EKONOMETRIA ECONOMETRICS 1(47) • 2015

ISSN 1507-3866
e-ISSN 2449-9994

Anna Sączewska-Piotrowska
University of Economics in Katowice
e-mail: anna.saczewska-piotrowska@ue.katowice.pl

POVERTY DURATION OF HOUSEHOLDS OF THE SELF-EMPLOYED

CZAS TRWANIA UBÓSTWA GOSPODARSTW DOMOWYCH PRACUJĄCYCH NA WŁASNY RACHUNEK

DOI: 10.15611/ekt.2015.1.03

Summary: This study is one of the first attempts to discover how long households in Poland remain in poverty (out of poverty) and whether the time spent in poverty (out of poverty) depends on the socio-economic group of household. The analysis is conducted using panel data collected in the project "Social Diagnosis" in 2000-2013. We analyze the survivor functions of staying in poverty (out of poverty) using the Kaplan-Meier method. The probability of survival for a long time in poverty is less than in the case of survival out of poverty. It should be noted that a small percentage of households remained in poverty for almost the entire period of the study. We compare the survival functions of staying in poverty (out of poverty) according to the socio-economic groups of households. For this purpose we use the log-rank test. In both cases the survivor functions are significantly different. Households of the self-employed survive longer out of poverty and simultaneously survive shorter in poverty than the other groups of households.

Keywords: Kaplan-Meier estimator, log-rank test, poverty duration.

Streszczenie: W artykule przedstawione są wyniki jednego z pierwszych w Polsce badań dotyczących długości przebywania w sferze ubóstwa (w sferze poza ubóstwem) z zastosowaniem metod analizy przeżycia. Analiza obejmuje gospodarstwa domowe ogółem oraz podzielone ze względu na grupę społeczno-ekonomiczną. Badanie jest przeprowadzane z wykorzystaniem bazy danych projektu „Diagnoza społeczna" z lat 2000-2013. W pierwszej kolejności analizowane są funkcje przeżycia Kaplana-Meiera, na podstawie których można stwierdzić, że prawdopodobieństwo przeżycia w ubóstwie w długim okresie jest mniejsze od prawdopodobieństwa przeżycia poza ubóstwem. Należy zaznaczyć, że tylko niewielki odsetek gospodarstw domowych pozostaje w sferze ubóstwa przez cały badany okres. W kolejnym kroku funkcje przeżycia w sferze ubóstwa (poza sferą ubóstwa) są porównywane w gospodarstwach domowych podzielonych ze względu na grupę społeczno-ekonomiczną. W tym celu używany jest test log-rank. Zarówno w przypadku przebywania w sferze ubóstwa, jak i przeżycia poza sferą ubóstwa, funkcje przeżycia różnią się w statystycznie istotny sposób. W porównaniu z innymi grupami gospodarstw domowych gospodarstwa pracujących na własny rachunek przeżywają dłużej w sferze poza ubóstwem i jednocześnie przeżywają krócej w sferze ubóstwa.

Słowa kluczowe: estymator Kaplana-Meiera, test log-rank, czas trwania ubóstwa.

1. Introduction

Poverty studies are mostly conducted in a cross-sectional manner. Adding a time variable to poverty studies allow to create a whole new perspective regarding the duration and dynamics of poverty. Survival analysis methods are one of the methods used in poverty dynamics and duration research. These methods are most common in demography, actuarial statistics and medicine, however they are used more frequently in other sciences. Bane and Ellwood [1986] were precursors of using survival analysis methods in poverty studies. Natalia Nehrebecka [2010] has analyzed time spent in poverty in Poland based on panel data from the CHER database (Consortium of Household Panels for European Socio-Economic Research) for 1997-2000. The study conducted by Nehrebecka is for the time being the only attempt to apply survival analysis methods in poverty studies in Poland. Other poverty studies that have taken into consideration the time variable, regarded only pointing out the groups of households in persistent poverty or focused on changes in the poverty status in two periods of time, thereby applying the transition matrices. Studies were conducted by Okrasa [2000], Topińska [2005], Sączewska-Piotrowska [2012] and Panek in the "*Social Diagnosis*" project [Czapiński, Panek 2003; 2005; 2007; 2009; 2011; 2013]. Since the last series of panels used in Nehrebecka's research fourteen years have passed, while the poverty phenomenon, to be fully explored, needs to be permanently studied and monitored.

It is worth mentioning that Poland is a country where during this period of years, relevant changes have occurred related to the socio-economic transformation and Poland accessing to the European Union. This makes it reasonable to conduct a study on poverty dynamics and its duration in Poland using event history analysis methods in a changing socio-economic reality.

The purpose of this article is to determine how long households in Poland survive in poverty (out of poverty) and whether the time spent in poverty (out of poverty) depends on the socio-economic group of the household. It is particularly important to answer the question of whether households of the self-employed survive longer in poverty and simultaneously survive shorter out of poverty than other groups of households. It is generally acknowledged that this group of households is in a worse material situation and this is why many people have a pessimistic approach to self-employment. This study will allow, to a certain extent, to answer the question whether these concerns are justified.

2. Data

This study of the duration of poverty in Poland is based on seven series of panels realized in 2000-2013 in the framework of the project "Social Diagnosis" [Council for Social Monitoring 2014]. The subsequent stages of the study involved all the households from the previous series and included a new representative sample. The

analysis of the duration of poverty refers to households participating in every stage of the panel.

Poverty analysis adopts the economic definition of poverty. As an indicator of a households' wealth we assume the net income of households in Poland in February/March 2000, 2003, 2005, 2007, 2009, 2011 and 2013. In order to take account of the differences in a household's size and composition, an equivalent income is calculated by dividing the household's income by its equivalent size, which is calculated using the modified OECD equivalence scale. This scale assigns 1 to the first adult of the household, 0.5 to each subsequent adult aged 14 or more and 0.3 to children (each person under 14). The household weighted equivalised income is adopted. The poverty threshold is set at 60% of the median equivalised income.

3. Methods

Survival analysis (also known as duration analysis) is the all-encompassing term for the statistical methods that examine time-to-event data. The dependent variable in survival analysis is the duration until event occurrence. Event time (also called survival time, episode, spell, duration or failure) is a non-negative random variable, which is denoted as T. The specific value of T is denoted as t. The primary quantity of interest in survival analysis is the survivor function, which can be expressed in terms of distribution function [Hosmer, Lameshow, May 2008, p. 16; Mills 2011, p. 9]. The distribution function of the random variable T is the probability that survival time T is less than or equal to some value t. This is denoted as

$$F(t) = P(T \leq t).$$

The survivor function, also known as a survival function, is specified as

$$S(t) = 1 - F(t) = P(T > t)$$

which expresses the probability that survival time T is greater than some time t. $S(t)$ denotes the proportion of subjects surviving beyond t. The survivor function has the following theoretical properties [Kleinbaum, Klein 2005, p. 9]:

- $S(t)$ is a non-increasing function,
- $S(0) = 1, \lim_{t\to\infty} S(t) = 0,$
- the graph of $S(t)$ is a smooth curve.

In practice, we observe events on a discrete time scale (days, weeks etc.) and therefore the graph of $S(t)$ is a step function. The graph often does not go all the way down to zero at the end of the study, because not every individual studied relates to the event.

Many events in economic or social research may occur more than once to an individual over the observation period. Households participating in the panel in the "Social Diagnosis" project could enter into poverty (or exit from poverty) several times, which means the events may be repeated. In practice, the survival time is the

waiting time for the occurrence of an event (usually for the first poverty entry or poverty exit) or the time between subsequent events (for example the time between the third and fourth poverty entry). Some authors point out that in the case of small mean spells per unit (less than two), it is recommended to limit analysis only to the first spell [Allison 2010]. Therefore in our analysis we take into consideration only the first spells (the waiting time for the first poverty exit and for the first poverty entry).

In the analysis of poverty duration there are many situations in which the story of the episode is not complete – there is the problem of left and right censored data. This means that certain episodes start and end outside of the study period. Our analysis takes into account only the spells that start within the observation period, which means that left-censored data are not included. From the seven series, the first two are used to construct the "inflow" condition. Consequently, up to five series are used for observing poverty exits and entries. For poverty exits, we demand that the household is not-poor in the first period, poor in the second period and from the third period we study whether the household exits poverty. The same situation occurs in the case of poverty entries (poor, not-poor and from the third period we observe poverty entries).

In the survival analysis nonparametric, semiparametric and parametric methods are used. Duration analysis often begins with nonparametric methods and these methods are used in our study. In nonparametric methods no assumption of the shape of the survivor function need be made. The most popular nonparametric estimator of the survivor function for uncensored and right-censored data is the Kaplan-Meier (product-limit) estimator.

Let T denote survival time with distribution function (d.f.) F and probability density function (p.d.f.) f and C denote censoring time with d.f. G and p.d.f. g. We observe n individuals. Each individual has a survival time T_i and a censor time C_i. On each of n individuals we observe the pair (Y_i, δ_i) where [Tableman, Kim 2005, p. 13]

$$Y_i = \min(T_i, C_i),$$

$$\delta_i = \begin{cases} 1 \text{ if } \text{uncensored, i. e. } T_i \leq C_i \\ 0 \text{ if } \text{censored, i. e. } T_i > C_i \end{cases}.$$

We observe n independent and identically distributed (i.i.d.) random pairs (Y_i, δ_i). Assume that the censoring time C_i is independent of the survival time T_i.

For a sample of size n, let $t_1 < t_2 < \cdots < t_m$ denote the ordered event times and let $t_0 = 0$. Let

- d_i = number of events occurring at time t_i,
- n_i = number of individuals at risk of event immediately before time t_i (including censored survival times at time t_i),
- $p_i = P(\text{survival at least } t_i | \text{survival up to } t_{i-1}) = P(T > t_i | T > t_{i-1})$,
- $q_i = 1 - p_i = P(\text{event in } t_i | \text{survival up to } t_{i-1}) = P(T = t_i | T > t_{i-1})$.

Remember the general multiplication rule for joint events A and B:

$$P(A \cap B) = P(B|A)P(A).$$

From the repeated application of this product rule for $t_k \leq t < t_{k+1}$ the survivor function can be expressed as

$$\begin{aligned} \mathrm{S(t)} &= P(T > t) \\ &= P(T > t|T > t_k) \times P(T > t_k|T > t_{k-1}) \times \ldots \times P(T > t_1|T > t_0) \\ &\times P(T > t_0) = \prod_{i=1}^{k} P(T > t_i|T > t_{i-1}) = \prod_{i:t_i \leq t} p_i \,, \end{aligned}$$

where $P(T > t_0) = 1$.

The estimates of p_i and q_i are

$$\hat{q}_i = \frac{d_i}{n_i},$$

$$\hat{p}_i = 1 - \hat{q}_i = 1 - \frac{d_i}{n_i}.$$

The Kaplan-Meier [1958] estimator of survivor function is

$$\hat{S}(\mathrm{t}) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{i=1}^{k} \left(1 - \frac{d_i}{n_i}\right).$$

Several estimators are used to approximate variance of $\hat{S}(\mathrm{t})$. One of the most common estimators is Greenwood's formula [1926]:

$$\widehat{\mathrm{var}}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} = \left(\hat{S}(t)\right)^2 \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}.$$

A common procedure using the Kaplan-Meier estimates is to compare two or more survivor curves. The most popular method testing equality of survivor functions is the log-rank (Mantel-Haenszel) test [Hosmer, Lemeshow, May 2008, pp. 51--53; Kleinbaum, Klein 2005, p. 82; Rodríguez 2005, pp. 6-8].

Let $t_1 < t_2 < \cdots < t_m$ denote the ordered event times in the total sample, obtained by combining all groups of interest. Let

- d_{ij} = number of events occurring at time t_i in group j,
- n_{ij} = number of individuals at risk of event at time t_i in group j.

We also let

$$d_i = \sum_{j=1}^{G} d_{ij},$$

$$n_i = \sum_{j=1}^{G} n_{ij}.$$

We estimate the expected number of events for each group under the assumption of equal survival functions as

$$\hat{e}_{ij} = \frac{d_i n_{ij}}{n_i}, \ \ j = 1,2, \dots, G.$$

We compare the observed and expected number of events for *G*-1 of the *G* groups. The observed and expected number of events in vector notation are

$$\mathbf{d}_i^T = (d_{i1}, d_{i2}, \dots, d_{iG-1}),$$
$$\hat{\mathbf{e}}_i^{\mathrm{T}} = (\hat{e}_{i1}, \hat{e}_{i2}, \dots, \hat{e}_{iG-1}).$$

The difference between these two vectors is

$$(\mathbf{d}_i - \hat{\mathbf{e}}_i)^T = (d_{i1} - \hat{e}_{i1}, d_{i2} - \hat{e}_{i2}, \dots, d_{iG-1} - \hat{e}_{iG-1}).$$

We have used the first *G*-1 of the *G* groups, but any collection of *G*-1 groups could be used.

To obtain a test statistic, we need an estimator of the variance-covariance matrix of $\mathbf{d}_i$. The elements of this matrix are obtained assuming that the observed number of events follows a multivariate central hypergeometric distribution. The diagonal elements of the variance-covariance matrix $\widehat{\mathbf{V}}_i$ are

$$\hat{v}_{jji} = \frac{n_{ij}(n_i - n_{ij})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \ \ j = 1,2, \dots, G-1$$

and the off diagonal elements are

$$\hat{v}_{jki} = -\frac{n_{ij}n_{ik}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \ \ j,k = 1,2, \dots, G-1, j \neq k.$$

Mantel and Haenszel proposed testing the equality of the *G* survival functions

$$H_0: S_1(t) = S_2(t) = \cdots = S_G(t)$$

by treating the quadratic form

$$Q = \left[\sum_{i=1}^{m}(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]^T \left[\sum_{i=1}^{m}\widehat{\mathbf{V}}_i\right]^{-1} \left[\sum_{i=1}^{m}(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]$$

as a χ^2 statistic with *G*-1 degrees of freedom.

4. Results

All the calculations are made in R using the `survival` package. We use the `survfit` function to compute the estimates of the Kaplan-Meier survival curves and the `survdiff` function to test the difference between the survival curves. The graphs are made using the `plot` function which is a part of a base package.

Analysis begins with the graphic presentation (Fig. 1) and estimates the calculation (Tab. 1) of the Kaplan-Meier survival function of staying in poverty. In our analysis there are used two-year time units.

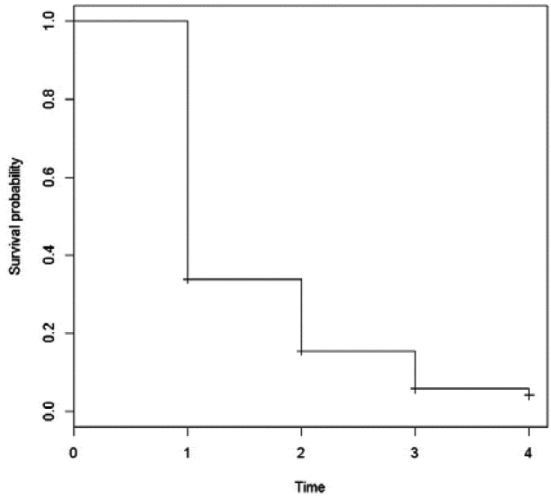

Fig. 1. Kaplan-Meier survival function of staying in poverty

Source: own study based on [Council for Social Monitoring 2014].

Table 1. Kaplan-Meier estimates of staying in poverty

Time in poverty (two-year time units)	Number of households at risk of poverty exit	Survival	Standard error
1	734	0.339	0.0175
2	118	0.155	0.0175
3	24	0.058	0.0167
4	7	0.042	0.0155

Source: own calculations based on [Council for Social Monitoring 2014].

The estimated survival function does not go to zero, we can therefore conclude that the largest observation is a right-censored value. The confirmation of this situation is symbol "+" at the end of curve. This symbol indicates right-censored cases.

At one period the Kaplan-Meier estimate is 0.339, which means that the estimated probability that a household will survive one period or more (i.e. two years or more) in poverty is 0.339. As can be seen the probability that a household will survive eight years or more is only 0.042. None of the households survived ten years or more in poverty.

In Figure 2 and Table 2 there are presented the results of the Kaplan-Meier estimates of staying out of poverty.

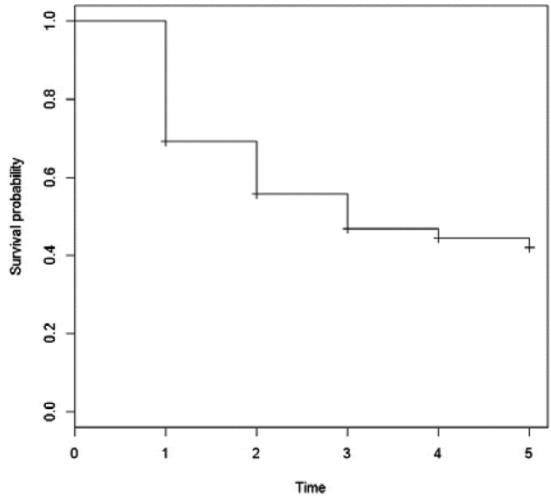

Fig. 2. Kaplan-Meier survival function of staying out of poverty

Source: see Fig. 1.

Table 2. Kaplan-Meier estimates of staying out of poverty

Time out of poverty (two-year time units)	Number of households at risk of poverty entry	Survival	Standard error
1	736	0.693	0.0170
2	241	0.558	0.0224
3	100	0.469	0.0278
4	40	0.445	0.0309
5	18	0.420	0.0378

Source: see Tab. 1.

In the case of survival out of poverty, the largest observation is a right-censored value. Almost 70% of households survive two years or more out of poverty and 42% of households survive ten years or more out of poverty.

The purpose of our analysis is to compare survivor curves between households divided into socio-economic groups. The survival functions of staying in poverty are plotted in Fig. 3.

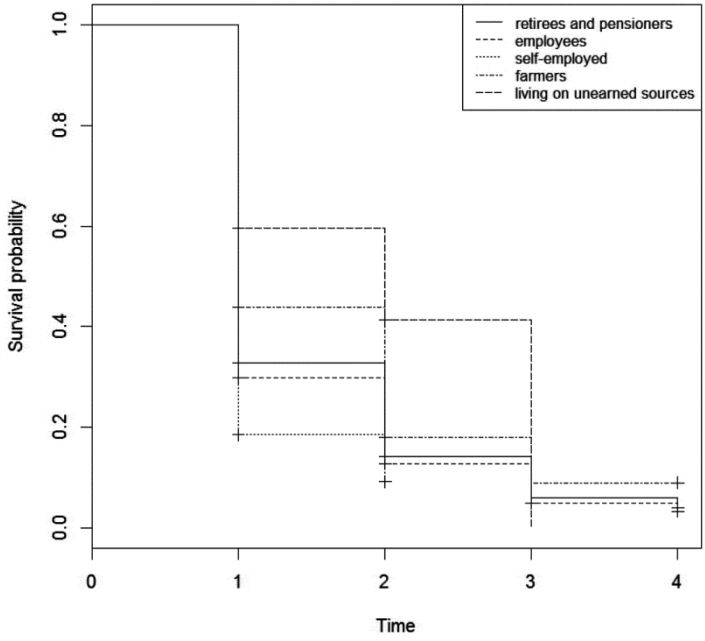

Fig. 3. Kaplan-Meier survival functions of staying in poverty according to socio-economic groups of the households

Source: see Fig. 1.

We can determine that the survival functions differ between the five groups. The line for the households living on unearned sources lies above the other lines and it can be concluded that this group survives the longest in poverty. Households of the self-employed survive the shortest in poverty, because the line of this group is the lowest situated one. We use the log-rank test to determine whether there is a statistically significant difference between the survival curves. The accurate results of the Kaplan-Meier estimates and the log-rank test are presented in Tab. 3.

The differences between the survival curves shown in Fig. 3 are also visible in various values of estimates in Tab. 3. It is evident that only in the case of households living on unearned sources the largest observation is uncensored data (the survival function goes to zero). In addition, we can see clearly that almost 60% of households living on unearned sources and simultaneously only 18.5% of households of the self-employed survive two years or more in poverty. The log-rank test confirms the differences between groups. We can see that the log-rank statistic is 23.5 and the corre-

sponding *p*-value is zero to three decimal places. We can conclude that there is a statistically significant difference between survival curves.

Table 3. Kaplan-Meier estimates and log-rank test results for the equality of the survival functions of staying in poverty according to socio-economic groups of the households

Time (two-year time units)	Households of				
	employees	farmers	the self-employed	retirees and pensioners	living on unearned sources
1	0.299	0.438	0.185	0.328	0.596
2	0.128	0.181	0.093	0.142	0.413
3	0.049	0.090		0.061	0.000
4	0.033			0.041	
$\chi^2 = 23.5$ on 4 degrees of freedom p-value = 9.85e-05					

Source: see Tab. 1.

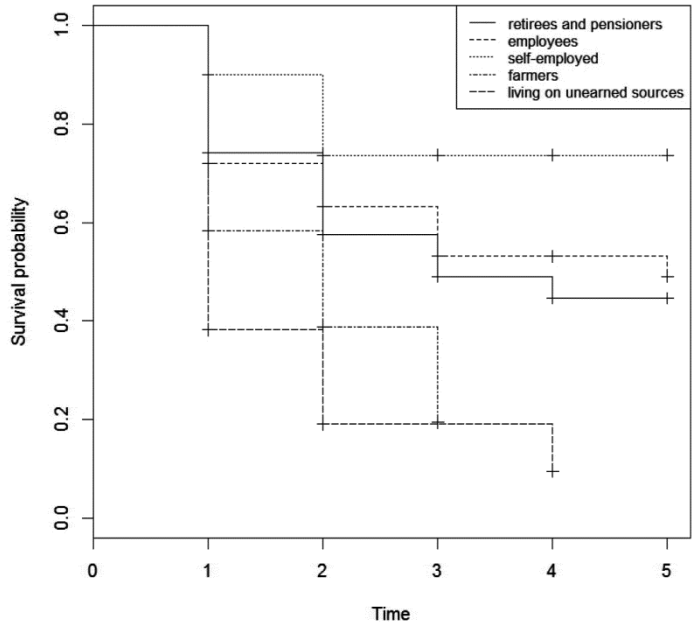

Fig. 4. Kaplan-Meier survival functions of staying out of poverty according to socio-economic groups of the households

Source: see Fig. 1.

The survival functions of staying out of poverty are plotted in Fig. 4. The results of the log-rank test are presented in Tab. 4.

Table 4. Kaplan-Meier estimates and log-rank test results for the equality of the survival functions of staying out of poverty according to socio-economic groups of the households

Time (two-year time units)	Households of				
	employees	farmers	the self-employed	retirees and pensioners	living on unearned sources
1	0.719	0.583	0.900	0.741	0.382
2	0.631	0.389	0.736	0.576	0.191
3	0.531	0.194	0.736	0.491	0.191
4	0.531		0.736	0.446	0.096
5	0.491				
χ^2=48.8 on 4 degrees of freedom p-value = 6.42e-10					

Source: see Tab. 1.

The largest observations in all survival functions are censored values. The figure shows that groups of households have different patterns of survival. On the one hand, for the households of self-employed, survival is relatively constant. On the other hand, for the households living on unearned sources, survival rapidly declines. It can be noted that 73.6% of households of the self-employed and only 9.6% of households living on unearned sources survive eight years or more out of poverty. We can conclude that the differences between the groups are large. The log-rank statistic is computed to be 48.8, which has a p-value of zero to three decimal places. The conclusion from the log-rank test is that there is a highly significant difference between the five survival curves of staying out of poverty.

5. Conclusions

The objective of this study was to discover how long households in Poland remain in poverty (out of poverty) and whether the time spent in poverty (out of poverty) depends on the socio-economic group of household. For this purpose we used nonparametric survival analysis methods.

Thanks to using the Kaplan-Meier method we determined that the probability of survival in poverty for a long time (eight years or more) is relatively small and at the same time the probability of survival for ten years or more out of poverty is ten times greater.

We compared the survival functions of staying in poverty (out of poverty) according to the socio-economic groups of households. For this purpose we used the log-rank test. In both cases the survivor functions were significantly different. Households of the self-employed survived longer out of poverty and simultaneously survived shorter in poverty than the other groups of households. In the worst situa-

tion were the households living on unearned sources – they survived the longest in poverty and the shortest out of poverty.

The obtained results are important from the point of view of the plan to combat poverty. The results suggest that the same group of households are at-risk of long-term survival in poverty and short-term survival out of poverty. Thus help from the government and non-governmental organizations should be directed mostly to this group of households. The aid granted at the right time will prevent long-term poverty and other phenomena related to poverty (for example, social exclusion).

Literature

Allison P.D., 2010, *Survival analysis,* [in:] G.R. Hancock, R.O. Mueller (eds), *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, Routledge, New York.

Bane M.J. Ellwood D.T., 1986, *Slipping into and out of poverty: The dynamics of spells*, The Journal of Human Resources, vol. 21, no. 1, pp. 1-23.

Council for Social Monitoring, 2013, *Social Diagnosis 2000-2013: Integrated Database*, http://www.diagnoza.com (accessed 23.04.2014).

Czapiński J., Panek T. (eds), 2003, 2005, 2007, 2009, 2011, 2013, *Social Diagnosis*, http://www.diagnoza.com (accessed 12.03.2014).

Greenwood M., 1926, *The Natural Duration of Cancer*, Reports on Public Health and Medical Subjects, vol. 33, pp. 1-26, Her Majesty's Stationery Office, London.

Hosmer D.W., Lemeshow S., May S., 2008, *Applied Survival Analysis. Regression Modeling of Time-to-Event Data*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Kaplan E.L., Meier P., 1958, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association, vol. 53, no. 282, pp. 457-481.

Kleinbaum D.G., Klein M., 2005, *Survival Analysis. A Self-Learning Text*, Springer, New York.

Mills M., 2011, *Introducing Survival and Event History Analysis*, SAGE Publications, Los Angeles–London–New Delhi–Singapore–Washington DC.

Nehrebecka N., 2010, *Analiza ubóstwa w Polsce w latach 1997-2000 z wykorzystaniem modeli hazardu*, Ekonomista, no. 1, pp. 95-116.

Okrasa W., 2000, *Who are Poland's long-term poor? Household risk-managing capabilities according to panel data 1993-1996*, Statistics in Transition, vol. 4, no. 5, pp. 841-882.

Rodríguez G., 2005, *Non-Parametric Estimation in Survival Models*, http://data.princeton.edu/pop509/NonParametricSurvival.pdf (accessed 6.06.2014).

Sączewska-Piotrowska A., 2012, *Badanie trwałości ubóstwa w Polsce*, [in:] A. Rączaszek (ed.), *Demograficzne uwarunkowania rozwoju społecznego*, Studia Ekonomiczne, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach no. 98, UE, Katowice, pp. 33-42.

Tableman M., Kim J. S., 2005, *Survival Analysis Using S. Analysis of Tim-to-Event Data*, Chapman & Hall/CRC, Boca Raton.

Topińska I., 2005, *Dynamika i trwałość ubóstwa w Polsce i na Węgrzech w latach dziewięćdziesiątych*,[in:] S. Golinowska, E. Tarkowska, I. Topińska (eds), *Ubóstwo i wykluczenie społeczne. Badania. Metody. Wyniki*, IPiSS, Warszawa, pp. 126-147.