Application of the Binary Logistic Regression Model in a Social Security Scheme in Albania

Erjola Cenaj[*], Raimonda Dervishi [2]

[1]*Department of Mathematical Engineering. Faculty of Mathematical and Physical Engineering, Polytechnic University of Tirana, Albania.*

[2]*Department of Mathematical Engineering. Faculty of Mathematical and Physical Engineering, Polytechnic University of Tirana, Albania.*

erjola_cenaj@yahoo.com

Abstract – Social security system continues to become more complex and difficult to understand. Countries vary considerably in the extent to which their social security apparatus is centralized and unified. The social security scheme in Albania is based on the system "pay as you go" according to which current payments to pensioners are funded by current contributions within the system. This creates a direct link between the number of persons contributing to the scheme, the amount of money they pay, and the benefits they receive. As a statistical modeling technique that can be applied to evaluate the simultaneous effect of a prediction group on the result of a specific variable that one or one of the two possible values, we use a binary logistic regression model in a social security scheme. We apply this method to predict the probability of non-inclusion in the social security scheme of individuals employed in Albania. From the database, we use a dependent variable and five independent variables. The dependent variable is categorical which takes two values and indicates whether an individual is part of a social security scheme or not, independent variables are: years of school, type of employment, age, sex, and zone. As a result of the application, we conclude that adding independent variables to the model increases the total percentage of the correct classification.

Keywords – binary logistic regression, contingency table, social security scheme

I. INTRODUCTION

Social security is the single largest element and foundation for the social welfare system of most countries. It grew out of the voluntary insurance agreements of medieval European craft guilds, was institutionalized by European governments in the late 19th and early 20th centuries, and soon spread to America [1]. It varies from one country to another. Since social security is the foundation of many social welfare systems, it is necessary to study them further. Approximately 140 countries have some type of social security scheme. Almost all of these countries have schemes covering work-related injuries and old-age and survivors'

pensions. More than half have provisions for sickness, and half have provisions for family allowances. The fewest schemes offered are for unemployment, although at least 40 countries have them [2].

In Albania, social security benefits are provided by the Institute of Social Security, an independent public institution whose organization and operation are regulated by the Administrative Council and the General Director. All social security contributions and benefits are tax-exempt [3]. Furthermore, the social security scheme in Albania is based on the system "pay as you go" according to which current contributions within the system fund current payments to pensioners. This creates a direct link between the number of persons contributing to the scheme, the amount of money they pay, and the benefits they receive. The higher this ratio is, the better it is for the social security system [4].

The way of managing the social security scheme over the years and its characteristics have been different before and after the 1990s. This study describes the service to predict the possibility of not being included in the social security scheme of employed individuals in Albania.
As a statistical modeling technique that can be applied to evaluate the simultaneous effect of a prediction group (age, sex, type of profession, area, years of education) on the result of a specific (categorical) variable that one or one of the two possible values (yes/no), we use binary logistic regression model in a social security scheme. Using this method, we conclusion that with adding independent variables to the model, the total percentage of the correct classification is increasing.

II. MATERIALS AND METHOD

The data that will be analyzed here is obtained from The World Bank's database [5]. These data are used to predict the possibility of not including an employed individual in the Social Security Scheme, in a group of n = 3310 individuals who are selected in two rounds. In the first round, 834 Primary Selection Units (PSUs) were randomly selected to represent the entire territory of the country. Then, 8 families for each PSU were selected to be interviewed in the second round through a procedure.

To address non-response cases or any other contact, 4 families for each PSU were selected as surrogates that provide the objective of 3310 completed questionnaires to families. The geographic scope of the analysis includes 12 prefectures of Albania, urban and rural areas as well as both genders [6]. From the survey data we make a classification in terms of inclusion or not in the scheme of individuals and table 1 is obtained.

Table 1. Distribution of individuals by participation in the social security scheme.

		Frequency	**% for frequency**
Participation in Social Security	Yes	2331	70.4
	No	979	29.6
	Total	3310	100.0

A dependent variable and five independent variables are used from the database. A dependent variable is a variable that takes two values and indicates whether an individual is part of a social security scheme or not [7].

First, we calculate the contingency table for the Independent Gender variable and the dependent variable in the Social Security Scheme to show if there is a difference between women and men and inclusion in the Social Insurance Scheme (Table 2).

From the contingency table, we notice that 35.38% are female respondents and 64.62% are males and from these 84.5% of women are included in the scheme and 62.7% of men are included in the scheme. So, a woman is expected to have more chances than a male to be included in the scheme.

Table 2. Contingency table of variables

			Gender		Total
			Male	**Female**	
Inclusion in the scheme	Yes	Frequency	1341	990	2331
		Gender (%)	62.7%	84.5%	70.4%
	No	Frequency	798	181	979
		Gender (%)	37.3%	15.5%	29.6%
Total		Frequency	2139	1171	3310
		Gender (%)	100%	100%	100%

The statistical model that is generally used for the analysis when the dependent variable is binary, meaning it has two categories, is the binary logistic regression model. It is commonly used when the outcome is coded as "1" or "0" and is not adequate for regular linear regression models.

Let denote, X the vector of predictors $\{X_1, X_2, \ldots, X_k\}$. To model this situation, we can use the binary logistic regression model [8]:

$$E(Y/X) = X\beta \qquad (1)$$

Since the expectation of a binary variable Y is P (Y= 1), the model (1) cannot fit the data over the whole range of the predictors.

Since a linear model

$$E(Y/X) = E(Y = 1/X) = X\beta \qquad (2)$$

can allow P (Y= 1) to exceed 1 or fall below 0.

In terms of the probability of Y = 1 given X, the values of predictors:

$$P(Y = 1/x) = \frac{e^{X\beta}}{1+e^{X\beta}} \qquad (3)$$

called the logistic function has an unlimited range for x while is restricted to the range from 0 to 1. Solving the equation (3) we obtain:

$$x = \log (\pi/(1 - \pi) = logit (Y = 1) \qquad (4)$$

III. RESULTS

As a result of the binary logistic regression model, referring to the data selected, we found out that Table 3 presents a summary of the binary logistic regression coefficients of processed outcomes, Wald statistics, and odds ratios [(Exp (B)] along with a 95% CI.

Table 3. Results of estimated parameters of the binary logistic regression model.

Variables of the Equation						
	B	S. E	Wald	df	Sig.	Exp(B)
Employment	0.722	0.037	378.357	1	0.000	2.058
Zone	0.147	0.047	9.706	1	0.002	1.159
Age	-0.033	0.004	57.530	1	0.000	0.967
Sex	-0.931	0.119	61.463	1	0.000	0.394
Years of School	-0.268	0.016	278.664	1	0.000	0.765
Constant	2.952	0.323	83.622	1	0.000	19.149

Furthermore, we obtain the equation:

Logit {Social Security Scheme =1| Employment, Age, Sex, Years of School, Zone} = 2,952 + 0,722 employment – 0,033 Age – 0,931 Sex – 0,268 Years of School + 0,147 Zone.   (5)

IV. DISCUSSION

This study described the service to predict the possibility of not being included in the social security scheme of employed individuals in Albania, in a group of n = 3310 individuals who are selected in two rounds. In the first round, 834 Primary Selection Units (PSUs) were randomly selected to represent the entire territory of the country. Then, 8 families for each PSU were selected to be interviewed in the second round through a procedure. The data are considered as categorical variables.

From the results of the method, it is evidenced that the model that considered all the five independent variables together was statistically significant, Chi-square = 1342.845, df = 5, p < .001. Moreover, Wald's statistics show that all variables significantly predict non-inclusion in the social security scheme.

V. CONCLUSION

Social security system continues to become more complex and difficult to understand. Countries vary considerably in the extent to which their social security apparatus is centralized and unified. Many statistical methods are used to analyze and predict this sector**.** In this paper, we apply a binary logistic regression model, to investigate the relationship between a prediction group (age, sex, type of profession, area, years of education) on the result of a specific (categorical) variable that one or one of the two possible values (yes/no). This study described the service to predict the possibility of not being included in the social security scheme of employed individuals in Albania.

From the results of the method, it is evidenced that full model which considered all the five independent variables together was statistically significant. The developed model correctly classified approximately 82% of the cases and all variables significantly predict non-inclusion in the social security scheme.

REFERENCES

[1] Lawrence H. Thompson, The Advantages and Disadvantages of Different Social Welfare Strategies, Social Security Bulletin, Vol. 57, No. 3, 1994

[2] https://www.britannica.com/money/social-security-government-program/Social-insurance

[3] https://www.euraxess.al/albania/information-assistance/social-security

[4] The World Bank, Albania - Living Standards Measurement Survey 2012

[5] Living Standard Measurement Survey for Albania 2017

[6] Cenaj E, Dervishi R, Muka Zh, "The use of logistic regression in a human health problem" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 12, December 2016

[7] Agresti A, Categorical Data Analysis, 2nd ed., University of Florida, 2003, pp.36-78

[8] Hosmer. D.V, Lemeshow.S, (2000). Applied Logistic Regression. Second Edition, John Wiley & Sons. pp 1-45