Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Two-stage logistic regression model

Mijung Kim [*]

Institute for Mathematical Sciences, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Republic of Korea

ARTICLE INFO

Keywords:
The National pension
Survivor's pension
Logistic regression
Decision tree
Predictive model assessment
Classification rate

ABSTRACT

In this article, a logistic regression model combined with decision tree for dealing with a significant interaction effect among the explanatory variables is suggested. Decision tree is applied for investigating the interaction among explanatory variables and grouping subjects based on χ^2 value for optimal split. Each group of subjects which is named cluster is determined by optimal split for the interacting explanatory variables. The suggested model incorporates this cluster as an explanatory variable for including significant interaction in the logistic regression model. This model shows better performances in assessment of predictive model than the logistic regression model or decision tree: better ranked classes, increased correct classification rate and R^2, improved Kolmogorov–Smirnov (K–S) statistic, and a better lift. National pension data are applied to this model, and as an application of the suggested model, strategies for reducing financial risks in managing and planning for pension financing are illustrated.

1. Introduction

Efficient financial planning is essential for the management of the National pension. Currently, society tends toward a low-birth rate as well as longevity, especially in Korea. To reflect the current society's trend in the management of the National pension, the prediction and classification of potential pensioners based on their probability scores for pension payment is important. This allows the possibility to determine the characteristics of the classified members, and thus to suggest strategies for reducing financial risks in planning for paying and collecting contributions by reflecting the classified members' characteristics. Of the several types of National pension, this study pertains to the survivor's pension (SP) that occurs as the result of diverse and unexpected causes.

The data are analyzed with the goal to predict the occurrence of the SP payment and determining which factors are influential to the SP payment among five basic factors (explanatory variables). The classification of the potential pensioners and finding their characteristics for providing strategies on managing and planning the financing of pension in relation to the basic but essential factors is also a goal of this study.

The National pension data consists of three groups: disability pensioners (DP), early old-age pensioners (EOP), and potential pensioners (PP) who are insured persons or qualify as a pensioner but do not belong to the former two groups. The logistic regression analysis revealed the PP group showed the largest probability of odds for the occurrence of SP among the three groups. Thus, the present study for SP is conducted with the PP group, which is the most influential for the occurrence of SP.

SP is given to the insured person's survivor, where the insured person has participated for over 10 years in the pension, and died without receiving benefits; a lump-sum death payment would be provided when the eligibility of the SP is not met.

The statistical model is fitted to the fatalities among the PP for the purpose of examining the significant factors of the five basic explanatory variables. The constructed model predicts and classifies the pensioners according to the predicted score for the occurrence of SP.

Several studies on pension have been performed with a logistic regression model (Bergh, Baigi, Månsson, Mattsson, & Marklund, 2007; Chen, Rosenheck, Greenberg, & Seibyl, 2007) or a probit model (Huberman, Iyengar, & Jiang, 2007).

Not only logistic regression but also decision trees are used for estimating class membership of a categorical dependent variable without any assumption of the explanatory variable (Breiman, Friedman, Olshen, & Stone, 1984; Buntine, 1992; Lewis, 2004). Few works have been published on the comparison of classification techniques in different areas (Camdeviren, Yazici, Akkus, Bugdayci, & Sungur, 2007; Kurt, Ture, & Kurum, 2008).

Logistic regression has been utilized for predicting the occurrence of an interesting event or estimating the probability score for occurrence of an interesting event (Agresti, 2002; Hosmer & Lemeshow, 1989). This model provides the information on the effects of the explanatory variables regarding the dependent variable. However, when the logistic regression model includes significant interaction effects, the main effect becomes complicated to explain. In addition, if many interaction effects exist, which interactions should be included in the model are difficult

[*] Tel.: +82 2 2123 4093; fax: +82 2 363 4845.
E-mail address: mjkim@yonsei.ac.kr

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.08.063

to determine. Conversely, decision tree allows explicit examination of the interaction effect, and to determine which interaction effects are most influential and thus provide the influential interactions to be involved in the model.

A two-stage logistic regression model for handling interaction effect is suggested in this paper in order to explain both the main and the interaction effects in the logistic model; influential interactions are selected via decision tree analysis, and a cluster variable of representing optimal trees as categories is involved in the logistic regression model as an explanatory variable. This two-stage logistic regression model incorporates interactions of explanatory variables and explains the main effect when interactions in the logistic regression model are present. This suggested model improves correct classification rate (CR) as well as the Kolmogorov–Smirnov (K–S) statistic which measures how well the classified classes are ordered, and improves Max rescaled R^2 which measures the correlation between the observed and the predicted value.

Comparison of the suggested model with traditional logistic regression models is shown in the following order.

Section 2.1 describes the data and logistic regression model for the occurrence of the SP payment. Section 2.2 discusses the benefits of sampled data with equalized frequency of the binary responses in applying the logistic regression model. Three logistic regression models are introduced.

Section 2.3 introduces the suggested model, motivation, and the how to incorporate the interaction effect with application of the decision tree to the logistic regression model. Comparisons of logistic regression models with the suggested model are also presented. Four possible logistic regressions are compared regarding performances.

2. Material and methods

2.1. Data and predictive model for SP

2.1.1. Data

The data for this study was collected from January 1988 to May 2007 at the National Pension Service in Korea (NPS). NPS records from the insured persons, insured duration, sex, income level, type of coverage, and age at death. These five variables are considered for classifying potential pensioners at NPS, and thus these are involved in the study as explanatory variables. Response variable is binary, where it takes 1 for the occurrence of the SP payment, and 0 for a lump-sum death payment.

The data description of the five explanatory variables and the dependent variable is provided in Table 1. Data analysis is performed with SAS (2000) V. 9.1, SAS Institute Inc. (1998, 2000, 2002).

The explanatory variables are sex (SEX), insured duration (DURA), level of income (WAGE), age at death (AGE), and type of coverage (CLASS).

Among the study subjects 88.74% were male and 11.23% were female. CLASS is the type of coverage at the time immediately before pension payment begins; 36.57%, 11.22%, and 52.21% from workplace based insured, voluntarily inured, and individually insured pensions, respectively. DURA is the length of insured duration from affiliation to the time of payment, in months. WAGE is the average adjusted level of income from 1 to 45, which was from 1 to 53 until April of 1995: adjustment is made with linear interpolation. AGE is the age of terminated pension due to death.

Table 1
Description on variables

Variable	Definition	Characteristic
Y	Dependent variable for occurrence of SP	1 = Survivor's pension payment, 0 = lump-sum death payment
SEX	Dummy variable	1 = Male, 0 = female
DURA	Insured duration	Number of months
WAGE	Level of income	Continuous value from 1 to 45
CLASS	Type of coverage	0 = Workplace based insured person, 1 = voluntarily insured person, 2 = individually insured person
AGE	Age	Age at death

2.1.2. Predictive model

The statistical model was fitted to the data from 343,528 fatalities among the potential pensioners for predicting SP payment and examining significant factors of the five explanatory variables, where nine subjects were eliminated from the 343,537 due to a nonspecific code for gender status.

The 343,528 fatalities are either survivor pensioners (296,375 members) or those who received a lump-sum death payment (47,162 members), and this is expressed with a binary dependent variable, which takes 1 for the SP payment and 0 for the lump-sum death payment. The data show a highly unbalanced frequency of the 2 categories; 86.27% consumes value 1, and 13.73% consumes value 0.

As a predictive model, the logistic regression model is fitted to the data, where for p explanatory variables it can be written as

$$\log\left(\frac{\Pr(Y=1|X_1,\ldots,X_p)}{\Pr(Y=0|X_1,\ldots,X_p)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p,$$

for the probability of the event to occur,

$$\Pr(Y=1|X_1,\ldots,X_p)=\frac{\exp(\beta_0+\sum_{i=1}^{p}\beta_iX_i)}{1+\exp(\beta_0+\sum_{i=1}^{p}\beta_iX_i)}\text{ and coefficient }\beta_i.$$

The logistic regression model provides each pensioner's predicted score for the occurrence of the SP payment. The subjects may be classified based on this score and thus the classes are ordered with the probability for occurrence of SP where the high ranked class pertains to the high score for SP payment, and vice versa. This information can be utilized for strategy and plan of pension management by investigating characteristics of class members based on their explanatory variables. However, highly unbalanced frequency in one category makes it difficult to classify pensioners since the relative frequency distribution of predicted scores is extremely skewed to one side, and thus the lengths of the class intervals are highly irregular. The survivor's National pension data of Korea are extremely skewed to the left in the distribution of the relative frequency of the scores.

To solve this problem, the current paper suggests sampling data to have equal frequency of binary responses, which is discussed in Section 2.2.

2.2. Simple random sample with equalized frequency of the binary responses and three predictive models

In this section, predictive models fitted to the original data and sampled data, are compared. Three logistic regression models are named as model (1), (2), and (3).

Model (1) is constructed for situations in which the main effects of the five explanatory variables are considered in the model fitted to the original data. Model (2) is constructed for situations where main and interaction effects of the five explanatory variables are considered in the model fitted to the original data. Model (3) is constructed for situations in which main effects of the five explanatory variables are considered in the model for the sampled data with equalized frequency of the two categories. With step-wise selection, all five variables and interactions were found to be significant as shown in Table 2.