# GENERAL LOGIT AND PROBIT MODEL IN PROBABILITY ANALYSIS OF DEATH

**Ondřej Šimpach**

**Abstract**

The probability of death and with it the hope of survival 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. The submitted study will provide a look at the alternative assessment of the probability of death of persons in general, who suffer from any disease. The modelling of the probability of death of persons with any disease is possible with the use of the LOGIT and PROBIT models of discrete selection, which are gaining considerable popularity at present in such applications as marketing, banking and insurance. 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. The LOGIT or PROBIT models are capable of estimating, with the use of the distribution function of logistic or normal distribution, the value of the probability of death of a person *x* years of age, where further supplementary information may create various forms of the probability function. On the basis of supplementary information about the population it is then possible to construct various probability scenarios with the utilisation of alternative variables.

**Key words:** LOGIT, PROBIT, probability of death, disease

**JEL Code:** C18, C35

## Introduction

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. The LOGIT or PROBIT models are capable of estimating, with the use of the distribution function of logistic or normal distribution, the value of the probability of death of a person *x* years of age, where further supplementary information may create various forms of the probability function. In the model presented the probability of death of a person *x* years old will be estimated for the course of the next *k* years from 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.

The authors Spector and Mazzeo (1980) put together an example, where they estimated the probability with which a student will succeed in the exam. Based on this example probabilistic LOGIT and PROBIT models were created, which are currently used by many authors in their calculations and publications, such as Hoyos et al. (2010) or Yang and Raehsler (2005) in microeconomic analysis.

# 1 Methodology

The explained variables of *Y* models 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 and PROBIT models of discrete selection will be applied, when the explained variables acquire values from the interval < 0 ; 1 >. The following variables may be used for the model:

- AGE – the precise current age of the person invited for a health check,
- CIRD – 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,

**Tab. 1: Replies of patients to doctor's questions during general examination**

| | $\Psi_1$ | $\Psi_2$ | $\Psi_3$ |
|---|---|---|---|
| Smoker | no | occasionally | regularly |
| Black coffee | no | occasionally | regularly |
| Alcohol | no | occasionally | regularly |
| Sleep | regular | irregular | poor |
| Nutrition | regular | irregular | poor |

Source: author's construction

and where instead of the verbal replies given there were recorded $\Psi_{i,j}$, 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.

**Tab. 2: Replies of patients to doctor's questions during general examination in the format "0/1"**

| | $\Psi_1$ | $\Psi_2$ | $\Psi_3$ |
|---|---|---|---|
| Smoker | $\Psi_{i,j}$ = 0/1 | … | … |
| Black coffee | … | | |
| Alcohol | … | | |
| Sleep | … | | |
| Nutrition | … | | |

Source: author's construction

From Table 2, in which the replies are recorded in the 0/1 format, emerges the CIRD for the *i*-th patient from the formula (1),

$$CIRD = \left( w_1 \cdot \sum_{i=1}^{5} \psi_{\mathrm{i},1} \right) + \left( w_2 \cdot \sum_{i=1}^{5} \psi_{\mathrm{i},2} \right) + \left( w_3 \cdot \sum_{i=1}^{5} \psi_{\mathrm{i},3} \right) \tag{1}$$

where $w_1$, $w_2$ and $w_3$ are the weights recommended on the basis of the doctors' opinion. The general rule, arising from the literature, is not here. It is possible to use $w_1 = 1$, $w_2 = 3.5$ and $w_3 = 7$.

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

The probability function for the LOGIT model (see e.g. Christensen (1990)) is

$$P_i = E(Y = 1 \mid \mathbf{X}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}' \mathbf{X}_i)}}, \tag{2}$$

modified for this study in the form

$$P_i = E(Y = 1 \mid \mathbf{X}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 AGE_i + \beta_2 CIRD_i + \beta_3 ILL_i)}}, \tag{3}$$

where *i* is the *i*-th person. Let us designate it

$$Z_i = \beta_0 + \boldsymbol{\beta}' \mathbf{X}_i, \tag{4}$$

and let us insert it for the purposes of this study

$$Z_i = \beta_0 + \beta_1 AGE_i + \beta_2 CIRD_i + \beta_3 ILL_i, \tag{5}$$

and 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 of 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

$$\ln\frac{P_i}{1-P_i}=Z_i=\beta_0+\boldsymbol{\beta}'\mathbf{X}_i, \tag{9}$$

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

$$\ln\frac{P_i}{1-P_i}=Z_i=\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i. \tag{10}$$

From the general entry of the logarithm of the credibility function

$$\ln L(\beta_0,\boldsymbol{\beta})=\sum_{i=1}^{N}\left[Y_i\ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right)+(1-Y_i)\ln\left(1-\frac{e^{Z_i}}{1+e^{Z_i}}\right)\right], \tag{11}$$

there arises after the substitution

$$\ln L(\beta_0,\boldsymbol{\beta})=\sum_{i=1}^{N}\left[Y_i\ln\left(\frac{e^{\beta_0+\boldsymbol{\beta}'\mathbf{X}_i}}{1+e^{\beta_0+\boldsymbol{\beta}'\mathbf{X}_i}}\right)+(1-Y_i)\ln\left(1-\frac{e^{\beta_0+\boldsymbol{\beta}'\mathbf{X}_i}}{1+e^{\beta_0+\boldsymbol{\beta}'\mathbf{X}_i}}\right)\right] \tag{12}$$

and for the purposes of this study

$$\ln L(\beta_0,\beta_1,\beta_2,\beta_3)=\sum_{i=1}^{N}\left[\begin{array}{l} DEATH\_K_i\ln\left(\frac{e^{\beta_0+\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i}}{1+e^{\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i}}\right) \\ +(1-DEATH\_K_i)\ln\left(1-\frac{e^{\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i}}{1+e^{\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i}}\right)\end{array}\right]. \tag{13}$$

For the PROBIT model (see e.g. Freese and Long, (2006)) we utilise the general entry of the logarithm of the credibility function

$$\ln L(\beta_0,\boldsymbol{\beta})=\sum_{i=1}^{N}\left[Y_i\ln(F(Z_i))+(1-Y_i)\ln(1-F(Z_i))\right] \tag{14}$$

after substitution

$$\ln L(\beta_0,\boldsymbol{\beta})=\sum_{i=1}^{N}\left[Y_i\ln(F(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i))+(1-Y_i)\ln(1-F(\beta_0+\boldsymbol{\beta}'\mathbf{X}_i))\right], \tag{15}$$

and for the purposes of this study

$$\ln L(\beta_0,\beta_1,\beta_2,\beta_3)=\sum_{i=1}^{N}\left[\begin{array}{l} DEATH\_K_i \ln\left(F(\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i)\right) \\ +(1-DEATH\_K_i)\ln\left(1-F(\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i)\right)\end{array}\right]. \tag{16}$$

The distribution function is then

$$F(\beta_0+\mathbf{\beta' X}_i)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta_0+\mathbf{\beta' X}_i} e^{\frac{-z^2}{2}}dz, \tag{17}$$

and after substitution for the purposes of this study

$$F(\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta_0+\beta_1 AGE_i+\beta_2 CIRD_i+\beta_3 ILL_i} e^{\frac{-z^2}{2}}dz. \tag{18}$$

## 2 Parameter estimation

Estimating the unknown parameters of nonlinear regression models is no problem today. The software uses an iterative method. The software itself selects the initial value. To estimate the parameters of LOGIT and PROBIT model is recommended to use Statgraphics Centurion or other statistical software package, because it is not necessary to create own script or source code. In the case that the analyst receives representative matrix data with a sufficient number of observations, it is possible to construct specific scenarios of probabilities of death of *x*-years old person. The data matrix can be requested (with certain limitations) from databases from health insurance corporations or it is possible directly to let assemble by particular doctor or medical facility.

## Conclusion

The model that was presented will be used in the future in various analyses. Suitable modification of the presented model would lead also to alternative options for calculating the probability of death of *x*-years old persons in life tables of the specific population, or even of the overall population. With the increasing number of additional information increases the possibility of using that model. Additional information that leads to the transfer of certain variables or demographic events to the situation "0" - it happened, "1" - did not happen, provide an opportunity to construct scenarios.

## References

FREESE, J., LONG, J.S.: „Regression Models for Categorical Dependent Variables Using Stata“, *College Station: Stata Press*, 2006.

HOYOS, D., MARIEL, P., MEYERHOFF, J.: „Comparing the performance of different approaches to deal with attribute non-attendance in discrete choice experiments: a simulation experiment“, *BILTOKI 201001*, Universidad del País Vasco - Departamento de Economía Aplicada III (Econometría y Estadística), 2010.

CHRISTENSEN, R.: „*Log-Linear Models*“ Springer-Verlag, New York, 1990.

SPECTOR, L.C., MAZZEO, M.: „Probit Analysis and Economic Education“, *Journal of Economic Education*. Spring, 11, 1980, pp. 37–44.

YANG, CH.W., RAEHSLER, R.D.: „An Economic Analysis on Intermediate Microeconomics: An Ordered PROBIT Model“, *Journal for Economic Educators*, Volume 5, No. 3, Fall 2005.

**Contact**

Ondřej Šimpach, Ing.

University of Economics Prague, Department of Demography

ondrej.simpach@vse.cz