Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Alternative Assessments of the Probability of Death with a Case Study for Persons with Celiac Disease in Selected East European Countries

Simpach Ondrej

Department of Demography, Faculty of Informatics and Statistics, University of Economics in Prague, Winston Churchill sq. 4, 130 67, Prague, Czech Republic (E-mail: ondrej.simpach@vse.cz)

Abstract. The probability of death depended in the past to a considerable extent on the level of advancement of the health service, the medical findings acquired and knowledge of the appropriate treatment processes. In the case of persons with Celiac Disease, which is a disease involving gluten intolerance, the hope of survival in the majority of countries was slim until the eighties of last century. These people died at a very young age thanks to ignorance of the diagnosis of their disease. However, as soon as it was possible to determine the diagnosis of Celiac Disease correctly there was a considerable breakthrough and progress rapidly changed the hope of survival for these people. This breakthrough occurred earlier in some countries and later in others. In this way treatment procedures were found for hitherto unknown diseases, or at least there was information on reducing the consequences of these diseases. The submitted study will provide a look at the alternative assessment of the probability of death of persons with Celiac Disease and the probability of death in general. The modelling of the probability of death is possible with the use of the LOGIT model. On the basis of supplementary information about the population it is then possible to construct various probability scenarios with the utilisation of alternative variables.
Keywords: Probability of Death, Celiac Disease, LOGIT, Alternative Assessment.

1 Introduction

In spite of the fact that medicine is constantly bringing people new information and the diagnosis of new diseases, there were and still are diseases for which the existing diagnosis is only partial and thus insufficient for the complete cure of the patients (Logan et al. [4]). In the second half of last century the diagnosis began to appear in some countries of a disease involving gluten intolerance, later described as Celiac Disease. This diagnosis, however, only spread to certain countries. There were countries not only in Europe, but also throughout the world, which did not have all the necessary information from science and research published abroad (Rubio-Tapia et al. [5]). This caused various information delays and the consequences of insufficient information about the diagnoses of certain diseases had an impact on the life expectancy of these people. The probability of the death of persons with specific diseases was thus raised in comparison with the probability of death of persons in the general population not burdened by any of the diseases with a still insufficient diagnosis.

2 Methodology and Model

The dependence on age of the probability of death of a person x years old can be explained with the use of further variables, both discrete and also categorical (Freese and Long [1]). The LOGIT models are capable of estimating, with the use of the distribution function of logistic distribution, the value of the probability of death of a person x years old, where further supplementary information may create various forms of the probability function. In the presented model the probability of death of a person x years old will be estimated for the course of the next k years after the medical examination (where k is any whole number) in the case where the person has some diagnosed disease, or in the case where the person is completely healthy. In the illustration case study the analysis will be used of the probability of death of persons with Celiac Disease in comparison with the probability of death for the population as a whole. The explained variables of Y will be alternative. When the value of variable Y equals 1, then the person will die within k years, and on the contrary when the value of variable Y equals 0, the person will survive k years. So that it would be possible also to determine the values of the probability of the occurrence of this phenomenon between the two extremes, the LOGIT model of discrete selection will be applied, when the explained variables acquire values from the interval <0 ; 1> (Hoyos et al. [2]). The following variables may be used for the model:
- AGE is the precise current age of the person invited for a health check,
- CIRD is the Constant of Increased Risk of Death, which acquires values from the interval $< l ; h >$, where l and h are whole numbers. The calculation of this constant arises for the i^{th} patient from Table 1, which is created during the medical examination and where instead of the verbal replies given there were recorded $w_{i,j}$, acquiring the values 0 and 1, where 0 = patient's reply does not coincide with the word given in the appropriate square and 1 = patient's reply coincides with the word given in the appropriate square.

	$\mathbf{V}_1$	$\mathbf{V}_2$	$\mathbf{V}_3$	$\mathbf{V}_1$	$\mathbf{V}_2$	$\mathbf{V}_3$
Smoker	no	occasionally	regularly	$v_{i,j} = 0/1$	...	...
Black Coffee	no	occasionally	regularly	...		
Alcohol	no	occasionally	regularly	...		
Sleep	regular	irregular	poor	...		
Nutrition	regular	irregular	poor	...		

Table 1. Replies of patients to doctor's questions during general examination (left) and 0/1 matrix replies (right)

From Table 1, in which the replies are recorded in the 0/1 format, emerges the CIRD for the i^{th} patient from the formula (1),

$$CIRD = (w_1 \times \sum_{i=1}^{5} v_{i,1}) + (w_2 \times \sum_{i=1}^{5} v_{i,2}) + (w_3 \times \sum_{i=1}^{5} v_{i,3}) \qquad (1)$$

where w_1, w_2 and w_3 are the weights recommended on the basis of the doctor's opinion, who provided data matrices for analysis. (We can use $w_1 = 1$, $w_2 = 3.5$ and $w_3 = 7$). The general rule, arising from the literature, is not here. This Constant can take values from interval <5 ; 35>, where the extreme value of 5 means, that the patient does not increase the risk of death because of its poor lifestyle and extreme value of 35 means, that the patient increases the risk of death in the worst way possible.

- ILL is a binary variable, acquiring the values 0 = the person does not have a diagnosed illness, or 1 = the person has a diagnosed illness. In the model with Celiac Disease the variable CEL will be used.

- DEATH_K is a binary variable, acquiring the values 0 = the person did not die within k years after the medical examination, or 1 = the person died within k years after the medical examination.

The probability function for the LOGIT model (Christensen [3]) is

$$P_i = E(Y = 1|\mathbf{X}_i) = \frac{1}{1 + e^{-(b_0 + \mathbf{b}^{'}\mathbf{x}_i)}} \tag{2}$$

modified for this study in the form

$$P_i = E(Y = 1|\mathbf{X}_i) = \frac{1}{1 + e^{-(b_0 + b_1 AGE_i + b_2 CIRD_i + b_3 CEL_i)}}, \tag{3}$$

where i is the i^{th} patient. Let us set

$$Z_i = b_0 + \mathbf{b}^{'}\mathbf{x}_i \tag{4}$$

and let us insert it for the purposes of this study

$$Z_i = b_0 + b_1 AGE_i + b_2 CIRD_i + b_3 CEL_i. \tag{5}$$

The subsequent expression

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}} = F(Z_i) \tag{6}$$

is the distribution function of the logistic distribution. The probability that a person aged x-years will not die within k years after the moment of the medical examination is

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \tag{7}$$

and therefore

$$\frac{P_i}{1 - P_i} = e^{Z_i}. \tag{8}$$

By calculating the logarithm we obtain LOGIT

$$ln(\frac{P_i}{1 - P_i}) = Z_i = b_0 + \mathbf{b}^{'}\mathbf{x}_i, \tag{9}$$

which is transferred for the purposes of this study into the form

$$ln(\frac{P_i}{1 - P_i}) = Z_i = b_0 + b_1 AGE_i + b_2 CIRD_i + b_3 CEL_i. \tag{10}$$

From the general assumptions, the logarithm of the credibility function

$$lnL(b_0, \mathbf{b}) = \sum_{i=1}^{N} [Y_i ln(\frac{e^{Z_i}}{1+e^{Z_i}}) + (1-Y_i) ln(1 - \frac{e^{Z_i}}{1+e^{Z_i}})] \tag{11}$$

there arises after the substitution

$$lnL(b_0, \mathbf{b}) = \sum_{i=1}^{N} [Y_i ln(\frac{e^{b_0+\mathbf{b}'\mathbf{x}_i}}{1+e^{b_0+\mathbf{b}'\mathbf{x}_i}}) + (1-Y_i) ln(1 - \frac{e^{b_0+\mathbf{b}'\mathbf{x}_i}}{1+e^{b_0+\mathbf{b}'\mathbf{x}_i}})] \tag{12}$$

and for the purposes of this study is

$$lnL(b_0, b_1, b_2, b_3,) = \sum_{i=1}^{N} [Y_i ln(\frac{e^{b_0+b_1 AGE_i+b_2 CIRD_i+b_3 CEL_i}}{1+e^{b_0+b_1 AGE_i+b_2 CIRD_i+b_3 CEL_i}}) +$$

$$+(1-Y_i) ln(1 - \frac{e^{b_0+b_1 AGE_i+b_2 CIRD_i+b_3 CEL_i}}{1+e^{b_0+b_1 AGE_i+b_2 CIRD_i+b_3 CEL_i}})]. \tag{13}$$

3 Data, Material and Case Study

For the study mentioned it is possible to use data from the databases of health insurance companies and medical statistics. There are few health insurance companies which record events to the necessary extent. Practical analysis will be carried out for selected periods of the nineties in the Czech Republic, Slovakia and Poland. The analysis will be restricted to persons with Celiac Disease and persons with no health complications and the results will be published separately for the male and the female gender. For the experiment of non-linear regression (Spector and Mazzeo [6] or Yang and Raehsler [7]), applied in the first part of this study about 200 observations of variables consisting of two samples were obtained for each country - Czech Rep., Slovakia and Poland. It is important to note, that this is not a representative selection for the application of standard methods of mathematical statistics. The selection was not taken at random. This is the data matrix, obtained by tentative minor research. Selection consists all individual invited in 1990 to general medical examination and their health status was checked in the future. For consecutive experiment of non-linear regression, applied in the second part of the study, approximately other 200 observations of patients, consisting of two samples were obtained for each country. It is a selection of patients invited in 1995 to the overall medical examination and their health status was checked in the future (but obtained from other sources than the first selection). We hope that there is minimum probability that some patients from the first sample are contained in the second sample. Estimating the unknown parameters of non-linear regression models is no problem today. To estimate the parameters of LOGIT model Statgraphics Centurion XVI version 16.1.11 and Gretl 1.8.7 build 2010-01-24 were used.
Based on the methodology showed above the estimates of unknown parameters of LOGIT models for males in 1990 and 1995 as well as for females in 1990 and

1995 were calculated for Czech Rep., Slovakia and Poland. Table 2 shows the results for the Czech Republic (top), for Slovakia (middle) and Poland (bottom). The first model for each country is always for males in 1990, the second model always for females in 1990, a third model always for males in 1995 and the fourth model always for females in 1995. Of the estimated models were constructed graphs showing the development of the probability of death of x-years old person. (See Figure 1 for Czech Rep., Figure 2 for Slovakia and finally Figure 3 for Poland.

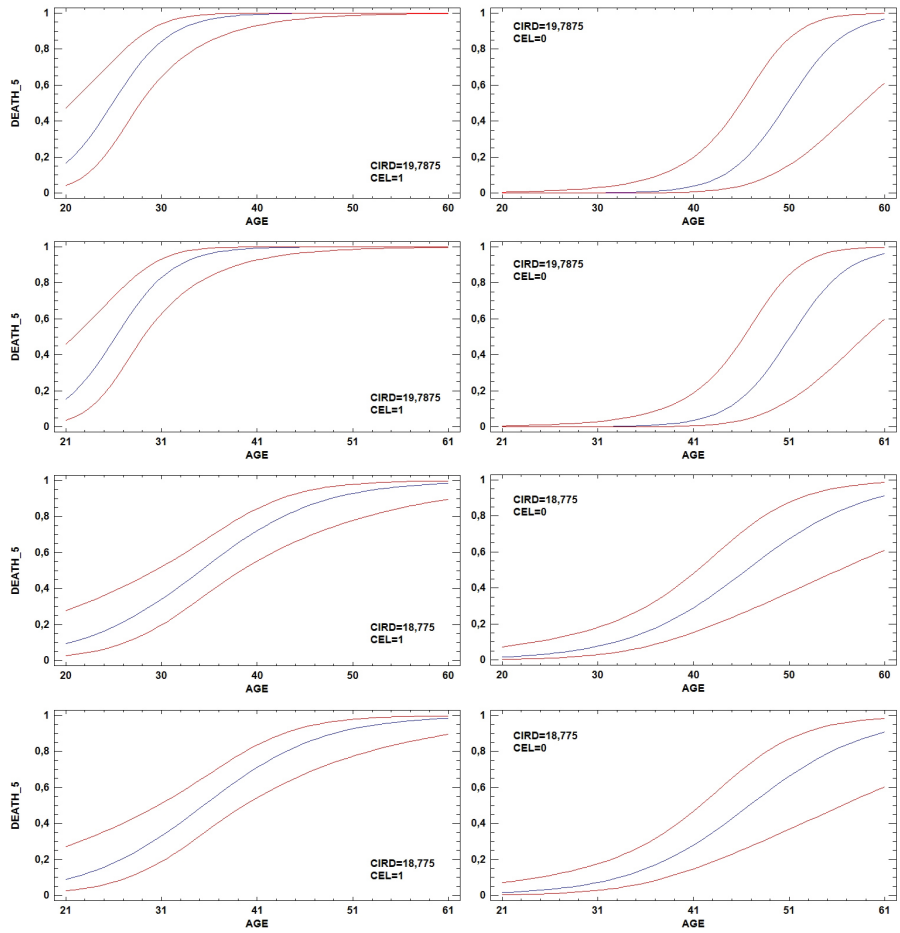

Fig. 1. Probability of death of x-years old person in the Czech Republic (males 1990 with CEL - 1^{st} row left, males 1990 without CEL - 1^{st} row right, females 1990 with CEL - 2^{nd} row left, females 1990 without CEL - 2^{nd} row right, males 1995 with CEL - 3^{rd} row left, males 1995 without CEL - 3^{rd} row right, females 1995 with CEL - 4^{th} row left and females 1995 without CEL - 4^{th} row right.)

Parameter	Estimate	St. Error	Odds Ratio	Factor	Chi-Sq.	DF	P-Value
Constant	-15,6423	3,71635					
AGE	0,333211	0,09654	1,40052	AGE	31,3889	1	0,0000
CIRD	0,370041	0,08211	1,40563	CIRD	39,0002	1	0,0000
CEL-0	-8,23231	1,90601	0,00029	CEL	78,8484	1	0,0000
Constant	-15,7816	3,71599					
AGE	0,329121	0,08534	1,38975	AGE	31,2349	1	0,0000
CIRD	0,361625	0,08578	1,43566	CIRD	38,8651	1	0,0000
CEL-0	-8,19201	1,89216	0,00027	CEL	78,8339	1	0,0000
Constant	-9,17489	1,90414					
AGE	0,161341	0,03913	1,17509	AGE	21,591	1	0,0000
CIRD	0,195672	0,04715	1,21613	CIRD	22,1363	1	0,0000
CEL-0	-1,84446	0,52223	0,15811	CEL	14,2464	1	0,0002
Constant	-9,26566	1,91888					
AGE	0,189633	0,04001	1,18655	AGE	22,0001	1	0,0000
CIRD	0,201122	0,04023	1,22366	CIRD	22,1963	1	0,0000
CEL-0	-1,83663	0,52889	0,16889	CEL	14,6398	1	0,0001
Constant	-14,771	3,61432					
AGE	0,319513	0,08661	1,37646	AGE	29,5255	1	0,0000
CIRD	0,337395	0,08146	1,40129	CIRD	36,9356	1	0,0000
CEL-0	-7,98723	1,89016	0,00033	CEL	76,0443	1	0,0000
Constant	-11,177	2,47924					
AGE	0,221478	0,05806	1,24792	AGE	22,5223	1	0,0000
CIRD	0,27303	0,06150	1,31394	CIRD	33,842	1	0,0000
CEL-0	-5,64065	1,16173	0,00355	CEL	64,0071	1	0,0000
Constant	-7,16711	1,60754					
AGE	0,133789	0,03569	1,14315	AGE	16,7686	1	0,0000
CIRD	0,146313	0,04042	1,15756	CIRD	15,711	1	0,0001
CEL-0	-1,90989	0,50245	0,14809	CEL	16,8345	1	0,0000
Constant	-7,05805	1,62331					
AGE	0,119955	0,03438	1,12745	AGE	14,1924	1	0,0002
CIRD	0,15557	0,04191	1,16832	CIRD	16,4657	1	0,0000
CEL-0	-1,87092	0,49519	0,15398	CEL	16,4205	1	0,0001
Constant	-10,565	2,35537					
AGE	0,23943	0,05954	1,27053	AGE	28,3616	1	0,0000
CIRD	0,2161	0,05382	1,24123	CIRD	24,6238	1	0,0000
CEL-0	-5,30962	1,12046	0,00494	CEL	59,8911	1	0,0000
Constant	-8,55789	1,92926					
AGE	0,163137	0,04635	1,1772	AGE	16,7961	1	0,0000
CIRD	0,217977	0,04990	1,24356	CIRD	28,4366	1	0,0000
CEL-0	-4,27303	0,84652	0,01393	CEL	51,4505	1	0,0000
Constant	-5,64467	1,42814					
AGE	0,103277	0,03250	1,1088	AGE	11,2856	1	0,0008
CIRD	0,122385	0,03745	1,13019	CIRD	12,1086	1	0,0005
CEL-0	-1,80914	0,47412	0,16379	CEL	16,6206	1	0,0000
Constant	-7,11323	1,63255					
AGE	0,123955	0,04111	1,13222	AGE	14,2396	1	0,0000
CIRD	0,23357	0,04263	1,17888	CIRD	16,4756	1	0,0000
CEL-0	-1,89992	0,50122	0,14536	CEL	16,4322	1	0,0000

Table 2. Estimations of unknown LOGIT models parameters

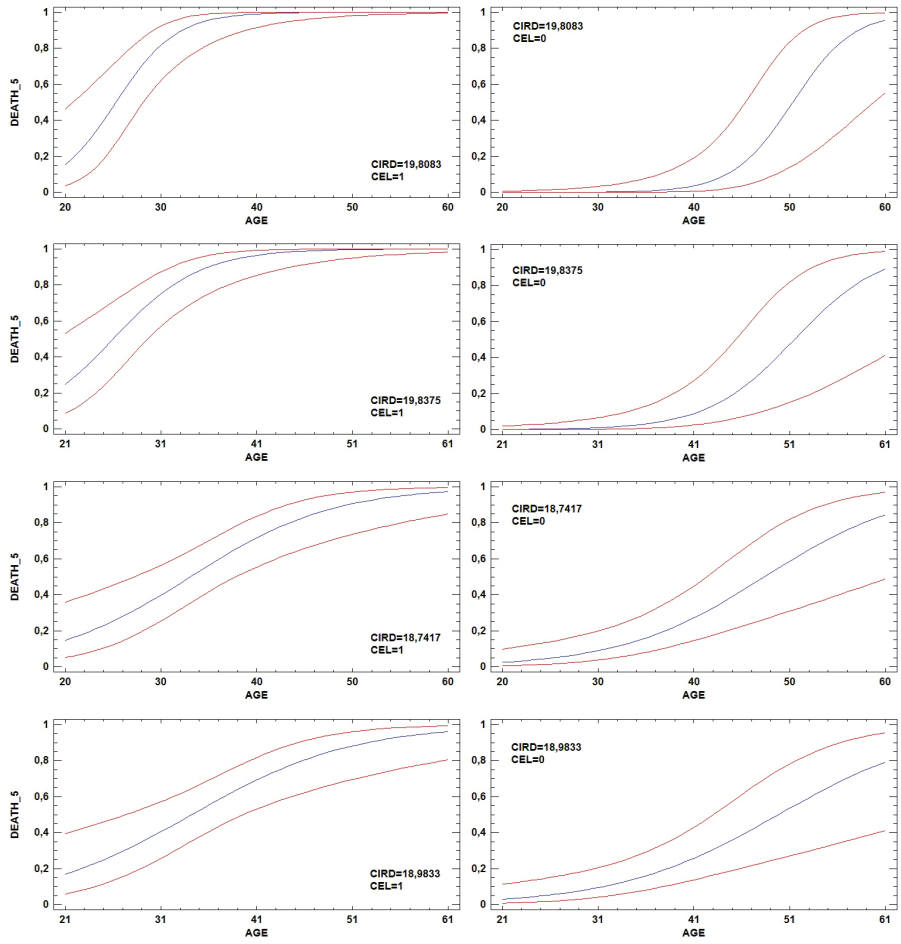

Fig. 2. Probability of death of x-years old person in Slovakia (see legend in FIG. 1.)

4 Conclusion

The aim of this study was to analyse the probability of death of x-year old persons in Czech Rep., Slovakia and Poland during next five years ($k = 5$) after the general medical examination in 1990 and 1995. The analyses were solved using LOGIT models and tried to confirm the hypothesis claiming, that the probability of death of x-year old person suffering from celiac disease decreased few years after the gaining of another medical knowledge from other countries. Even if some assumptions for the application of methods of mathematical statistics are broken, it is possible to say, that the key hypothesis was confirmed. Looking at Fig. 1, 2 and 3 we can see only slight differences between the presented countries. Their development of the compared statistics in the past should be similar.

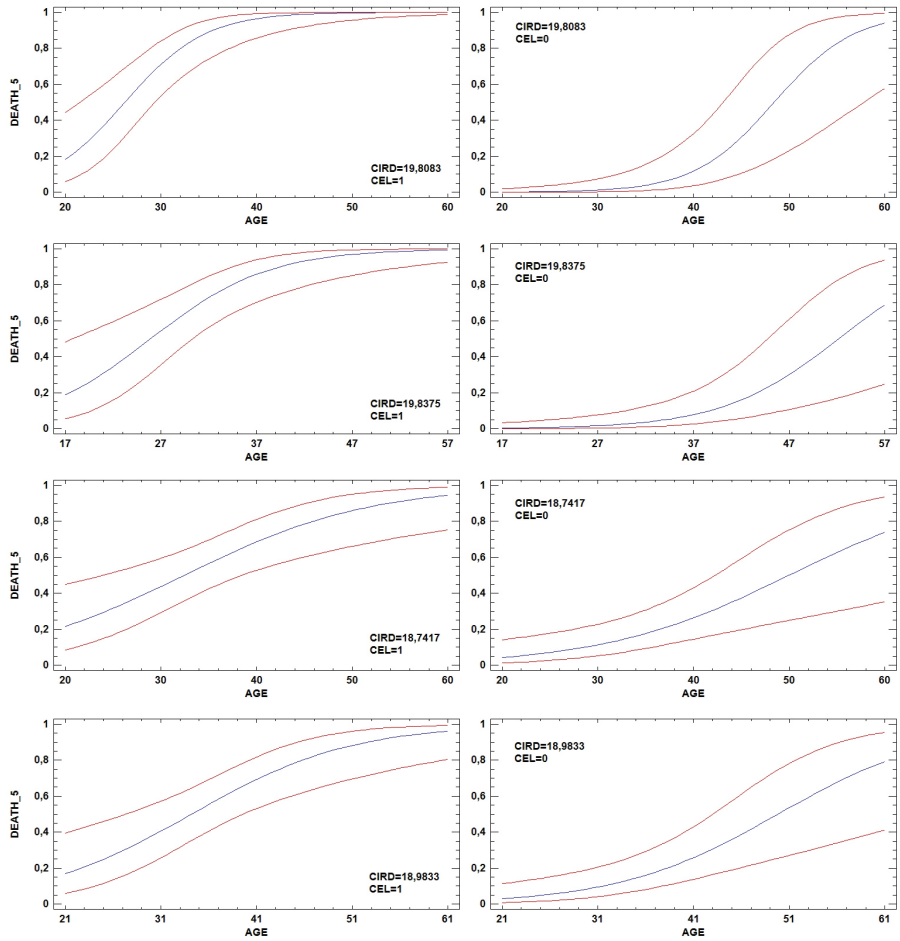

Fig. 3. Probability of death of x-years old person in Poland (see legend in FIG. 1.)

References

1.J. Freese and J. S. Long. Regression Models for Categorical Dependent Variables Using Stata, *College Station: Stata Press,* 2006.
2.D. Hoyos, P. Mariel and J. Meyerhoff. Comparing the performance of different approaches to deal with attribute non-attendance in discrete choice experiments: a simulation experiment, *BILTOKI 201001*, Universidad del Pais Vasco, 2010.
3.R. Christensen. Log-Linear Models *Springer-Verlag*, New York, 1990.
4.R. F. Logan, E. A. Rifkind, I.D. Turner and A. Ferguson. Mortality in celiac disease. *Gastroenterology*, 97(2), 265–271, 1989.
5.A. Rubio-Tapia et al. Increased Prevalence and Mortality in Undiagnosed Celiac Disease. *Gastroenterology*; 137 : 88–93. 2009.
6.L. C. Spector and M. Mazzeo. Probit Analysis and Economic Education, *Journal of Economic Education*. Spring, 11, pp. 37-44. 1980.
7.CH. W. Yang and R. D. Raehsler. An Economic Analysis on Intermediate Microeconomics: An Ordered PROBIT Model, *Journal for Economic Educators*, Volume 5, No. 3, Fall. 2005.