

# Classification of the Insurance sector with logistic regression

Bahadtin Ruzgar and Nursel Selver Ruzgar

**Abstract**— In statistical case studies where categorical results such as “successful-unsuccessful”, “ill-not ill” and “good-fair-bad” are obtained as a result of evaluation of data, the logistic regression is a rather suitable statistical method. In this study, the data for the years 2004, 2005 and 2006 from 53 companies that are active in the insurance sector in Turkey were evaluated by using logistic regression method. However, since the data were not sufficient for all the insurance companies, twelve insurance companies were eliminated from the evaluation. Forty-one companies used for the analysis were divided into two groups depending on their activity area. Seventeen companies were evaluated by using the data on individual accident, health and life branches; and twenty-four companies were evaluated by using data on fire, transportation, engineering, agriculture, all-risks, obligatory traffic, obligatory highway transportation, individual accident and other accident and health branches. The success ranking of companies is made as companies in the first 10 and companies between 11 and 20. Whether such classification of 41 companies collides with the classification of “successful” and “unsuccessful” companies according to geometrical mean and median was determined with a comparison. The first six-month data of 2006 year were used for control and the classification obtained from models was compared to real classification of companies.

**Keywords**—Classification, Discriminant analysis, Logistic regression.

## I. INTRODUCTION

The researchers and model designers always endeavor to convert the data they obtain from real events or experiments to functional structures by means of various models. Although to establish a mathematical model is rather difficult, doing so ensures producing very beneficial information. Classification of the data used in models constitutes the very important part of the statistical analysis and it is widely used by various science branches mainly in health. The following are examples of such logistic regression studies. The comparison of the mobile nursing system that was established between the years of 1977-1985 in America to restrict the health expenses to the former system was examined with multi logistic regression analysis [1]. The

American data obtained from extraordinary events such as wars, elections, political crisis and epidemic diseases were used to determine differences between the periods when such events occurred and other periods by means of logistic regression [2]. Binary logistic regression was used to calculate the retirement age of people depending on age, sex, economical and social statuses [3]. Between the years of 1980-1995, the data of bankrupted American companies were examined; 237 of the bankrupted companies were handled as samples for the year 1992, and the financial and non-financial values of their final bankruptcy resolutions were examined with logistic analysis and their classifications were estimated [4]. The health insurance classification of insured and uninsured low-income children in America between the years of 1995-1999 and the classification of uninsured ones according to their sex, age and economical status were made by using the logistic regression model [5]. The national health researches of the Australian households were made by using 2001 data; the rate of switching to private health insurance and the reasons for switching including economical, social and health factors were examined by means of multi logistic regression analysis [6]. In classification of car accidents in America, the logistic regression analysis was performed using the variables such as literacy rate, economical status and sex [7]. In Japan, 57 big parent companies that were very important for the Japanese economy between the years 1998-2001 were classified as “financially under stress” and “peaceful” [8]. Logistic modeling was used to determine whether a Treatment Center established for purpose of treating visually disabled or blind people to help them find new jobs was beneficial for such people [9]. Between the years 1980-2004, the disability risk and disability risk insurances were examined in America and a classification was made by using the logistic regression according to workability limits, non-workability situations and the need to get health care for people who retired due to a physical disability [10]. In another study, the logistic regression was used in determination and classification of car insurance tariffs of insurance companies [11]. When theoretical studies related to the logistic regression are examined, it is apparent that the development of the coefficient estimation methods has caused the widely used logistic regression models to be examined in a more detailed manner. In the logistic coefficient estimation procedure, the popular discriminate function approach was used [12].

$\chi^2$  Likelihood rate ( $G^2$ ), pseudo likelihood estimations, consistency benefit and hypothesis tests were examined in the logistic regression [13]. The distribution of fault terms and

Manuscript received January 31, 2007; Revised version received April 9, 2008.

B. Ruzgar is with the Actuaries Department, Banking and Insurance School, Istanbul, Marmara University, Turkey, (phone: 90-216-414 99 89; fax: 90-216-347 50 86; e-mail: bruzgar@marmara.edu.tr).

N. S. Ruzgar, is with Vocational School of Social Science, Marmara University, Istanbul, Turkey, (phone: 90-212-517 20 16; fax: 90-216-517 20 12; e-mail: nruzgar@marmara.edu.tr).