unmarried singles and married couples without children demand less insurance, which is consistent with our results to some degree. Both studies show a high lapse rate for singles. On the other hand, Lee *et al.* (2010) showed that high unemployment above a certain level lowers insurance policy loan. It is speculated that the surrender will also decrease as the tendency of policy loans decreases.

RES Hammond *et al.* (1967) argued that residents of urban areas make more insurance purchases. They are expected to have a lower lapse rate, which is consistent with our findings. On the contrary, Milhaud *et al.* (2010) made no conclusion whether the insurance needs of urban residents of the West Coast of the United States are higher than those of the East Coast. The apartment type of housing was found to induce a higher surrender rate in this study, in contrast to the findings of Anderson and Nevin (1975), who found that apartment residents purchase life insurance. Ferber and Lee (1980); Gandolfi and Miners (1996) found that the home-owner tends to maintain higher levels of insurance purchases. We can infer from this that the home-owner tends to surrender insurance less often.

IC Many empirical studies show that income variables affect the purchase or withdrawal of insurance. Except for those studies making mixed findings (Anderson and Nevin, 1975; Ferber and Lee, 1980), most suggest that life insurance demand increases as individuals' incomes rise (Russell *et al.*, 2013), as seen in Table 6. In this study, we observe that the higher income, the lower is the insurance surrender rate, which is consistent with the literature.

EXP Concerning lapse, no empirical studies have analyzed the impact of consumer spending variables on the surrender. Ferber and Lee (1980) revealed that the higher the specificity used for expense among household budget, the higher is demand for insurance, which is consistent with the findings of this work. This is because our analysis shows that the higher expenditure, the lower is the surrender rate.

AST Fitzgerald (1989) showed that the bequest motive is positively correlated with insurance demand. We infer from this that the bequest expectation is negatively correlated with the insurance surrender. This is consistent with our results.

INS To the best of our knowledge, we have found no research that included insurance-related explanatory variables.

SAT Burnett and Palmer (1984) discussed life insurance ownership based on psychographic traits. According to Burnett and Palmer (1984), characteristics such as fatalism, socialization preference, community involvement, self-esteem, and being opinionated affect life insurance ownership. Other studies of lapse have observed no variables such as psychological factors due to data limitations.

5. Conclusion

Policyholders' decision to surrender insurance can be motivated by many factors including the product features, business cycles (e.g. new business), and policy characteristics (e.g. insured's age and contract age). It is also affected by macroeconomic factors such as changes in market interest rates according to the rational expectations hypothesis (interest rate hypothesis), and by personal funding needs due to unemployment (emergency fund hypothesis). The policyholder's personal, life cycle, irrational, and psychological behavior are not fully represented by these variables.

In this work, we noted such a limitation and included variables that explicitly represent household, personal, and financial characteristics including household variables such as family life cycle, household composition, and type of residence; personal and microeconomic variables such as personal life cycle, income, expenditure, assets, and liabilities; variables for risk aversion, such as holding other insurance or pension plans; and psychological factors such as physical health, mental health, and life satisfaction. As the data contain a large number of variables, regression analysis on the surrender was performed both with an ordinary logistic regression and a lasso penalized logistic regression model, the latter of which can reduce collinear variables.

Our analysis identified the following factors that drive the decision to surrender a policy contract. Factors that raise the surrender rate include the presence of young dependents, divorce status, apartment housing, non-ownership of homes, the number of children that incur child care expenditure, no expected bequest from spouses, a non-holder of national health insurance, and expected retirement pension benefits. Factors that lower the surrender rate include the existence of old dependents, unemployed family members, urban residence, home-ownership, high income, high expenditure, expected bequest from spouses, and good relationships with siblings.

The data on which our analysis is based are somewhat limited because they are observed for elderly private health insurance policyholders about to retire. The analysis would be more elaborate if we could further observe well-known determinants of surrender such as the insured's age and contract age in combination with variables used in this study. Future research could study the effect of insurance literacy on surrender, as Weedige *et al.* (2019) investigated such an effect on insurance purchase. It could also conduct long-term longitudinal surveys and combine them with time series data on market interest rates, stock market indices, and insurance business cycles to provide a more accurate understanding of how far our findings can complement the increasingly popular emergency fund hypothesis.

Appendix: Descriptive statistics

Description and descriptive statistics of explanatory variables are shown in Table A.1

References

Albizzati MO and Geman H (1994). Interest rate risk management and valuation of the surrender option in life insurance policies, *The Journal of Risk and Insurance*, **61**, 616–637.

Altman DG and Andersen PK (1989). Bootstrap investigation of the stability of a Cox regression

Table A.1: Description and descriptive statistics of explanatory variables

Variable	Description	Code	Data type (unit)	Mean (standard deviation)	Proportion of true-response
Number of household menbers aged 65+	Number of household members at advanced ages (aged 65 and older)	w00num06	Integer	0.454 (0.705)	
Household composition (married couple, siblings, and children)	Household composed of married couple, their siblings, and children	w00gtype	Dummy (true/false)		0.001
Household composition (household head and grandchildren)	Household composed of household head and his/her grandchildren	w00gtype	Dummy (true/false)		0.001
Number of unemployed family members	Number of unemployed family members in the last year	h00d039	Integer	0.136 (0.353)	
Marital status (divorce)	Marital status of respondent: divorced	w00mar	Dummy (true/false)		0.039
Residence (Seoul)	Residential area of household: Seoul	w00area	Dummy (true/false)		0.169
Residence (province)	Residential area of household: province	w00area	Dummy (true/false)		0.555
Type of housing (apartment)	Type of housing: apartment	w00htype01	Dummy (true/false)		0.500
Home-ownership (self-owned)	Self-owned house of household	h00b003	Dummy (true/false)		0.828
Home-ownership (rent with deposit)	Rented house of household with deposit	h00b003	Dummy (true/false)		0.063
Private transfer income	Amount of private transfer income of household in the last year	h00d016 h00d016	Continuous (1000 KRW)	981.827 (2658.621)	
Other expenditure	Amount of other expenses of household in the last year	h00c011	Continuous (1000 KRW)	3798.808 (4530.661)	
Non-consumption expenditure (living subsidy excluding remittance to children)	Non-consumption expenses of household in the last year: living subsidy excluding remittance to children	h00c012h	Continuous (1000 KRW)	230.755 (1922.035)	
Number of children that incur childcare expenditure	Number of children in household who are eligible for childcare expenses in the last year	h00c033	Integer	0.016 (0.142)	
Other assets	Amount of other assets of household	h00e006	Continuous (1000 KRW)	10561.870 (15956.070)	
Expected bequest/gift from spouse (yes)	Inheritance from spouse is expected in the future	p00i080	Dummy (true/false)		0.204
Expected bequest/gift from spouse (no spouse)	Inheritance from spouse is not expected in the future due to the absence of spouse	p00i080	Dummy (true/false)		0.071
Subscribed to national health insurance (no)	Respondent is not subscribed to Korea's national health insurance	p00f001	Dummy (true/false)		0.009
Expected benefit of private retirement pension (yes)	Benefit from private retirement pension is expected in the future	p00e031	Dummy (true/false)		0.016
Relationship with siblings	Life satisfaction regarding relationship with siblings	p00k091	Integer (Five-point Likert scale)	3.598 (0.737)	

Codes are variable names defined in the dataset source (National Pension Service, 2019).
'KRW': Korean Republic won
List of included items in this category can be found at (National Pension Service, 2019).

model, *Statistics in Medicine*, **8**, 771–783.

Anderson DR and Nevin JR (1975). Determinants of young marrieds' life insurance purchasing behavior: an empirical investigation, *The Journal of Risk and Insurance*, **42**, 375–387.

Bacinello AR (2003). Fair valuation of a guaranteed life insurance participating contract embedding a surrender option, *The Journal of Risk and Insurance*, **70**, 461–487.

Bacinello AR (2003). Pricing guaranteed life insurance participating policies with annual premiums and surrender option, *North American Actuarial Journal*, **7**, 1–17.

Bacinello AR (2005). Endogenous model of surrender conditions in equity-linked life insurance, *Insurance: Mathematics and Economics*, **37**, 270–296.

Bajaj MVR (2017). *On the Drivers of Lapse Rates in Life Insurance*(Master Thesis), Universitat de Barcelona.

Bauer D, Gao J, Moenig T, Ulm ER, and Zhu N (2017). Policyholder exercise behavior in life insurance: the state of affairs, *North American Actuarial Journal*, **21**, 485–501.

Berekson LL (1972). Birth order, anxiety, affiliation and the purchase of life insurance, *The Journal of Risk and Insurance*, **39**, 93–108.

Boj del Val E, Claramunt Bielsa MM, and Varea Soler X (2020). Role of private long-term care insurance in financial sustainability for an aging society, *Sustainability*, **12**, 8894.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Brown JR, Goda GS, and McGarry K (2012). Long-term care insurance demand limited by beliefs about needs, concerns about insurers, and care available from family, *Health Affairs*, **31**, 1294–1302.

Burnett JJ and Palmer BA (1984). Examining life insurance ownership through demographic and psychographic characteristics, *The Journal of Risk and Insurance*, **51**, 453–467.

Chesney M and Loubergé H (1986). Risk aversion and the composition of wealth in the demand for full insurance coverage, *Swiss Journal of Economics and Statistics*, **122**, 359–370.

Cook PJ and Graham DA (1977). The demand for insurance and protection: the case of irreplaceable

commodites, *The Quarterly Journal of Economics*, **91**, 143–156.
Derksen S and Keselman HJ (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables, *British Journal of Mathematical and Statistical Psychology*, **45**, 265–282.
Duker JM (1969). Expenditures for life insurance among working-wife families, *The Journal of Risk and Insurance*, **36**, 525–533.
Eling M and Kiesenbauer D (2014). What policy features determine life insurance lapse? an analysis of the German market, *The Journal of Risk and Insurance*, **81**, 241–269.
Eling M and Kochanski M (2013). Research on lapse in life insurance: what has been done and what needs to be done?, *The Journal of Risk Finance*, **14**, 392–413.
Ferber R and Lee LC (1980). Acquisition and accumulation of life insurance in early married life, *The Journal of Risk and Insurance*, **47**, 713–734.
Fier SG and Liebenberg AP (2013). Life insurance lapse behavior, *North American Actuarial Journal*, **17**, 153–167.
Fischer S (1973). A life cycle model of life insurance purchases, *International Economic Review*, **14**, 132–152.
Fitzgerald JM (1989). The taste for bequests and well-being of widows: a model of life insurance demand by married couples, *The Review of Economics and Statistics*, **71**, 206–214.
Gandolfi AS and Miners L (1996). Gender-based differences in life insurance ownership, *The Journal of Risk and Insurance*, **63**, 683–693.
Giovanni DD (2010). Lapse rate modeling: a rational expectation approach, *Scandinavian Actuarial Journal*, **2010**, 56–67.
Green RK and Hendershott PH (2001). Home-ownership and unemployment in the US, *Urban Studies*, **38**, 1509–1520.
Greene MR (1963). Attitudes toward risk and a theory of insurance consumption, *The Journal of Insurance*, **30**, 165–182.
Grosen A and Jørgensen PL (2000). Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options, and bonus policies, *Insurance: Mathematics and Economics*, **26**, 37–57.
Hammond JD, Houston DB, and Melander ER (1967). Determinants of household life insurance premium expenditures: an empirical investigation, *The Journal of Risk and Insurance*, **34**, 397–408.
Hanoch G and Levy H (1969). The efficiency analysis of choices involving risk, *The Review of Economic Studies*, **36**, 335–346.
Hong J (2020). The effect of life insurance settlement on insurance market and consumer welfare, *Communications for Statistical Applications and Methods*, **27**, 689–699.
Judd CM, McClelland GH, and Ryan CS (2008). *Data Analysis: A Model Comparison Approach*(2nd Ed), Routledge.
Kim C (2005). Modeling surrender and lapse rates with economic variables, *North American Actuarial Journal*, **9**, 56–70.
Krstajic D, Buturovic LJ, Leahy DE, and Thomas S (2014). Cross-validation pitfalls when selecting and assessing regression and classification models, *Journal of Cheminformatics*, **6**, 10.
Lee WJ, Park KO, and Kim HK (2010). Statistical prediction for the demand of life insurance policy loans, *Communications for Statistical Applications and Methods*, **17**, 697–712.
Linnemann P (2003). An actuarial analysis of participating life insurance, *Scandinavian Actuarial Journal*, **2003**, 153–176.

Linton MA (1932). Panics and cash values, *Transactions of the Actuarial Society of America*, **33**, 265–394.

Mantis G and Farmer RN (1968). Demand for life insurance, *The Journal of Risk and Insurance*, **35**, 247–256.

Milhaud X and Dutang C (2018). Lapse tables for lapse risk management in insurance: a competing risk approach, *European Actuarial Journal*, **8**, 97–126.

Milhaud X, Loisel S, and Maume-Deschamps V (2010). *Surrender Triggers in Life Insurance: Classification and Risk Predictions*, Laboratoire de Sciences Actuarielle et Financiere(Working Paper).

Neves C, Fernandes C, and Melo E (2014). Forecasting surrender rates using elliptical copulas and financial variables, *North American Actuarial Journal*, **18**, 343–362.

Oshiro TM, Perez PS, and Baranauskas JA (2012). How many trees in a random forest?, *Machine Learning and Data Mining in Pattern Recognition*, MLDM 2012, Lecture Notes in Computer Science, Springer, **7376**, 154–168.

Outreville JF (1990). Whole-life insurance lapse rates and the emergency fund hypothesis, *Insurance: Mathematics and Economics*, **9**, 249–255.

Russell DT, Fier SG, Carson JM, and Dumm RE (2013). An empirical analysis of life insurance policy surrender activity, *Journal of Insurance Issues*, **36**, 35–57.

Schott FH (1971). Disintermediation through policy loans at life insurance companies, *The Journal of Finance*, **26**, 719–729.

National Pension Service (2019). *Korean Retirement & Income Study UserGuide Ver. 7.1.*

Shah AD, Bartlett JW, Carpenter J, Nicholas O, and Hemingway H (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a caliber Study, *American Journal of Epidemiology*, **179**, 764–774.

Shefrin H (2002). *Beyond Greed and Fear: Understanding Behavioral Finance and the Psychology of Investing*, Oxford University Press.

Showers VE and Shotick JA (1994). The effects of household characteristics on demand for insurance: a tobit analysis, *The Journal of Risk and Insurance*, **61**, 492–502.

Statistics-Korea (2015). *Percentage of Home ownership*.

Steffensen M (2002). Intervention options in life insurance, *Insurance: Mathematics and Economics*, **31**, 71–85.

Stekhoven DJ and Bühlmann P (2012). MissForest– non-parametric missing value imputation for mixed-type data, *Bioinformatics*, **28**, 112–118.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, and Altman RB (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.

Weedige SS, Ouyang H, Gao Y, and Liu Y (2019). Decision making in personal insurance: impact of insurance literacy, *Sustainability*, **11**, 6795.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Received January 29, 2021; Revised March 22, 2021; Accepted April 21, 2021