USING SURVIVAL ANALYSIS IN ECONOMICS

Ana-Gabriela BABUCEA
Faculty of Economics and Business Administration
Constantin Brancusi University of Tirgu-Jiu
babucea@utgjiu.ro

Daniela-Emanuela DANACICA
Faculty of Economics and Business Administration
Constantin Brancusi University of Tirgu-Jiu
daniela.danacica@gmail.com

Abstract

The main objective of this paper is to examine methodological and applicative problems of survival analysis in the analysis of socio-economic phenomena.

Although at the beginning the survival analysis was used to study death as an event specific to medical studies, as from the '70s these statistical techniques have been increasingly used in economics and social sciences. Besides the fact that survival data are not normally distributed, they often contain incomplete information, censored subjects. Censoring of subjects may be on the right or left. It is vital to include censored subjects in the statistical analysis. But, according to Greene (2003), a very large number of censored subjects may affect the accuracy of statistical tests. We presented in this paper methodological aspects of Kaplan-Meier analysis, and statistical significance testing for the resulted survival curves. We have also concentrated on the Cox regression, and we have set out the concept of hazard, baseline hazard, hazard rate, hazard rate interpretation. An application of the survival analysis in unemployment is presented.

Keywords: survival analysis, Kaplan-Meier method, censored subjects
JEL classification: C41, J01, J21

1. METHODOLOGICAL PROBLEMS OF SURVIVAL ANALYSIS

The first use of survival analysis and duration models comes from medical research. Survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature. Although at the beginning the survival analysis was used to study death as an event specific to medical studies ([1], [19]) and demographical studies ([3], [5], [8]), as from the '70s these statistical techniques have been increasingly used in economics and social sciences. Survival data requires a different statistical analysis compared to the quantitative data due to their particularities. Besides the fact that survival data are not normally distributed, they often contain incomplete information, censored subjects. Censoring and abnormality distribution of survival data raises

specific methodological and statistical techniques required for an adequate analysis. Censoring of subjects may be on the right or left. It is vital to include censored subjects in the statistical analysis. But, according to [9] a very large number of censored subjects may affect the accuracy of statistical tests. A detailed analysis of survival is presented by [24] and [11], [15].

The survival analysis can be used for socio-economic research to investigate complex phenomena such as unemployment, employment, inflation, supply and demand for bank loans, life expectancy of the products, the producer and consumer, etc. Survival analysis, adapted in conventional econometric modelling data, received the title of duration models ([16], [14], [18]).

The advantages of survival analysis can be grouped as follows:

a) descriptive, for the sample of subjects examined;
b) predictive for the representative population of subjects;
c) as a comparative method being objective and presenting a high accuracy.

Using survival analysis involves the simultaneous observance of the following conditions:

a) the researcher must analyze empirical data for all subjects in the study. Subjects who didn't achieve the pre-established event will be censored on the right. All right censored subjects are lost from the statistical observation.

In Figure no. 1 we have the graphical representation of survival data. On the abscissa we have presented the observation period of subjects, measured in months, and on the ordinate we have presented the observed subjects. Horizontal segments represent the periods of tracking subjects, and x is marked with the event default. It is noted that subjects B and C are right censoring, assuming that the survival time was 80 months.

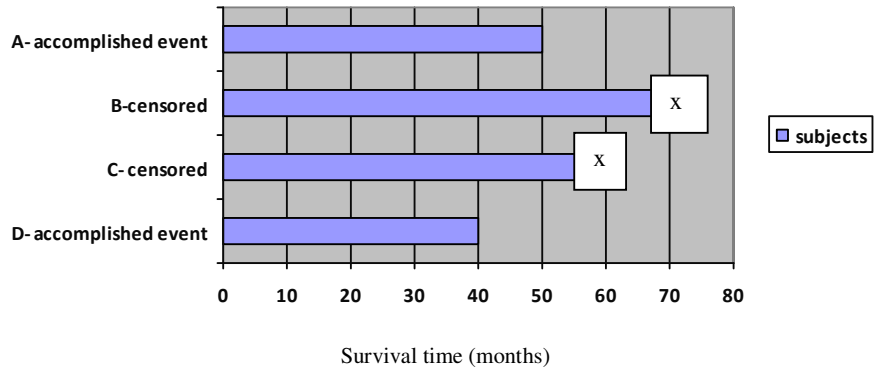

Figure no. 1 Graphical representation of survival data

b) It happens sometimes that subjects are not taken into study at the same time. If subjects do not entry simultaneously in the study, we will have progressively censored data. (Figure no. 2)

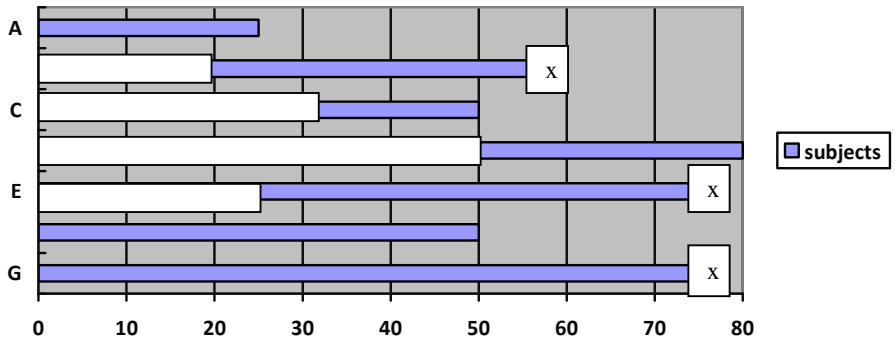

Figure no. 2 Graphical representation of survival data

The study began with A, F and G subjects; then entered in the study the subject B, 15 months after its start, followed by C to 25 months from the start and subjects D and E at 40 and 20 months from the start of the study. Subjects B, E and G have not accomplished the event during the study.

1.1. KAPLAN –MEIER METHOD

The Kaplan-Meier method is a nonparametric (actuarial) technique for estimating time-related events (the survivorship function). Ordinarily it is used to analyze death as an outcome, in biostatistics, but in recent years these techniques have also gained popularity in the social sciences or industrial statistics (an economist might measure the length of time people remain unemployed after a job loss or an engineer might measure the time until failure of machine parts). A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant.

An important advantage of the Kaplan-Meier curve is that the method can take into account "censored" data — losses from the sample before the final outcome is observed (for instance, if a patient withdraws from a study). On the plot, small vertical tick-marks indicate losses, where patient data has been censored. When no truncation or censoring occurs, the Kaplan-Meier curve is equivalent to the empirical distribution.

Kaplan Meier method presupposes a greater reduction in calculus volume than the actuarial method, because survival is estimated every time when the pre-established event for a subject occurs (employment in our case), thus neglecting the registrations lost of sight along the survey.

The stages of Kaplan-Meier method are:

- listing the time when the pre-established event occurs, since subject's involvement in the survey (participation time);
- finding for every participation time the number of subjects that continue to participate in the survey – those who did not achieve the pre-established event (employment in our case);

- establishing the number of subjects who achieved the pre-established event within *nx* time interval;
- the calculus of the probability of occurrence of the pre-established event, for each participation interval *(dx)* according to the formula: *qx=dx/nx*, where *x* is the participation duration;
- as for the actuarial method, the calculus of survival probabilities for *x* duration is: *px=1-qx*, and the cumulated survival probability is *Px=px(px-1)(px-2)....p2p1*.

The survival rate is expressed as the *survival function* $S(t)$:

$$S(t) = \frac{\text{number of the unemployment spells surviving until time t or longer}}{\text{total number of unempl. spells observed}} \tag{1}$$

The survival function $S(t)$ denotes the probability of unemployment duration until time t or longer and is given by

$$S(t) = P(T \geq t) = 1 - F(t), \tag{2}$$

where T denotes survival time – duration of unemployment spell, and $F(t)$ is distribution function of T. $F(t)$ measures the probability time of survival – unemployment duration up to time t.

The product limit method of Kaplan and Meier (1958) is used to estimates S:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{3}$$

where t_i is the survival time – duration of unemployment spell at the point i, d_i is the number of ends of unemployment spell up to time t_i and n_i is the number of cases of unemployment spells at risk just prior to t_i. The survival function is based upon probability that a case of unemployment spell survives at the end of a time interval, on the condition that the individual was present at the start of the time interval. The survival function is the product of these conditional probabilities.

The Kaplan-Meier technique is usually only useful as a method of preliminary evaluation, since it is purely a descriptive method for the evaluation of one variable. The survival curve of this method is scalariform because the proportion of subjects who have the chance to continue observation without the occurrence of the pre-established event changes exactly at the moments when the pre-established event is achieved. The survival level is of 100% from the curve origin until the moment of the first occurrence of the event (employment in our case), where it drops to the new calculated value, that constitutes a new level during which survival is constant, until the next event achieved. Therefore, every step corresponds to the occurrence of one or several pre-established events.

1.2. COX REGRESSION

Cox regression can be used to determine whether a characteristic of subjects affecting the survival and, if so, how much and in what direction (to increase or decrease). Survival prediction can be difficult if not taking into account all factors that influence it. It is therefore necessary to identify those variables that affect the survival and that can be used in the calculation of a predictive indicator of survival. A method to determine such an indicator, and associated survival curve, is called Cox proportional-hazard regression. Cox proportional-hazard model is a semi-parametric method that enables to determine the effect of different variables on the hazard. Assuming that we have "*n*" units' individual under observation, then the model has the form:

$$\lambda_i(t) = e^{x_i'\beta} \cdot \lambda_0(t) = c_i \cdot \lambda_0(t), \quad i = 1, 2, \ldots, n \tag{4}$$

where $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ik}\right)'$ is the vector of variables factor, $\beta = \left(\beta_1, \beta_2, \ldots, \beta_k\right)'$ is the vector of regression coefficients, $\lambda_i(t)$ is the hazard calculated for each individual *i* and $\lambda_0(t)$ is the baseline hazard**.** The baseline hazard function corresponds in our case to the probability of employment when all the explanatory variables are 0. The proportional hazards assumption is the assumption that affect parameters multiply the hazard. The ratio of the hazard function is named the hazard ratio. A comprehensive introduction to the Cox model is given, by [11] and [15]. Hazard and the probability of survival are interrelated; the relationship is complex, but sufficient to be able to determine a parameter with the other. If the hazard is constant for the entire duration of study, means that the risk of death is independent of the duration of survival of an individual.

The proportional hazards assumption is crucial for the Cox regression model. The proportional hazards assumption can be checked using the *log-minus-log* curve or with the help of *partial (Schönfeld) residuals.* In the first case, if baseline hazards are proportional, then the lines corresponding to individual layers must be parallel; in the second case proportional hazards assumption requires that in the *Schönfeld* graph there should be no *pattern*. If proportional hazard assumption is violated, then we shall build a Cox model with non-proportional hazard by entering an interaction between the specific covariate and time.

2. MODELLING TIME OF UNEMPLOYMENT WITH SURVIVAL ANALYSIS

In this section we analyzed the influence of exogenous variables *gender*, *age* and *educational level* for the duration of unemployment in Gorj County, using the survival analysis. We used the data made available by the National Agency for Employment (NAE) Bucharest; the sample was made of 29544 registrations, with information concerning the start date and the end date of unemployment spells, gender, age and level of education and the reason of unemployment leaving for each registered person during 1st January 2005 – 4th August 2006. 39.6 percent represents female unemployment and 60.4% represents male unemployment. The minimum duration of unemployment spells is 0 days and the maximum is 675 days; the average duration of the unemployment spells is 155 days. The fluctuation around

the mean duration of the unemployment spells is presented by the dispersion measured, with a skewness of 0.731 and a kurtosis of 0.195.

Regarding the factor *gender*, we noticed that the male unemployment in Gorj County for the analyzed period is higher than the female unemployment, and for the unemployed men it lasts longer than for unemployed women (the more the unemployment period lasts, the more differences between male and female unemployment increase). Taking into account the fact that the number of women in Gorj County that are able to work is higher than the number of men in 2005 and 2006, we draw the conclusion that differences between the number of women registered as unemployed and the number of men are a direct consequence of the continuous reorganization, after 1992, of the mining sector, thermo energetic and oil tanker in the Gorj County area, with negative effects on men belonging to all educational levels, employed in these jobs.

As for the factor *age*, the average age of the persons registered in the database is of 33 years, and the median is of 32 years. Most of the unemployed registered in the database are aged between 15-35 years; the youngest subject is 15 years and the oldest is 62. The high number of young unemployed registered in Gorj County shows that young people cannot find a job after finishing their studies, as the labor market in the county is not ready to receive them. The age distribution is positively skewed. In our analysis we divided the variable age into five groups (group 1 -15-24 years, group 2 - 25-34 years, group 3 - 35-44 years, group 4 - 45-54 years and group 5- 55-64 years), according to the Romanian Year Book 2005 and the methodology of Romanian Institute for Statistics). As we can notice from Table 1, there is a positive correlation between age and duration of unemployment: with the age increasing the duration of unemployment increase.

Regarding the *level of education*, 7.8% registered in the database are university graduates, 2.5% graduated from post high school, 21.1% graduated from specialty high school, 13.5% graduated from theoretical high school, 0.3% are special education graduates, 24.7% have vocational school, 5% graduated from foremen school, 6.1% are apprenticeship complementary education graduates, 16.4% graduated only from secondary school, the educational level for 2.1% is unfinished secondary school, and 0.5% are without education. In data processing we have grouped persons by their educational level in 5 groups: group 0 - without graduated school, group 1- unfinished secondary school, secondary school, vocational school, apprenticeship complementary education, special education, with the maximum number of 10 years of study, group 2- theoretical high school, specialty high school, with 12 respectively 13 years of study, group 3 – foremen school and post high school with 14 years of study and group 4 corresponding to university education, (with short form – college), with 15, 16 and respectively 17 years of study (according to the methodology of the Ministry of Education Romania). As we can notice from Table 1, there is a negative correlation between level of education and duration of unemployment.

In table 1 we have descriptive statistics for the duration of unemployment spells in days and the variables gender, age, and level of education.

Table no. 1 - Descriptive statistics for the duration of unemployment spells (in days)

	N	MEAN	STD. DEV.	95% CONFIDENCE INTERVAL FOR THE MEAN
Total	29544	155.30	132.10	(153,54, 157,0)
Factor Gender				
Male	17831	157.24	138.17	(154.93-159.54)
Female	11713	151.78	118.17	(149.15-154.42)
Factor: Education				
Level 0 – without education	141	245.53	188.34	(209.09-281.98)
Level 1	14643	173.42	131.89	(170.91-175.94)
Level 2	10231	143.54	129.42	(140.65-146.42)
Level 3	2212	139.71	142.93	(132.48-146.95)
Level 4	2317	106.59	108.15	(101.57-111.61)
Factor: Age				
15-24 years	9505	116.30	94.56	(114.24-118.36)
25-34 years	7161	157.30	127.13	(153.94-160.65)
35-44 years	6841	187.11	150.36	(182.69-191.52)
45-54 years	5401	198.51	159.29	(193.05-203.97)
55-64 years	630	214.56	159.86	(198.29-230.83)

The result of the Kruskal –Wallis test allowed us to reject the null hypothesis. The differences noticed for each of the levels of the factors gender, age and level of education, regarding the mean duration of unemployment spells are statistically significant.

2.1. RESULTS OF THE KAPLAN – MEIER ANALYSIS

In figure 3 there is presented the survival curve for the women (0) and men (1) in the database. The results suggest a significant difference in probabilities of remaining unemployed between female and male; the median unemployment duration for female is 180 days and for male is 185 days. After 600 days the curves coincide.

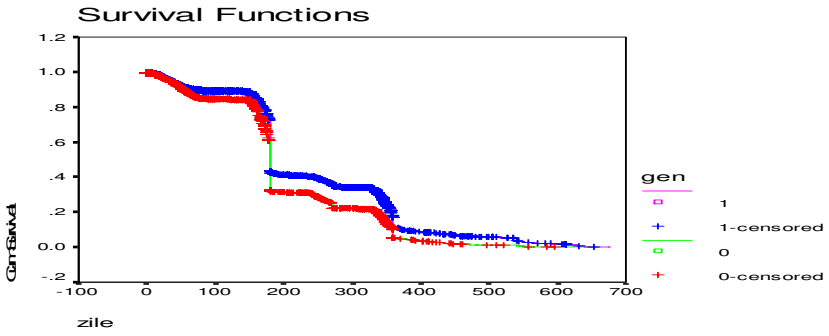

Figure no. 3 Survival function estimates for female and male unemployed

In Figure no. 4 there is presented the survival curve for the age groups 15-24 years, 25-34 years, 35-44 years, 45-54 and 55-64 years. Applying Kaplan-Meier analysis we have:

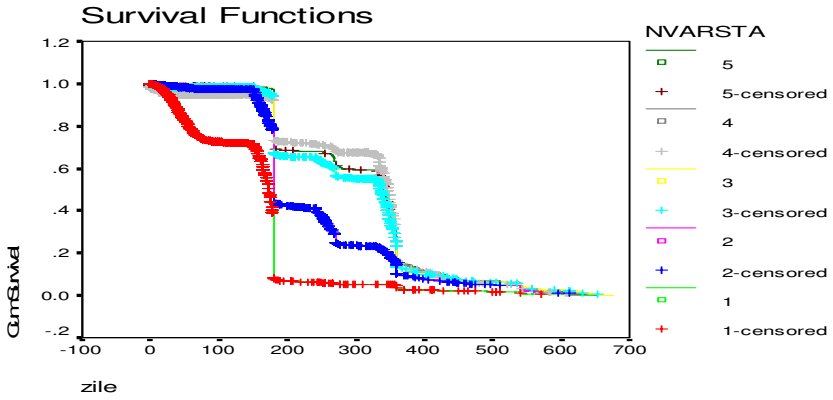

Figure no. 4 Survival function estimates for the age groups 15-24 years, 25-34 years, 35-44 years, 45-54 and 55-64 years

We can notice that the probability of remaining unemployed increased with age. The older persons are disadvantage on the labor market of Gorj County. The median unemployment duration for the age group 15-24 years is 173 days; for the age group 24-34 years is 180 days, for the age group 35-44 years is 336 days, for the age group 45-54 is 346 days and for the age group 55-64 is 346 days. The differences observed are statistically significant.

In Figure no. 5 there is presented the survival curve for the level of education. Applying Kaplan-Meier analysis we have:

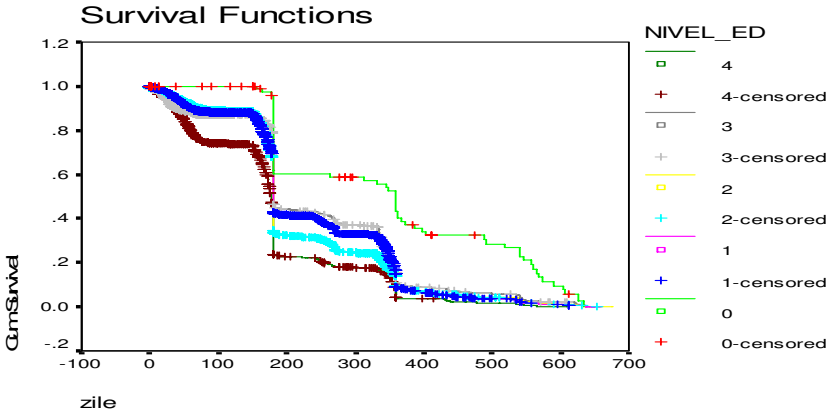

Figure no. 5 Survival function estimates for the five groups of education

We can notice that the probability of remaining unemployed is higher for the persons without education, followed by the persons with foremen school and post high-school and the lowest probability of remaining unemployed is for the persons with university education. We can also notice that after 600 days unemployment curves start to coincide and the educational level no longer influences the probability of finding a job.

The result of the log rank test with Chi-Squared distribution under the null for all three factors, confirm the results derived graphically from the Kaplan-Meier estimates of the survival functions.

2.2. COX ANALYSIS RESULTS

The empirical analysis was performed with the SPSS 10.0 program package. The factor gender was coded as 1 for the male unemployed and 0 for the female unemployed. The age variable was divided into 5 intervals, in conformity with the Romanian Year Book 2005 and with the Romanian statistical methodology. As for the variable education, we have grouped educational levels as it follows: group 0 - without education; group 1: unfinished secondary school, secondary school, vocational school and apprenticeship complementary education and special education; group 2: theoretical high school, specialty high school, group 3: foremen school and post high school; group 4: college, university education. It is obvious that the duration of unemployment is higher for the lower levels of education and lower for the higher levels of education.

The reference category of covariates was the last category, and the Enter method was selected; all of the predictors were specified in the model simultaneously.

The results of the omnibus tests of the model coefficients are given in Table 2. The obtained results allow us to reject the null hypothesis $H_0 : \beta = 0$.

Table no. 2 - Omnibus tests of the model coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
123126.523	499.055	3	.000	476.913	3	.000	476.913	3	.000

a Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 123603.436
b Beginning Block Number 1. Method = Enter

In Table 3 are presented the results of the Cox regression analysis: B is the estimate vector of the regression coefficients, $\text{Exp}(B_p)$ is the predicted change in the hazard for each unit increase in the covariate.

Table no. 3 - Variables in the equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gender	.333	.028	144.537	1	.000	1.395	1.321	1.472
Age	.068	.011	36.620	1	.000	1.070	1.047	1.094
Education			334.682	4	.000			

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Education(1)	-.708	.043	264.921	1	.000	.493	.452	.537
Education(2)	-.411	.044	86.401	1	.000	.663	.608	.723
Education(3)	-.274	.059	21.803	1	.000	.760	.678	.853

As we can notice from table 3 the hazard for the unemployment spell to end is 39.5% higher for the unemployed female that for the unemployed male. With increased age, the hazard is reduced by 0.7% each year. All levels of education have significant hazard ratios of less than 1; the hazard ratio is the lowest for the level 1 - unfinished secondary school, secondary school, vocational school and apprenticeship complementary education, special education -0.493 and the highest for level 3 - foremen school and post high school (0.760). As we expected, the hazard ratio increased with higher levels of education.

The proportional hazards assumption is very important for the Cox regression. Using a graphically examination, the proportional hazard assumption is not violated if the survival curves for different group of individuals do not cross. We performed the log-minus-log plot and we observed that in our case the survival curves are crossed.

For the next step we performed the partial residual plot and we noticed the fact that the baseline hazards are not proportional and R squared linear indicates a positive correlation between partial residual and time, therefore the proportional hazard assumption does not hold. For the next step we used a model that includes the covariate age and the interaction term between time and age. The results of the omnibus test for the model coefficients allow us to reject the null hypothesis. We noticed from the analysis that the estimates for all the variables are almost similar to the Cox proportional hazards model from Table 3, the hazard ratios for the levels of education are slightly lower than before. The comparison between Cox proportional hazards model and Cox regression model with time-dependent covariate gives similar conclusions for all the three factors, gender, age and level of education for the analysed period.

3. CONCLUSION

The main objective of this paper was to examine methodological and applicative problems of survival analysis in the analysis of socio-economic phenomena. First section of the paper deals with methodological problems specific to survival analysis. Although at the beginning the survival analysis was used to study death as an event specific to medical studies, as from the '70s these statistical techniques have been increasingly used in economics and social sciences. Besides the fact that survival data are not normally distributed, they often contain incomplete information, censored subjects. We presented in this section methodological aspects of Kaplan-Meier analysis, statistical significance testing for the resulted survival curves and the peculiarities of various statistical tests used. We have also concentrated our attention on the Cox regression, and we have set out the concept of *hazard*, *baseline hazard, hazard rate, hazard rate interpretation*. We pointed out that the proportional hazards assumption is crucial for the Cox regression model. The proportional hazards assumption can be checked using the *log-minus-log* curve or with the help of *partial (Schönfeld) residuals.* In the second section of paper we wanted to present the applied side

of survival analysis. We focused on studying the factors influencing the duration of unemployment. The empirical analysis is based on data offered by the National Agency for Employment of Romania. The data set contains information about all the subjects registered by the National Agency for Employment in 2005 and 2006. Cox proportional hazard models and Cox regression with time- dependent covariate conducted to similar results: persons with a higher education have the best chances to become re-employed compared with the persons without education or with level 1 of education. As for age, young people aged between 15-24 years remain unemployed for 173 days on the average, unlike the group 45-54 year or 55-64 year who remain unemployed for 346 days on the average. Regarding the variable gender, the duration of unemployment is smaller for women than for men. Also the chance to be re-employed is higher for women than men, in the analyzed period.

References

[1] Armitage P, B. G (1959) *Statistical Methods in Medical Research*. Blackwell.

[2] atsiramos, Konstantinos (2006), *Unemployment Insurance in Europe: Unemployment Duration and Subsequent Employment Stability*. IZA Discussion Paper no. 2280.

[3] Balakrishnan, N. (1991). *Handbook of the Logistic Distribution*. Marcel Dekker, Inc.

[4] Berkson J, R.P. Gage (1950). *Calculation of Survival Rates for Cancer*. Proceedings of Staff Meetings of the Mayo Clinic;25:270-286.

[5] Carroll Nick (2005). *Explaining Unemployment Duration in Australia.* http://econpapers.repec.org/article/blaecorec/.

[6] *Cutler*, S J şi *Ederer* F (1958). *Maximum utilization of the life table method in analyzing. survival*. J. Chronic Dis. 8:699-712.

[7] Dănăcică, D.E., Babucea A.G. (2007). *Modelling Time of. Unemployment – A Cox Analysis Approach*, SOR'7 International Conference, Nova Gorica, Slovenia, 2007, ISBN 978-961-6165-25-9, pag. 273-279.

[8] Foley, M.C. (1997). *Determinants of Unemployment Duration in Russia*, Yale Economic Growth Center Discussion Paper 779:39.

[9] Gehan, E.A. (1969). *Estimating Survival Function for the Life Table*. Journal of Chronic Diseases, 21 629-44.

[10] Greene, William H. (2003). *Econometric Analysis.* New York: Prentice-Hall.

[11] Ham, J.C. and Rea, Jr (1997). *Unemployment Insurance and Male Unemployment Duration in Canada"*. Journal of Labor Economics, 5(3): 325-53.

[12] Kettunen, Juha. 1994. The Effects of Education on the Duration of Unemployment. *Labour* 2: 331-352.

[13] Kettunen, Juha. 1997. Education and Unemployment Duration. *Economics of Education Review* 2: 163-170.

[14] Kiefer, N.M. (1988). *Economic Duration Data and Hazard Functions*. Journal of Economic Literature 26: 646-679.

[15] Klein, J. P., and M. L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Verlag.

[16] Klein, John P., and Melvin L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Verlag.

[17] Lancaster, T. and S. J. Nickell (1980). *The Analysis of Re-employment Probabilities for the Unemployed*, Journal of the Royal Statistical Society, A, 143, 141-165.

[18] Moffitt, Robert A. 1999. New developments in econometric methods for labor market analysis. In: O. Ashenfelter, and D. Card (eds). *Handbook of Labor Economics*. Chapter 24: 1367-1397.

[19] Narendranathan, W and M. Stewart (1993). *Modelling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Baseline Hazards*, Journal of the Royal Statistical Society, Series C, Applied Statistics, 42(1), pp. 63-83.
[20] Pike, M.C. (1966). *Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient*. British Journal of Cancer 1976;34:585-612.
[21] Popelka John (2004). *Modelling Time of Unemployment via Cox Proportional Model*, paper presented at Applied Statistics 2005 International Conference, http://ablejec.nib.si/AS2005/Presentations.htm.
[22] Schomann, Klaus, and Phillip J. O'Connell. 2002. Education, training and employment dynamics: Transitional labour market in the European Union. *Labour Markets and Employment Policy Series.* Cheltehham, U.K., and Northampton, Mass.: Elgar.
[23] Tansel, A. and H. M. Tasci (2005), *Determinants of Unemployment Duration for Men and Women in Turkey*. IZA Discussion Paper no. 1258.
[24] Therneau, Terry M. and Patricia M. Grambsch (2001). *Modelling Survival Data: Extending the Cox model.* New York: Springer Verlag.