A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

Iwona Markowicz[*], *Beata Stolorz*[**]

THE PROBLEM OF THE CENSORSHIP OF DATA IN THE RETROSPECTIVE RESEARCH OF UNEMPLOYMENT

Abstract. In the paper results of study on the influence of inadequate information to the estimation of survival curve using the Kaplan-Meier method will be presented. Authors carried out the time of being unemployed analysis for people unregistered from the District Labour Office in Szczecin in I quarter 2007. Methods of survival analysis allowing censored data applying were used.
Key words: Kaplan-Meier estimator , log-rank test, Gehan test, unemployment

I. INTRODUCTION

Methods of survival analysis are increasingly used in studies of socioeconomic phenomena (see Bednarski (2005); Hozer (2002); Markowicz, Stolorz (2007b). On account of lack of the need for a knowledge of the distribution of the studied random variable, particular importance is attached to non-parametric or semi-parametric models. The authors performed a review of analysis methodology of events and their application in the study of companies functioning period as part of the implementation of a MNiSW grant (N 111 011 31/1109). The condition of survival model analysis application is an appropriate database enabling the determination of the length of a defined state for separate units of a community under study. These are usually retrospective studies using the compiled records. An example of this type of database is the register of the unemployed. The aim of the study was to indicate what effect the data classification can have on the interpretation of results.

II. STATISTICS. CLASSIFICATION OF UNITS FOR CENSORSHIP

The conducted analysis was based on the statistics obtained from the District Labour Office in Szczecin and concerns the unemployed unregistered in I quarter 2007. We assumed complete observations to be those unregistered due to taking up a job while other reasons resulted in the classification of data into

[*] Ph.D., Department of Econometrics and Statistics, Faculty of Economics and Management University of Szczecin.
[**] Ph.D., Department of Econometrics and Statistics, Faculty of Economics and Management University of Szczecin.

the censored observation group. However, some reasons for unregistrating from the Labour Office may be connected with taking up a job. The questionable reasons include: going abroad, non-appearance in the Office on time. This is why the authors decided to assume two different approaches to the censorship of data (Table 1). The first method considered only the unregistration due to taking on a job (including self-employment) to be complete observations. In the other method of censorship, the list of reasons for taking up a job was extended. This may have a significant influence on the results of studies on the length of unemployment.

Table 1. Data classification method in type I and II censorship

Censorship I	Reasons for unregistration	Censorship II
Complete observations	Taking up a job	Complete observations
Censored observations	– Going abroad, – Non-appearance in the Office on time, – The application of the interested party for deleting their name from the register	
	Other reasons	Censored observations

Source: authors' study.

Table 2 presents the structure of the unemployed unregistered from the District Labour Office in Szczecin in I quarter 2007 according to gender, education and age. Using the described methods of censorship, a total of 4237 individuals were studied. These did not include the unemployed sent by the Office to placements with companies, training courses or vocational preparation. In these cases there is no formal employment and therefore the individuals are not included in the analyzed data. Particular variants of gender and education and age divisions were numbered in Table 2, which is going to be used in subsequent presentation of the analysis results.

Table 2. The number of the unemployed according to gender, education and age group considering two censorship methods

Feature		Censorship I		Censorship II	
		Observations			
		complete	censored	complete	censored
1	2	3	4	5	6
Gender	women	907	1300	2027	180
	men	1039	991	1851	179
Education	none or incomplete primary, primary, junior high, (1)	502	938	1298	142
	basic vocational, (2)	360	500	767	93
	secondary, (3)	208	198	374	32

Table 1 (cont.)

1	2	3	4	5	6
	4-year vocational secondary, vocational secondary, post-secondary/vacational college (4)	409	360	706	63
	higher (including Bachelor's degree) (5)	467	295	733	29
Age	<18-25) (1)	431	569	955	45
	<25-35) (2)	779	719	1426	72
	<35-45) (3)	310	369	619	60
	<45-55) (4)	361	474	718	117
	<55-60) (5)	61	112	131	42
	<60-65) (6)	4	48	29	23
In total		1946	2291	3878	359

Source: authors' own study.

III. PROBABILITY OF NOT FINDING A JOB

The probability of staying unemployed in the following months after registration was estimated using the Kaplan-Meier estimator (Product-Limit Estimation) (Kaplan, Meier (1958), pp. 457–481), which takes into consideration the existence of censored data (see Domański, Pruska (2000), pp. 203–204; Frątczak, Gach-Ciepiela, Babiker (2005), pp. 30–36, Bednarski (2005), pp. 385–396). It is a nonparametric method, not requiring the need for compartment construction for the time variable, but only the classification of episodes according to their length (Frątczak, Gach-Ciepiela, Babiker (2005), pp. 65-69). Every time point, at which at least one event occurred, was assigned with the risk value. In survival analysis, it is important to include units which belong to the population at a given time unit. These units are not included in subsequent calculations after the censorship. In the case of the study of unemployment, the chances of finding a job are influenced by the total number of the unemployed at a given time period, regardless of whether they will be unregistered in the future due to taking up a job or for another reason. Fig. 1 presents the course of survival curves using two censorship methods. Treating some reasons for the unregistration of the unemployed as taking up a job causes a change in the course of the survival curve and indicates a faster way out of unemployment.

Fig. 2 presents a comparison of two survival curves for the unemployed according to gender using two data censorship methods. The analysis of the results using the extended censorship method (II) indicates a faster way out of unemployment both for women and men. There was a change in the curves' position for particular genders towards each other. The II type of censorship was more beneficial for men.

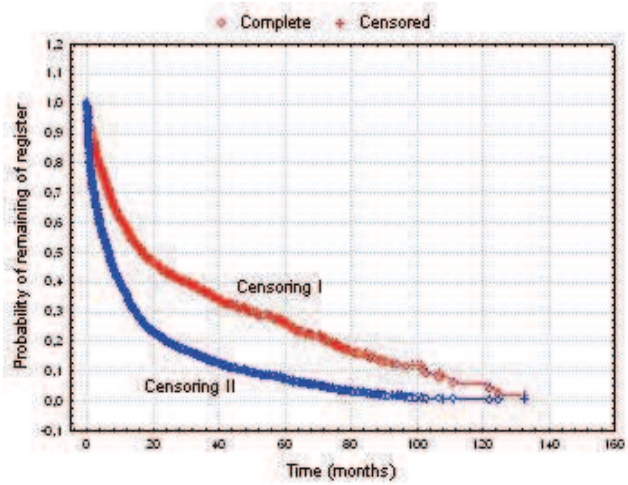


Fig.1. Estimators of Kaplan-Meier survival curves considering two methods of censorship - total number of unemployed people

Source: authors' own study.

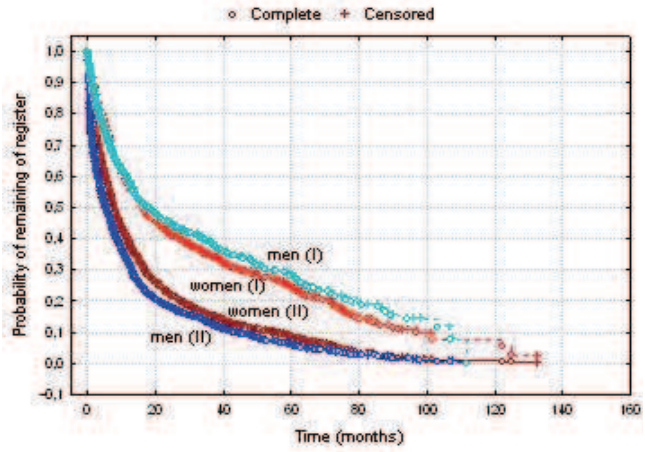


Fig.2. Estimators of Kaplan-Meier survival curves considering two methods of censorship - unemployed with respect to sex

Source: authors' own study.

Fig. 3 presents the Kaplan-Meier estimators for both censorship methods according to education for men and women. The use of II censorship method indicates a lower probability of unemployment after a given time period. There was also a decrease in the length of finding a job by the unemployed in different education categories.

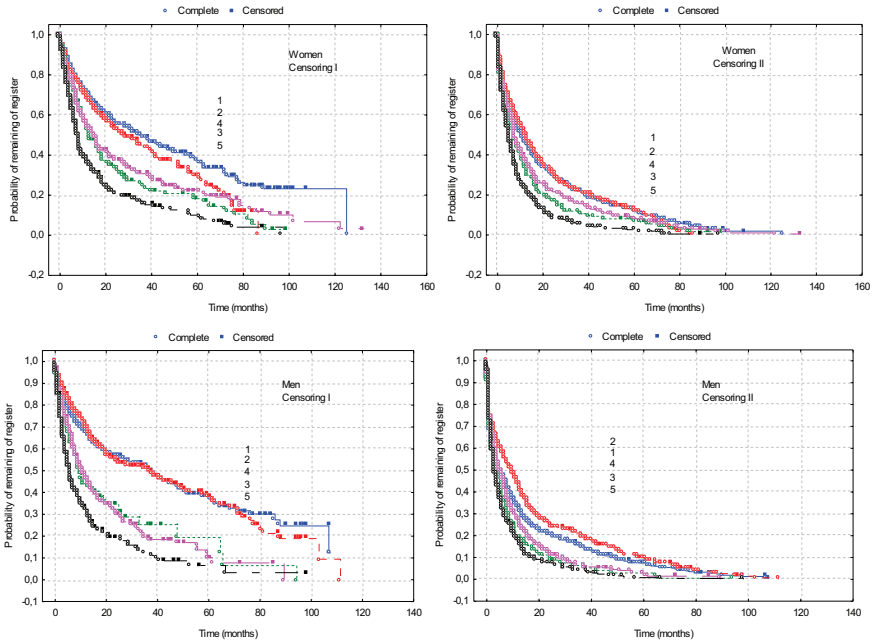
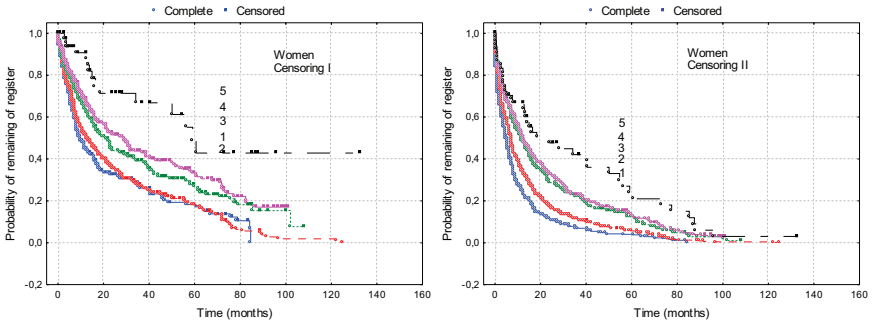


Fig.3. Estimators of Kaplan-Meier survival curves considering two methods of censorship - unemployed with respect to education

Source: authors' own study.

Another characteristic of the unemployed under study is age[1]. The probability of not finding a job by the unemployed in particular age groups, according to gender, using both censorship methods, is presented in Fig. 4.



[1] See Markowicz, Stolorz (2007a).

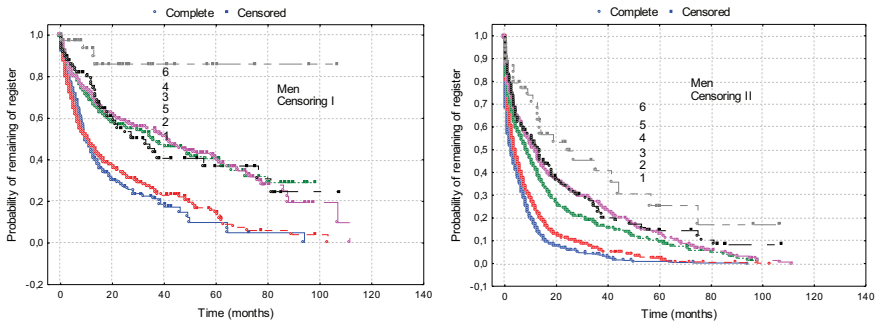


Fig.4. Estimators of Kaplan-Meier survival curves considering two methods of censorship - unemployed with respect to age

Source: authors' own study.

In all of the studied cased, the second method of censorship indicates a shorter period of finding employment. A decrease in differences between particular age and education categories is also apparent. Another change was in the order of survival curves, where in both cases the first to find employment were the people with higher education, whereas the last were the elderly people.

IV. THE STUDY OF THE SIGNIFICANCE OF THE SURVIVAL CURVE COURSE

Communities under study can be divided into groups. Then there is a possibility of estimating the survival function for each of the group and studying the significance of differences between them. The survival time can be compared in two or more tests. As their distribution is unknown, the nonparametric test has to be used. These types of tests are based on the range order of survival time. In the case of the following tests can be applied: the Gehan generalization of Wilcoxon test, the Cox-Mantel test, the F Cox test, the long-rank test as well as the Peto generalization and Wilcoxon test Peto[2]. Unfortunately, there are no commonly accepted methods of the test choice in a given situation. This is because it depends on the number of tests, on the existence of the censored data as well as on the knowledge of the data distribution[3]. In the majority of the tests, the calculated statistics for a large essay asymptotically aim at regular distribution. The majority of the tests give reliable results only in the case of large essays, whereas the efficacy of the tests in the case of smaller essays is less well known. Due to the differences in the assumed weights, there are differences between the test values. It is assumed,

[2] See Namboodiri, Suchindran (1987), pp. 71-91; Cox, Oakes (1984), pp. 123-125.
[3] Lawless (1982), pp. 425-427.

for example, that the Gehan (Wilcoxon) test gives good results in the study of the differences in the initial course of the survival curves. The differences in the final course of the survival curves are better shown in the log-rank test. The analysis in the paper was conducted on the basis of two tests (Gehan and log-rank), which resulted in two identical verification decisions. Table 3 present significant differences in the course of Kaplan-Meier curves for two coding methods in the appropriate female and male education and age groups.

Table 3. Significant differences in coding methods – women and men

Education	Age - women						Age - men						
	total	18-25	25-35	35-45	45-55	55-60	total	18-25	25-35	35-45	45-55	55-60	60-65
Total	+	+	+	+	+	+	+	+	+	+	+	+	+
None or incomplete elementary, elementary, junior high	+	+	+	+	+	+	+	+	+	+	+	+	+
Vocational elementary	+	+	+	+			+	+	+	+	+		
Secondary	+	+	+				+	+	+				
4-year secondary, vocational secondary, post-secondary	+	+	+		+		+	+	+	+	+		
Higher (including Bachelor's degree)	+	+	+				+		+				

Source: authors' own study.

Table 4 present significant differences in the course of the Kaplan-Meier curves for gender variants according to education, age and censorship methods. Situations, in which the differences in the course of curves for men and women were statistically significant, are marked with *M*. The period of seeking employment in both cases was shorter for men.

Table 4. Significant differences in the course of Kaplan-Meier curves for gender variants according to education and age and censorship methods

Education	Censorship method		Age	Censorship method	
	I	II		I	II
Total		M	total		M
None or incomplete elementary, elementary, junior high		M	18-25		M
Vocational elementary		M	25-35	M	M
Secondary		M	35-45		M
4-year secondary, vocational secondary, post-secondary		M	45-55		
Higher (including Bachelor's degree	M	M	55-60	M	

Source: authors' study.

The application of Gehan test (also log-rank) concerning the significance of differences in the probability function graph of not finding a job – conclusions: significant differences in the rate of leaving unemployment using I and II censorship method (I-women, II-men):

– the censorship method has no effect on the probability function graph of unregistration from District Labour Office for higher age and education groups,

– the classification method of reasons for unregistration has an effect on the achieved results in the case of the unemployed: young (up to 35 years), poorly educated (up to junior high), older and poorly educated, young and well-educated,

– gender: there are no significant differences in the probability function graph of not finding a job by women and men using I censorship method and there is a significance of differences using II censorship method,

– the differences in the rate of leaving unemployment between women and men are significant only for the higher education group - using I censorship method and in all education variants - using II censorship method,

– the differences in the rate of leaving unemployment between women and men are significant in the 25-35 and 55-60 age groups – using I censorship method, and in the 18-25, 25-35 and 35-45 age groups – using II censorship method.

V. CONCLUSIONS

The methods of history of events analysis, commonly used in demography are increasingly applied in other fields of socioeconomic science. The tools used here require an appropriate method of collecting statistical data. The data need to be individual and must include the length of a person's remaining in a particular state. The conducted study shows that the reason for unregistration of the unemployed from the District Labour Office may be ambiguous and therefore the conducted analysis may produce different results.

REFERENCES

Bednarski T. (2005), Ocena przydatności BAEL dla charakterystyki rozkładu czasu poszukiwania pracy na przykładzie danych z lat 2001-2002, *Studia Ekonomiczne*, nr 4, Instytut Nauk Ekonomicznych PAN, Warszawa, s. 385-396.

Cox D. R., Oakes D. (1984), *Analysis of Survival Data*, Chapman and Hall, London.

Domański C., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

Frątczak E., Gach-Ciepiela U., Babiker H. (2005), *Analiza historii zdarzeń. Elementy teorii, wybrane przykłady zastosowań*, SGH, Warszawa

Hozer J. (red.), (2002), *Badania statystyczne w ubezpieczeniach,* Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.
Kaplan E. L., Meier P. (1958), Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* 53, s. 457-481.
Lawless J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
Markowicz I., Stolorz B. (2007a), Identyfikacja determinant czasu oczekiwania na pracę bezrobotnych w Szczecinie, *Wiadomości Statystyczne*, nr 12, Warszawa, s.57-65.
Markowicz I., Stolorz B. (2007b), *Determinants of Labour Seeking Time Resulting From Labour Demand on Szczecin Labour Market*, [w:] The labour demand in the modern economy, Economics & Competition Policy, No.10, Szczecin.
Namboodiri K., Suchindran C. M. (1987), *Life Table Techniques and Their Applications*, Academic Press Inc., New York.

Iwona Markowicz, *Beata Stolorz*

PROBLEM CENZUROWANIA DANYCH W BADANIU RETROSPEKTYWNYM BEZROBOCIA

W artykule przedstawione zostaną wyniki badania wpływu niepełnej informacji na estymację krzywej przeżycia metodą Kaplana-Meiera. Autorki przeprowadziły analizę czasu pozostawania bez pracy przez osoby bezrobotne, które zostały wyrejestrowane z Powiatowego Urzędu Pracy w Szczecinie w I kwartale 2007 roku. Zastosowano metody analizy przeżycia, które dopuszczają występowanie danych cenzurowanych.