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].

χ^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).

approach of parameter values to real values were examined in the logistic regression [14]. Traditionally, a researcher or experimenter desires to find whether there is a relation between two or more variables and to express such relation with an equation [18]. For instance, an engineer may want to know the relation between the pressure and temperature, an economist between the income level and consumption expenses, an insurer between the number of policies sold and profitability, and an educator between the absent days of students and their success ranks. An equation showing the relation between two (or more) variables not only demonstrates the functional form of the relation between variables but it also estimates any variable if the value of another is known [16]. Determination the relation between two or more variables is generally necessary for two types of information. These are, firstly, the reliability of estimations on values of any variable by means of observation results on another variable and, secondly, the rate of some determinative factors related to observed differences in variable values. In other words, if two rational variables are connected to each other, the information on one of these variables can be used to estimate the values of other variable. For this reason, the functional type, direction and rank of the relation between variables must be known. When a dependent variable is a classified variable which depends on two situations while independent variables can be continuous, discrete or classified, logistic regression has a quite functional relation and it is suitable for category classification by using the structure of regression analysis.

II. METHODLOGY

The purpose of the study is to control the explained success performances of 53 insurance companies and to check whether the companies that are divided into two groups according to geometrical mean and median are in correct classification in relation to rates having deviated end values. In this study, the logistic regression analysis was performed by considering the policy numbers and total premium productions of 53 insurance companies, but only 41 companies were evaluated [17], for the years 2004 and 2005 on basis of 12-month branches. The statistics of companies have been regularly broadcasted in Internet and the companies are being classified according to the changes in their statuses when compared to their statuses of the previous year. The success ranking of companies is made for the companies in the first 10 and the companies ranking between 11 and 20. A comparison was made to determine whether such classification of 41 companies (17 companies of group I and 24 companies of group II) collides with the classification of "successful" and "unsuccessful" companies according to geometrical mean and median. The first six-month data of the year 2006 were used for control, and the classification obtained from models was compared to real classification of companies.

III. RESULTS

17 companies (Group I) out of 41 companies were examined by considering the data on individual accident (IA), health (H) and life (L) branches and 24 companies (Group II) were examined by considering the data on fire (F), transportation (T), engineering (E), agriculture (A), all-risks (AR), obligatory traffic (OT), obligatory highway transportation (OHT), individual accident (IA) and other accident (OA) and health (H) branches. A classification was made with 3 different regression models by using the data of 17 companies in Group I and 24 companies in Group II for 2004 and 2005 years. Firstly, when they were classified by considering their success percentages when compared to previous year for 2004 and 2005 years, the companies in the first 20 were classified as "successful" and others as "unsuccessful". Besides, for data of 2004 and 2005 years, separate models were found. Secondly, the policy numbers of companies for 2004 and 2005 years were considered and as they were formed from deviated data, they were classified as" successful-unsuccessful" according to geometrical mean and median. For each 2004 and 2005 years, the companies of which policy numbers were within and above the geometrical mean were classified as successful and below the geometrical means as unsuccessful. The companies within and above the median depending on their policy numbers were classified as successful and below the median as unsuccessful. In this way, the logistic regression equations were found. Thirdly, the premium fees of companies (in YTL) for 2004 and 2005 years were considered and the successful-unsuccessful classification was made again with the geometrical mean and median.

For each 2004 and 2005 years, the companies of which premium production was within and above the geometrical mean were classified as successful and below the geometrical means as unsuccessful. The companies within and above the median depending on their premium productions were classified as successful and below the median as unsuccessful. The logistic regression equations that were found in Table I, Table II and Table III, the logistic regression equations belong to 17 companies in Group I and in Table IV, Table V and Table VI, the logistic regression equations belong to 24 companies in Group II are given.

When we examine the statistics related to the logistic regression, we find that the Cox-Snell R^2 (CS-R^2) and Nagelkerke R^2 (Nag R^2) values that show the degree of relation between the dependent and independent variables in the logistic regression models are higher and that the -2LogL (-2log likelihood=-2LL) statistic is lower. When the model exactly represents the data, the likelihood is 1 and the -2LL statistics is zero. For this reason, the lower -2LL statistic always shows a better model [18, 19]. When the statistics related to testing of meaningfulness of model are examined, the Chi-square (χ^2) statistics, the -2LL statistics and the Blok Chi-square (B χ^2) statistics must be considered. The χ^2 statistics tests the logistic regression model in general. The χ^2 statistics firstly shows the fault only when there is a fixed term in the model and then it determines whether all the logistic coefficients except the fixed term are equal to zero. The χ^2 statistic conforms to χ^2 distribution with the degree of freedom that is equal to difference between the parameter number of examined model and parameters of model with

fixed term [20]. In logistic regression, the -2LL statistics shows the fault of model when an independent variable is added to model. For this reason, it is the measure of unexplained variance in a dependent variable and the non-meaningful statistic is a desirable situation. The B χ^2 statistic shows the change in the χ^2 statistics when a block variable is added to model [21]. When the relation measure in a logistic regression analysis is examined, we see that a widely accepted statistical logistic regression that is similar to R^2 statistics does not exist. In the regression analysis, R^2 shows the percentage of explained variance of dependent variable but the variance of a dependent variable in the logistic regression analysis depends on the probability distribution of that variable. For this reason, R^2 in regression analysis must not be confused with R^2 in logistic regression. The mostly used R^2 statistics for the logistic regression are CS-R^2 statistic and Nag R^2 statistic [18], [22]. CS-R^2 statistic may have a value higher than zero. Its value below 1 strengthens the interpretation of the statistic. Nag R^2 statistic was developed to ensure CS-R^2 statistic to have values between 0 and 1. The statistic closer to 1 means the relation is high.

After 17 companies in Group I are classified according to their success percentages rate when compared to previous year using the classification scheme of "successful- unsuccessful" and then checked to see if they are in the first 20 of all insurance companies, it is seen in Table I that they are in correct classification with 94.12% for the year 2004 and 82.35 % for the year 2005. In the 2004-year logistic regression model, Company 4, Anadolu Hayat ve Emeklilik, is estimated as unsuccessful, whereas, in reality, it was successful. In the 2005-year logistic regression model, Company 4, Anadolu Hayat ve Emeklilik, and Company 6, Aviva Hayat ve Emeklilik, companies are estimated as unsuccessful, whereas, in reality, they were successful. Company 11, Garanti Emeklilik, is estimated as successful, whereas, in reality, it was unsuccessful. In the 2005-year logistic regression model, Nag R^2 is 0.505 and -2LL is 11.610 and these demonstrate that the model is not in a good classification.

Table I. The logistic regression values of 17 companies in the first 20 of Group I (successful) and of other companies (unsuccessful) for 2004 and 2005 years.

$First20_{2004}$=-7.351484+4.579242E-04*IA -6.741806E-6*L					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec.Actual/Predicted
0.54	2	16.41	0,000273	94.12	4 (1/0)
Step	-2 LL	CS-R^2	Nag R^2	Overall %	The cut value
1	4.18	0.62	0.88	94.1	0.500
$First20_{2005}$=-1.276982+3.363836E-05*IA-1.554465E-03*H					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec.Actual/Predicted
0.37	2	8.16	0.0167	82.35	4(1/0), 6(1/0), 11(0/1)
Step	-2 LL	CS-R^2	Nag R^2	Overall %	The cut value
1	11.61	0.34	0,505	76.5	0.5

When 17 companies in Group I are classified as successful and unsuccessful according to geometrical mean and median in respect to their policy numbers of 2004 and 2005 years, it is seen that this classification is made in 100% correctness. Nag R^2 is 1 and -2LL is 0.000 in the logistic regression model made according to geometrical mean and median for 2004 year. However, though the classification is not wrong, in the logistic regression model of 2005 according to median, Nag R^2 is 0,185 and -2LL is 20,973. This model is not a suitable logistic regression model.

Table II. The logistic regression values of 17 companies in Group I that are classified as successful and unsuccessful according to geometrical mean and median of their policy numbers for 2004 and 2005 years.

$Policy\_num_{2004}$= -367.218+2.41603E-03*L (GM) $Policy\_num_{2004}$=-248.6836+9.521043E-04*L+8.481388E-04*IA (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class.%	Misclass. Rows Sec. Actual/Predicted
0.60 /0.63	1/2	22.1/23.5	0/0	100/100	-/-
Step	-2LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/0	0.73/0.75	1/1	100/100	0.5/0.5
$Policy\_num_{2005}$=-84.39449+1.015181E-03*L+8.428923E-04*IA(GM) $Policy\_num_{2005}$=-41.95804+3.689567E-04*L+8.586143E-04*IA (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class.%	Misclass. Rows Sec.Actual/Predicted
0.63/0.63	2/2	23.5/23.5	0/0	100/100	-/-
Step	-2LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/20.9	0.74/0.14	1/0.19	100/58.8	0.5 /0.5

Table III. The logistic regression values of 17 companies in Group I that are classified as successful and unsuccessful according to geometrical mean and median of their premium productions for 2004 and 2005 years.

Premium $prod._{2004}$= -4,66442+ 3,984531E-04*IA (GM) Premium $prod._{2004}$= -2.04204+ 1.879997E-04*IA (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class.%	Misclass. Rows Sec.
0.5/0.4	1/1	17.4/11.4	0/0	88.2/82.4	13(0/1),16(1/0),13(0/1), 14(1/0), 16(1/0)
Step	-2LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	4.8/10.2	0.7/0.5	0.9/0.7	88.9/76.5	0.5/0.5
Premium $prod_{2005}$= -2.407429+ 9.659029E-06*L+ 8.835993E-04*H+ 3.710223E-05*IA (GM) Premium $prod._{2005}$= -.8682666+ 7.672565E-06*L (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class.%	Misclass.Rows Sec.
0.4/0.28	3/1	9.05/5.8	0.03/0.02	70.6/70.6	3(1/0), 6(1/0), 8(1/0), 9(0/1), 11(0/1)/1(1/0), 6(1/0), 8(1/0), 9(0/1), 12(1/0)
Step	-2LL	CS- R^2	Nag R^2	Overall%	The cut value
1/1	18.5/18.1	0.23/0.2	0.3/0.4	70.6/64.7	0.5/0.5

When the 2004 and 2005 year premium productions of 17 companies in Group I are examined in YTL, they are classified as successful and unsuccessful according to geometrical mean and median of their premium productions and the logistic regression model is applied, it is seen that the classification can not provide a suitable separation. Though Nag R^2 values are high for 2004 year, -2LL values are high and percentages of correct estimation values are between 80-

90 %. Besides, there are wrong estimations made with the logistic regression model. For 2005 year, Nag R^2s are rather low and -2LLs are rather high. For this reason, as the wrong classifications estimated with the logistic regression models are many, the correct classifications percentages are about 70%. This shows that the logistic regression can not be used for premium productions.

Table IV. The logistic regression values of 24 companies in the first 20 of Group II (successful) and of other companies (unsuccessful) for 2004 and 2005 years.

$First20_{2004}$=-63.82845+9.545645E-04*OHT+7.222392E-04*T					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.6	2	31.76	0	100	-
Step	-2 LL	CS- R^2	Nag R^2	Overall %	The cut value
1	0	0.7	1	100	0.5
$First20_{2005}$= -109,7679+ 2,055966E-03*OHT					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.59	1	31.76	0	100	-
Step	-2 LL	CS- R^2	Nag R^2	Overall %	The cut value
1	0	0.734	1	100	0.5

Table V. The logistic regression values of 24 companies in Group II that are classified as successful and unsuccessful according to geometrical mean and median of their policy numbers for 2004 and 2005 years.

$Policy\_num_{2004}$= -644.5532+0.0141135*F+ 0.0136799*T (GM) $Policy\_num_{2004}$= -644.5532+0.0141135*F+ 0.0136799*T (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.6/0.6	2/2	32.6/32.6	0/0	100/100	-/-
Step	-2 LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/0	0.75/0.74	1/1	100/100	0.5/0.5
$Policy\_num_{2005}$= -109.7679+ 2.055966E-03*OHT (GM) $Policy\_num_{2005}$= -271.143+ 1.729338E-03*OHT+ 3.164847E-03*T-1.054694E-02*OT (Me)					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.6/0.6	1/3	31.8/33.3	0/0	100/100	-/-
Step	-2 LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/0	0.7/0.8	1/1	100/100	0.5/0.5

When 24 companies in Group II are classified in respect to their percentage success rates when compared to previous year their classification of successful-unsuccessful by considering whether they are in the first 20 or not among all other companies- it is seen that classification for 2004 and 2005 years are made correctly with 100%. Nag R^2 is 1 and -2LL is 0.000 for 2004 and 2005 years and this shows that the logistic regression models are correct. Besides, none of the companies is classified wrong in reality and estimation. When 24 companies in Group II are classified as successful and unsuccessful according to geometrical mean and median in respect to their policy numbers of 2004 and 2005 years, it is seen that this classification is made in 100% correctness. When the policy numbers in years of 2004 and 2005 are examined separately, it is determined that Nag R^2 is 1 and -2LL is 0. This shows that the logistic regression models are in correct classification.

When the 2004 and 2005 year premium productions of 24 companies in Group II are examined in YTL, the successful-unsuccessful classification of 24 companies is made by considering the geometrical mean and median of their premium production. According to the geometrical mean, the correct classification percentage in the logistic regression model of 2004 year is 91.67%. Though Nag R^2 is 1 and -2LL is 0.000, wrong classification is made for two companies. Company 5, Aviva, is estimated as successful in reality it is unsuccessful whereas Company 13 Guven, is estimated as unsuccessful in reality it is successful. According to the median, the classification according to the both logistic regression models for 2004 and 2005 year data is made in 100% correctness.

Table VI. The logistic regression values of 24 companies in Group II that are classified as successful and unsuccessful according to geometrical mean and median of their premium productions for 2004 and 2005 years.

$Premium\_prod_{2004}$= -5.58078+7.9489E-05*F+ 6.62827E-05*T (GM) $Premium\_prod_{2004}$= -63.828+ 9.5456E-04*OHT+7.222E-04*OT (M)					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.5/0.6	2/2	23.8/31.8	0/0	91.7/100	5(0/1), 13(1/0)/-
Step	-2 LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/0	0.7/0. 7	1/1	100/100	0.5/0.5
$Premium\_prod_{2005}$=-125.984+ 2.5504E-03*OHT-3.433E-03*E (GM) $Premium\_prod_{2005}$=-491.868+ 3.3655E-03*OHT+7.819E-03*T-3.860124E-03*F (M)					
R^2	D.F.	χ^2	Model Prob	Correctly Class. %	Misclass. Rows Sec. Actual/Predicted
0.6/0.6	2/3	33.1/33.27	0/0	100/100	-/-
Step	-2LL	CS- R^2	Nag R^2	Overall %	The cut value
1/1	0/0	0.7/0.8	1/1	100/100	0.5/0.5

When a classification is made according to the first six-month of 2006 by considering the data of 2004 year (companies in the first 20 are deemed as successful and others as unsuccessful), it is estimated that 16 of 17 companies (94%) are in correct classification and 1 (6%) of them in wrong classification. When a classification is made by considering the data of 2005 year, it is estimated that 14 of 17 companies (82%) are in correct classification and 3 (18%) of them are in wrong classification. When the companies above according to the first six-month of 2006 the geometrical mean are grouped as successful and ones below the geometrical mean as unsuccessful by considering the geometrical mean of policy numbers of 17 companies in Group I; it is estimated that 15 of 17 companies (83%) are in correct classification and 2 of them (12%) are in wrong classification. When a classification is made by considering the data of 2005 year, it is estimated that 16 of 17 companies (94%) are in correct classification and 1 (6%) of them is in wrong classification. When the companies above the median are classified as successful and ones below the median as unsuccessful, it is estimated that 13 of 17 companies (76.5%) are in correct classification and 4 of them (23.5%) are in wrong classification.

Table VII. The comparison of the first six-month real data of 17 companies in Group I for 2006 year to the estimated classification values found with application of the logistic regression equations estimated for 2004 and 2005 years to the values of 2006 year.

Company Name (Grup I)	First 10-20	2004	2005	Geo. Pol. Num	2004	2005	Geo Pol Prod	2004	2005	Med Pol Num	2004	2005	Med Pol Prod	2004	2005
Acıbadem Saglık ve Hayat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ak Emeklilik	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
American Life	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anadolu Hayat ve Emek.	1	0	0	1	1	1	1	0	1	1	1	1	1	0	1
Ankara Emeklilik	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aviva Hayat Ve Emek.	1	1	0	1	0	1	1	1	0	1	0	1	1	1	0
Axa Oyak Hayat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Basak Emeklilik	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Birlik Hayat	0	0	0	1	1	1	0	0	0	1	0	1	0	0	1
Demir Hayat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Garanti Emeklilik	0	0	1	1	1	1	1	0	0	1	0	1	1	0	1
Genel Yasam	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Guven Hayat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isvicre Hayat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Koc Allianz Hayat-Emek.	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Vakif Emeklilik	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
Yapı Kredi Emeklilik	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Wrong Cell Number		1	3		2	1		3	3		4	1		4	4

When a classification is made by considering the data of 2005 year, it is estimated that 16 of 17 companies (94%) are in correct classification and 1 (6%) of them is in wrong classification. When the companies above according to the first six-month of 2006 the geometrical mean are grouped as successful and ones below the geometrical mean as unsuccessful (by considering geometrical mean of premium productions of 17 companies in Group I); it is estimated that 14 of 17 companies (82%) are in correct classification and 3 of them (18%) are in wrong classification (in classification made by considering the data of 2004 year). When a classification is made by considering the data of 2005 year, it is estimated that 14 of 17 companies (82%) are in correct classification and 3 (18%) of them is in wrong classification. When the companies above the median are classified as successful and ones below the median as unsuccessful, it is estimated that 13 of 17 companies (76.5%) are in correct classification and 4 of them (23.5%) are in wrong classification (in classification made by considering the data of 2004 year). When a classification is made by considering the data of 2005 year, it is estimated that 13 of 17 companies (76.5%) are in correct classification and 4 (23.5%) of them is in wrong classification. 10 of 17 companies in Group I (59%) are classified correct with all methods.

Depending on the first 6-month data of 2006 year, in situation where the first 20 companies are deemed as successful and others as unsuccessful, when a classification is made from the logistic regression equation found by considering the 2004 year data, it is estimated that 18 (75%) of 24 companies are in correct classification and 6 (25%) of them are in wrong classification. When a classification is made from the logistic regression equation by considering the 2005 year data, it is estimated that 20 (83%) of 24 companies are in correct classification and 4 (17%) of them are in wrong classification.

Table VIII. The comparison of the first six-month real data of 24 companies in Group II for 2006 year years to the estimated classification values found with application of the logistic regression equations estimated for 2004 and 2005 years to the values of 2006 year.

Company Name (Grup II)	First10-20	2004	2005	Geo Pol Num	2004	2005	Geo Pol Prod	2004	2005	Med Pol Num	2004	2005	Med Pol Prod	2004	2005
Aig	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ak Sigorta	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Anadolu	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ankara	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Aviva	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0
Axa Oyak	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Basak	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0
Birlik	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Finans	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Garanti	1	0	0	1	1	1	1	1	0	1	1	0	1	0	0
Generali	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gunes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Guven	1	0	1	1	0	1	1	0	1	1	0	0	1	0	0
Hur	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ihlas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isık	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isvicre	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Koc Allianz	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ray	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Seker	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T.Genel	1	0	0	1	1	0	1	0	0	1	1	0	1	0	0
Teb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Toprak	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Yapı Kredi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Wrong CellN		6	4		1	2		2	2		2	4		3	4

When the companies above according to the first six-month of 2006 the geometrical mean are grouped as successful and ones below the geometrical mean as unsuccessful by considering the geometrical mean of policy numbers of 24 companies in Group II; it is estimated that 23 of 24 companies (96%) are in correct classification and 1 of them (4%) are in wrong classification (in classification made by considering the data of 2004 year). By considering the geometrical mean of the first 6-month policy numbers of 24 companies in Group II in 2006, the companies above the geometrical mean are grouped as successful and ones below the geometrical mean as unsuccessful. When a classification is made from the logistic regression equation by considering the 2004 year data, it is estimated that 23 (96%) of 24 companies are in correct classification and 1 (4%) of them are in wrong classification. When a classification is made from the logistic regression equation by considering the 2005 year data, it is estimated that 22 (92%) of 24 companies are in correct classification and 2 (8%) of them are in wrong classification. Similarly, when the companies above the median are grouped as successful and

ones below the median as unsuccessful and a classification is made from the logistic regression equation by considering the 2004 year data, it is estimated that 22 (92%) of 24 companies are in correct classification and 2 (8%) of them are in wrong classification. When a classification is made from the logistic regression equation by considering the 2005 year data, it is estimated that 20 (83%) of 24 companies are in correct classification and 4 (17%) of them are in wrong classification.

When the companies above according to the first six-month of 2006 the geometrical mean are grouped as successful and ones below the geometrical mean as unsuccessful (by considering geometrical mean of premium productions of 24 companies in Group II); it is estimated that 22 of 24 companies (92%) are in correct classification and 2 of them (8%) are in wrong classification (in classification made by considering the data of 2004 year).

When a classification is made by considering the data of 2005 year, it is estimated that 22 of 24 companies (92%) are in correct classification and 2 (8%) of them is in correct classification. Similarly, when the companies above the median are classified as successful and ones below the median as unsuccessful, it is estimated that 22 of 24 companies (92%) are in correct classification and 2 of them (8%) are in wrong classification (in classification made by considering the data of 2004 year). When a classification is made by considering the data of 2005 year, it is estimated that 20 of 24 companies (83%) are in correct classification and 4 (17%) of them is in wrong classification. 17 of 24 companies (71%) in Group II are classified correct with all five methods.

IV. CONCLUSION

In this study, the applicability of the logistic regression in classification of general sizes of insurance companies in Turkey, establishment of future oriented strategies and correct recognition of strategies of companies is tried to be demonstrated. 41 companies that have sufficient data from 53 insurance companies in Turkey were examined by means of the logistic regression to categorize whether they are being "successful and unsuccessful".

In examination, firstly, the companies were listed in order according to their success percentages when compared to their previous year successes (for the years 2004 and 2005), the first 20 companies were categorized as successful and others as unsuccessful and the logistic regression was applied. The logistic regression equations proved sufficient in classification of Group I companies and highly sufficient in classification of Group II companies. Moreover, from the logistic regression equations estimated for the years 2004 and 2005, it was examined whether a correct classification was made for the first 6-month of real data from the year 2006, and it was determined that the models are valid to a great extent. On the other hand, it was determined that the policy numbers of companies could be used as a dependent variable. For policy numbers, the geometric mean and median was used in categorical separation and the logistic regression equations were found for the years 2004 and 2005. It was seen that the logistic regression equations formed a suitable separation for Group I and Group II companies. Furthermore, for the first 6-month real data of the year 2006, estimations were made from the logistic regression models of the years 2004 and 2005 and a comparison was made to verify whether the real classifications and estimated classifications of the first 6-month real data of the year 2006 were the same.

As a result of that comparison, it was determined that the estimations confirmed the same separation in high probability. In conclusion, the logistic regression was deemed to be a good method in categorizing of insurance companies as "successful and unsuccessful".

REFERENCES

[1] W. G. Weissert, J. M. Elston, G. G. Koch, *Risk of institutionalization*,, 1977–1985 U.S. Department of Health and Human Services, 1990.

[2] G. King, L. Zeng, "Logistic regression in rare events data (Periodical style)," *Political Analysis,* vol. 9, no. 2, pp. 137–310

[3] J. B. Williamson, T. K. McNamara, "Why some workers remain in the labor force beyond the typical retirement age?" *Center for Retirement Research Working Papers*, no. 47, Boston College CRR WP 2001-9 Chestnut Hill, MA., 2001.

[4] R. A. Bran v, A. R. Leach, "Predicting bankruptcy resolution". *J. Bus Fin and Account*, vol. 29, no. 3, 2002, pp. 497–507.

[5] R. Fisher, J. Campbell, "Health insurance estimates for states (Published Conference Proceedings style)" U.S. Census Bureau Government Statistics Proceedings of the American Statistical Association Annual Meeting, New York, 2002.

[6] J. Temple, "Explaining the private health insurance coverage for older Australians (Periodical style)," *People and Place*, vol. 12, no.2, 2004, pp. 13–24.

[7] T. Lee, Y, Yeh, R, Liu, *Can corporate governance variables enhance the prediction power of accounting-based financial distress prediction models?*, Financial Distress and Bankruptcy Prediction An introduction Magnus School of Business Publisher , 2006.

[8] K. Sullivan, "*Transportation work: Exploring car usage and employment outcome in the LSAL data*," NCSALL Occasional Papers Edu Res Dev Cen Prog, Cambridge MA, 2003.

[9] M. E. Capella, McDonnall, "Predictors of competitive employment for blind and visually impaired consumers of vocational rehabilitation services (Periodical style)" *J. Visual Impairment and Blindness*, vol. 99, no. 5, 2005, pp. 303–315.

[10] A. Chandra, A. A. Samwick, *Disability risk and the value of disability insurance*, In: D. Culter, D. Wise (eds) "*Health in older ages: The causes and consequences of declining disability among the elderly*," University of Chicago Pres Chicago, 2005.

[11] M. M. Galiano, A. Christmann, "Insurance: An R-program to model insurance data (Report style)" *University of Dortmund, SFB-475, Technical Report*, 2004, pp. 5–8.

[12] J. Cornfield, *Joint dependence of the risk of coronary heart disease on serum cholesterol and sistolic blood pressure: A diskriminant function analysis*, Federation Proc., no. 21, 1962, pp. 58–6.

[13] G. Robert, N. K. Rao, S. Kumar, "Logistic regression analysis of sample data (Periodical style)," *Biometrika*, vol. 79, no. 35, 1987, pp. 58.

[14] D. E. Duffy, "On continuity-corrected residuals in logistic regression (Periodical style)," *Biometrica,* vol. 77, 1990, pp. 287–293.

[15] H. Tatl dil, *Uygulamalı cok degiskenli istatistiksel analiz,* Engin Ankara, 1992, pp. 225–232.

[16] S. Menard, *Applied logistic regression analysis,* Sage Pub. USA, 1995, pp. 37–42.

[17] Sigorta Verileri , Avaliable: http://www.tsrsb.org.tr/tsrsb/Istatistikler /Genel+Sektör+verileri/Türk+sigorta+sektörü+veriler

[18] A. Agresti, *Categorical Data Analysis*, Wiley Pub., Florida, 2002, pp.. 165–257.

[19] K. Ozdamar, *Paket programlar ile istatistiksel veri analizi,* Kaan Kitabevi, Eskisehir, 2004, pp. 601–608.

[20] S. Weisberg, *Applied linear regression*, Wiley Pub, Canada, 2005, pp. 255–265.

[21] A. S. Albayrak, *Uygulamalı cok degiskenli istatistik teknikleri*, Asil Yayin Ankara, 2006, pp. 439–462.

[22] S. F. Hair, R. E. Anderson, R. L. Tahtam, W. C. Black, *Mutivariate data analysis*, Prentice-Hall, NJ , 1998, pp. 276–281, pp. 314–321.

B. Ruzgar was born in Musulca, Edirne, Turkey in 1962. He received the MS degree in Mathematics from Marmara University, Istanbul, Turkey in 1986 and Ph. D. degree in Quantitative Methods from Mamara University, Istanbul, Turkey, in 1992.

He worked as a Mathematician, Research Assistant at Department of Management, Marmara University, Istanbul Turkey. Currently, he works as an Assistant Professor at Baking and Insurance School, Marmara University, Istanbul, Turkey. He has five books, an author of more than 15 papers in refereed journals and more than 50 papers in conference proceedings. His research interests are fuzzy logic, applied statistics, quantitative methods.

Assis. Prof. Dr. Ruzgar is a Member of Mathematics Association of Turkey, Member of Operation Research Society of Turkey, Member of Informatics Association of Turkey and Member of Association of Econometrics.

N. S. Ruzgar was born in Pinarhisar, Kirklareli, Turkey in 1962. She received the MS degree in Mathematics from Istanbul Technical University, İstanbul, Turkey in 1989 and Ph. D. degree in Quantitative Methods from Istanbul University, Istanbul, Turkey, in 1998.

She worked as a Mathematician, Lecturer, and Assistant Professor at Computer and Electronic Education Department of Technical Education Faculty, Marmara University, Istanbul Turkey. Currently, she works as an Associte Professor at Vocational School of Social Sciences, Marmara University, Istanbul, Turkey. She has four books, an author of more than 15 papers in refereed journals and more than 60 papers in conference proceedings. Her research interests are system simulation, applied statistics, quantitative methods, and distance education.

Assoc. Prof. Dr. Ruzgar is a Member of Mathematics Association of Turkey, Member of Operation Research Society of Turkey, Member of Informatics Association of Turkey and Member of Association of Econometrics.