Using Kaplan – Meier curves for preliminary evaluation the duration of unemployment spells

Babucea, Ana-Gabriela and Danacica, Daniela-Emanuela

Constantin Bracusi University of Targu Jiu

2007

USING KAPLAN – MEIER CURVES FOR PRELIMINARY EVALUATION THE DURATION OF UNEMPLOYMENT SPELLS

BABUCEA ANA GABRIELA
DĂNĂCICĂ DANIELA

"Constantin Brâncuşi" University of Targu-Jiu, Romania

ABSTRACT: *In this study we present the results of basic data analysis of the duration of unemployment spells in Gorj County using Kaplan – Meier curves The database includes individual information about all the subjects registered at the county agency of Gorj county during January 1st, 2002- August 31st, 2006. The database has individual information about all the subjects registered at NAE during January 1, 2002 - August 31, 2006. Statistical data analysis as part of the ASO project "The role of education for duration of unemployment", is based on data offered by the National Agency for Employment of Romania (NAE) and was made with SPSS for Windows V. 10. 0.5.*

1. THE DATABASE

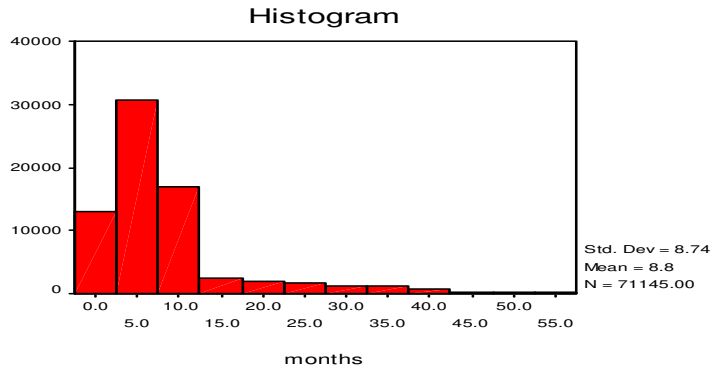
The sample contains 80961 registrations, with information concerning the start date and end date of the unemployment spells, sex, age, educational level and the reason of unemployment leaving for each registered person. Because NAE Romania does not allow disclosure of personal information concerning the persons registered as unemployed, the database contains an identification number for every registered person. 9816 registrations are considered "missing values" because we don't know their date of unemployment end, as these data were no longer studied.

The minimum duration of unemployment spells is of 0 days and the maximum duration is of 1723 days. For the analysis we express the duration of unemployment in months, we have its minimum duration – 0 months and its maximum duration of 57 months, with an average of 8.8 months and median of 6 months.

Table 1: *Distribution of unemployed by duration of unemployment spells (months)*

Group interval (months)	No. of registrations	Percent %
0-6	43378	53.6
6-12	17282	21.3
12-18	3333	4.1
18-24	2202	2.7
24-30	1648	2.0
30-36	1666	2.1
36-42	980	1.2
42-48	340	0.4
48-54	260	0.3
54-peste	56	0.1
Total	71145	87.9
Missing values	9816	12.1
Total	**80961**	**100.0**

Figure 1: *Histogram for the variable duration of unemployment spells (months)*



In table 2 are present the descriptive statistics for the duration of unemployment spells in months and the variables sex, educational level and age.

Table 2: *Descriptive statistics for the duration of unemployment spells (in months)*

	N	Mean	Std.Dev.	95% confidence interval for the mean
Total	**71145**	**8.82**	**8.74**	**(8.75, 8.88)**
FACTOR SEX				
Male	47691	9.32	9.56	(9.23-9.41)
Female	33270	8.03	7.17	(7.94, 8.11)
FACTOR: EDUCATION				
Level 0 – without education	440	12.78	9.26	(11.92-13.65)
Level 1 Unfinished secondary school, secondary school, vocational school and apprenticeship complementary education Special education	35683	9.16	8.53	(9.08-9.25)
Level 2 Theoretic high school, speciality high school	25456	8.77	9.01	(8.66-8.88)
Level 3 Foremen school and post high school	5012	9.69	10.16	(9.41-9.97)
Level 4 college, university education	4554	5.05	5.44	(4.89-5.21)
FACTOR: AGE				
15-24 years	24015	6.03	6.27	(5.95-6.11)
25-34 years	18960	9.30	9.38	(9.17-9.44)
35-44 years	15338	10.53	9.31	(10.38-10.68)
45-54 years	11727	11.17	9.46	(11-11.35)
55-64 years	1105	12.47	10.56	(11.84-13.05)

We used the non-parametric Kruskal-Wallis test in order to test the null hypotheses that the mean duration of unemployment spells is the same for each of the levels of the factors sex, age and level of education. The null hypothesis is strongly rejected for each of the factors since the p-values are lower that 10^{-6}.

Of 80961 persons registered in the database of Gorj County as unemployed, during 1.01.2002-31.08.2006, 19369 persons found a job, until August 31st, 2006; in the database, the reason of their unemployment leaving was filled in with "employed". The average duration of unemployment until finding a job is of 6 months, median of 2 months, maximum value 57 months and minimum value 0 months. Of these 6390 persons (33%) are women and 12979 (67%) are men. If the percentage difference between the male unemployment and the female unemployment registered in the database is of 17.8%, the percentage difference between the number of men who become employed and the number of women in the same situation is of

34%, which shows that although there are more unemployed men, and they stay unemployed about one more week than women, on the average, however they are preferred by employers. Regarding the distribution by age, 30.1% of employees are from the 25-34 years group, followed by the group of 15-24 years with 25.7 percent, 35-44 years with 25.4 percent, 45-54 years with 17.5 percent and over 55 years group with 1.3%. 0.4 percent of employees are persons without education, 44.7 percent have no more than 10 years of education, and 37.4 percent are graduates of high school; 8.6 percent of employees are persons with apprenticeship complementary education, and 8.9 percent are persons with university education.

2. USING 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 analyse 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 $Px=px(px-1)(px-2)....p2p1$.

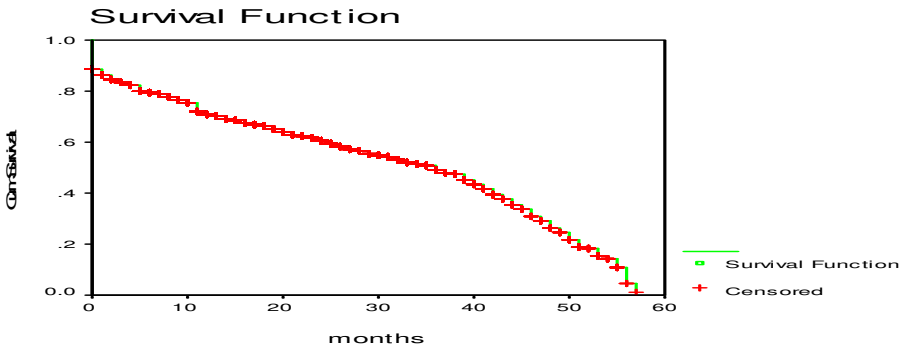
The Kaplan-Meier technique is usually only useful as a method of preliminary evaluation, since it is purely a descriptive method for the evaluation of one variable. The survival curve of this method is scalariform because the proportion of subjects who have the chance to continue observation without 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.

For our survey the pre-established event is employment, this event being ascribed the value 1, the number of the subjects who achieved the event at the end of the analysed period being of 19369, representing only 23.9% of the total of subjects; the rest of 61592 subjects representing 76.1% of the total either did not achieve the event, or their track has been lost (they don't have the date of unemployment leaving), they have been censored at the right side, being ascribed the value 0.

In figure 2 there is presented the survival curve Kaplan–Meier for the entire database, the pre-established event being employment; the censored subjects are marked with red.

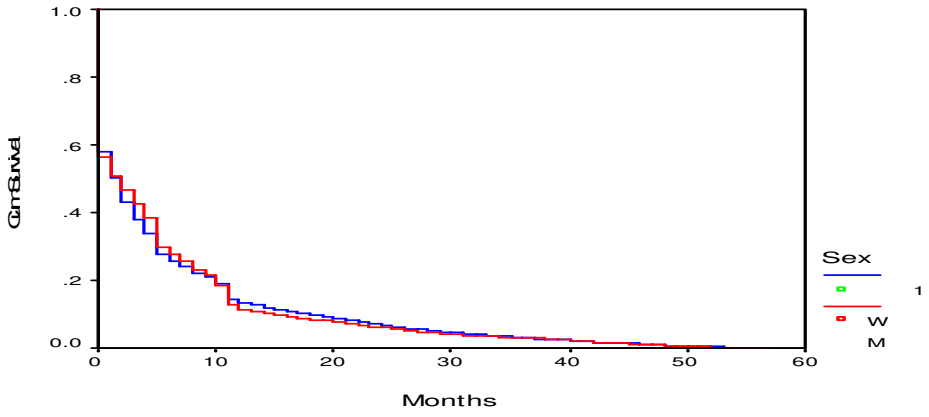
Figure 2: *Survival function estimates for all unemployed persons*



In order to analyse to potential influence of sex, age and educational level variables on the duration of unemployment until the moment of employment we have created a new database only for the subjects who have achieved the pre-established event, **the employment** (19369 registrations).

In figure 3 there is presented the survival curve for the women and men in the database. The qualitative sex variable has been codified, 1 representing men, 0 representing women. For the first 10 unemployment months we can notice a difference between the two curves, with a slight advantage for men over women, between 10 months and 25 months the curves invert, women hold a slight advantage to men; after 30 unemployment months the curves coincide.

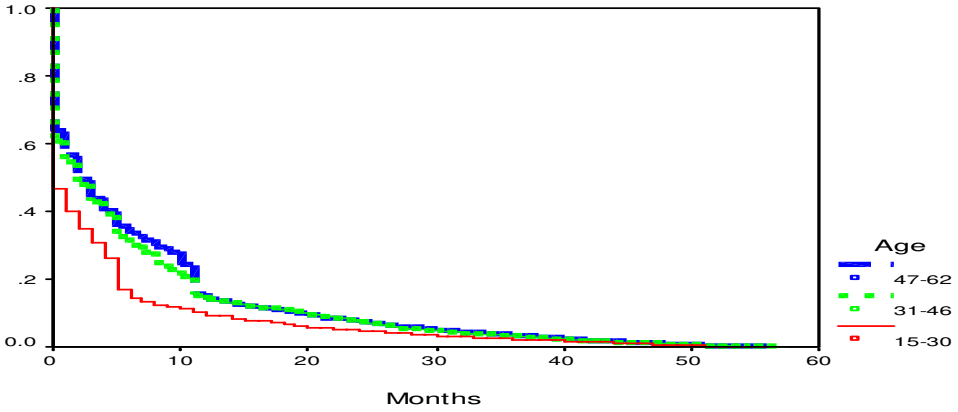
Figure 3: *Survival function estimates for male and female unemployed*



In figure 4 there is presented the survival curve for the age groups 15-30 years, 31-46 years and 47-62 years. We can notice that within 30 unemployment months the young age group 15-30 years holds a net advantage as compared to the other two age groups, after 3 unemployment years the curves start to coincide with each other.

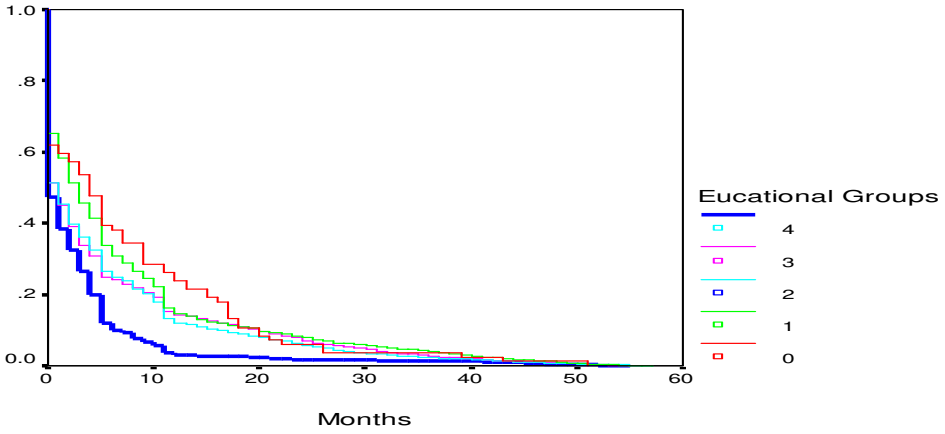
The age group 31-46 years has an advantage over the age group 47-62 years within 12 months; afterwards differences between these two curves are minor, starting to coincide.

Figure 4: *Survival function estimates for the age groups 15-30, 31-46 and 47-62 years*



In figure 5 there is presented the survival curve for the four groups concerning the educational level described in this research. We can notice in the figure that for the group 4, university education level (faculty or college) the probability of unemployment at time t or later decreases much more rapidly, indicating that the unemployed with the higher education have better opportunities in the labour market, as we expected. Group 0, without education, followed by group 1, unfinished secondary school, secondary or vocational school, registers the lowest employment probability. We can notice in the figure that after 40 unemployment months curves start to coincide and the educational level no longer influences the probability of finding a job.

Figure 5: *Survival function estimates for the four groups of education*



Testing the statistical signification for Kaplan Meier method presupposes the choice of one of the two hypotheses*: the null hypothesis, which supposes that curves should be the same for two or several levels of a specified factor*, or the alternative hypothesis, which supposes that they should be different. With this purpose we used the so-called log rank test

with Chi-Squared distribution under the null. For the age and education factors, the highly significant *p*-values (lower that 10^{-6}.) confirm the results derived graphically from the Kaplan-Meier estimates of the survival functions. But for the sex variable the value of *p* is not significant, the null hypothesis being accepted.

3. CONCLUSION

In respect of the duration of unemployment, persons with university education level remain unemployed for 5 months on the average, unlike persons without education, who remain unemployed for 13 months on the average, and persons with maximum 10 years of study, who remain unemployed for 9 months on the average.

After the performed analysis, the conclusion which emerges is that the persons with university education level are advantaged on the labour market in Gorj County; of 5185 persons registered in the database as unemployed with university education level, 1724 of them, representing 33.25% leave unemployment by becoming employed, of 5838 unemployed with post high school + foremen school, 28.47% become employed, of 28555 unemployed who are high school graduated, 25.38% become employed, of 40863 unemployed with maximum 10 years of study, 21.16% become employed, and of 520 unemployed without education that are registered in the database, only 16.15% leave unemployment by becoming employed.

As for age, for the group 15-24 years 20.75% leave unemployment by becoming employed, 30.76% of the young people aged between 25-34 years registered in the database leave unemployment by becoming employed, 32.07% of the persons aged between 35-44 years become employed during the analysed period, 28.82% is the percent corresponding to the age group 45-54 years respectively 23.07% for the age group over 55 years.

The age group 15-24 years is disadvantaged on the labour market by the lack of experience, a considerable number of them become unemployed after graduation. But the duration of unemployment is on the average the smallest for the age group 15-24 years, 6 months, compared to 9 months for the group 25-34 or 13 months for the group over 55 years.

Regarding the variable gender, of 33270 women registered in our database 19.21%, leave unemployment by becoming employed and of 47691 men registered 27.21% leave unemployment by becoming employed. But the duration of unemployment is smaller for women with about a month on the average. The differences between men and women, although existent, don't allow us to point out a clear advantage for one of these two categories.

4. BIBLIOGRAFY

Cadariu Andrei Achimas (2004). *Methodology of research in medical science* http://www.info.umfcluj.ro/resurse/Laborator/Metodologie/LabStoma/Materiale/CursMetodologie.pdf.

Chan Y.H (2004). *Biostatistics 203. Survival Analysis.* Singapore Med J 2004 Vol. 45(6): 249.

Carroll Nick (2005). *Explaining Unemployment Duration in Australia.* http://econpapers.repec.org/article/blaecorec/v_3A82_3Ay_3A2006_3Ai_3A258_3Ap_3A298-314.htm

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

Ministry of Labour, Social Solidarity and Family (2006). *Statistics.* http://www.mmssf.ro/website/ro/statistici/somaj53.pdf