sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 3
ISSN 1507-3866; e-ISSN 2449-9994

THE HAZARD FUNCTION AND ITS ROLE IN A NON-PARAMETRIC DURATION ANALYSIS OF ENTERPRISES IN THE ŁÓDZKIE VOIVODESHIP

Artur Mikulec
University of Lodz, Lodz, Poland
e-mail: artur.mikulec@uni.lodz.pl
ORCID: 0000-0001-8249-2296

DOI: 10.15611/eada.2019.3.06
JEL Classification: C10, C14, C41

Abstract: In a duration analysis of enterprises, as a rule there are determined four basic functions related to the time of their duration, i.e.: the density function; the distribution function; the survival function, and the hazard function. It turns out that the hazard function and its cumulative version are the key to understanding modern survival analysis. The aim of the paper is to indicate the best method of the estimation of the values of individual functions in survival analysis based on other functions. The paper provides compiled and classified information on particular functions used in the non-parametric duration analysis of enterprises. It examines some theoretical and practical problems related to the determination of, among others, the hazard function and the cumulative hazard function on the basis of data in cohort tables and the results of the estimation of the survival function with the use of the Kaplan-Meier method. The considerations included in the paper are illustrated with the results of analyses for enterprises established in the Łódzkie Voivodeship in 2001-2015 (including those which went into liquidation).

Keywords: enterprises, duration analysis, hazard function, cumulative hazard function, mean absolute percentage error, mean absolute percentage error.

1. Introduction

In a duration analysis of enterprises, as a rule there are determined four basic functions related to the time of their duration, i.e.: the density function – defining the probability of enterprise liquidation (unconditional); the distribution function – describing the probability of enterprise persistence; the survival function defining the probability of enterprise survival, and the hazard function – explaining the intensity of enterprise liquidation. In total there may be even seven functions, as the hazard function can take the ordinary or cumulative form, and the probability of

liquidation and survival can also take the form of conditional probability. Several methods of calculating the values of each of them are known, and moreover there exist numerous relations (dependencies) among the above mentioned functions.

The aim of the paper is to present theoretical-empirical considerations on the methods of determining the values of particular functions in survival analysis (under limited access to information) and the results showing which of them is the best (i.e. burdened with the smallest error). The paper provides compiled and classified information on the particular functions used in the non-parametric duration analysis of enterprises. It examines some theoretical and practical problems related to the determination of, among others, the hazard function and the cumulative hazard function on the basis of data in cohort tables and the results of the estimation of the survival function with the use of the Kaplan-Meier method. Moreover, it gives a short overview of the statistical programs and packages available in R environment – assisting with the non-parametric modeling of the hazard function. The article (to the best of the author's knowledge) can be seen as the first in Polish literature on this subject to provide a comparison of the various methods of estimating the values of the functions in survival analysis on the basis of other functions. The considerations included in the paper are illustrated with the results of analyses for enterprises established in the Łódzkie Voivodeship in 2001-2015 (including those which went into liquidation).

2. Function properties in the duration analysis of enterprises

The duration of an enterprise is a particular kind of non-negatively defined continuous random variable T, ($T \geq 0$), which defines the time which elapses from the moment of the enterprise's foundation until the moment of occurrence of an event – usually its liquidation. It is assumed that the examined entity can experience the analysed event only once. In order to characterize the distribution of time duration, one of the given below probabilistic functions with specific properties should be used [Balicki 2006, pp. 26-34, 53-67, 78-85; Frątczak et al. 2014, pp. 37--40; Jackowska 2013, pp. 16-27].

The density function of the probability of random variable is function f: $R \rightarrow R$ fulfilling two conditions $\forall_{t^* \in R} f(t^*) \geq 0$, $\int_0^\infty f\left(t^*\right) dt^* = 1$. It represents the probability of occurrence of an event of liquidation in the sense that $f(t^*)\Delta t^*$ can be understood as an approximation of probability that the event will occur in the time t^*. For any real t^*_1 and t^*_2, where $t^*_1 < t^*_2$ we have that $P\left(t_1^* < T \leq t_2^*\right) = \int_{t_1^*}^{t_2^*} f\left(t^*\right) dt^*$. Function $f(t^*)$ is an unconditional density function, thus the only condition to be satisfied is for an enterprise to exist in time $t^* = 0$. The density function is used in empirical research for the approximation of the empirical distribution of the number of events in specified intervals of time duration.

The distribution function of random variable is function F: $R \rightarrow [0, 1]$ defined by formula $F\left(t^*\right) = P\left(T \leq t^*\right)$, satisfying the following conditions: $F(0) = 0$, $F(\infty) = 1$, non-decreasing for $t^* > 0$, at least right-sided continuous. In practice the size of $F(t^*)$ defines the probability of an event that an enterprise will not survive time t^* – will experience the event of liquidation in period of time $(0, t^*]$.

The survival function is function S: $R \rightarrow [0, 1]$ given by formula $S(t^*) = 1 - F(t^*) = P(T > t^*)$, satisfying the following conditions: $S(0) = 1$, $S(\infty) = 0$, non-decreasing for $t^* > 0$, at least right-sided continuous. In practice the size of $S(t^*)$ defines the probability of an event that an enterprise survives time t^* – will function for a time longer than t^*.

The following relations occur among the density of the probability function, distribution function, and survival function: $F(t^*) = 1 - S(t^*)$, $f\left(t^*\right) = \frac{dF(t^*)}{dt^*}$, $F(t^*) = \int_0^{t^*} f(u)du$, $f\left(t^*\right) = \frac{-dS(t^*)}{dt^*}$, $S(t^*) = \int_{t^*}^{\infty} f(u)du$.

The distribution of time duration is also defined by the intensity function of liquidation / the risk function, also called **the hazard function** (ordinary), given by formula:

$$h\left(t^*\right) = \lim_{\Delta t^* \to 0^+} \frac{P\left(t^* \leq T \leq t^* + \Delta t^* | T \geq t^*\right)}{\Delta t^*}. \tag{1}$$

From the above formula it follows that intensity in moment t^* is approximately equal to the above conditional probability for one unit of time. Moreover for Δt^* small enough, the probability of an enterprise experiencing the event of liquidation in a short period of time $[t^*, t^* + \Delta t^*]$, on condition that until moment t^* it did not experience an event, is approximately proportional to the length of interval Δt^* and it is equal to $h(t^*)\Delta t^*$. Therefore intensity can be interpreted as an enterprise's susceptibility to experiencing an event of liquidation in time t^*, on condition that an enterprise persists until moment t^*. It can also be said that $h(t^*)$ measures the speed of decreasing of survival function of enterprises (dependent on the length of time duration t^*). In theory the hazard function can be an increasing, decreasing or constant function. As a rule, the examined entities maintain the same level of intensity throughout the whole period of time e.g. in demography it assumes a 'bathtub' shape. In an analysis of enterprise duration, it takes an inverted-U shape [Markowicz, Mikulec 2018]. The function increases quickly after an enterprise's foundation, reaches a certain maximum intensity of liquidation, and then decreases along with the enterprise's age. It has the following properties: $h(t^*) \geq 0$ for $t^* > 0$,

$$\int_0^{\infty} h\left(t^*\right)dt^* = \infty, \; h\left(t^*\right) = \frac{-d\ln S(t^*)}{dt^*}, \; h\left(t^*\right) = \frac{f(t^*)}{1 - F(t^*)}, \; h\left(t^*\right) = \frac{f(t^*)}{S(t^*)}.$$

In the literature on this subject we can also find the **cumulative hazard function** / the cumulative risk function / the cumulative intensity function of liquidation, also known as the logarithmic survival function in the form:

$$H\left(t^*\right) = \int_0^{t^*} h(u)du, \tag{2}$$

which satisfies the following conditions: $H(0) = 0$, $H(\infty) = \infty$, is non-decreasing for $t^* > 0$. It measures the total risk of liquidation from the moment of foundation until moment t^*. The relations between the cumulative hazard function and the survival function can be expressed in the following way: $H(t^*) = -\ln S(t^*)$, $S\left(t^*\right) = \exp\left[-\int_0^{t^*} h(u)du\right]$.

A compilation of all the relations (dependencies) between the functions in the form of tables can be found in the works by: Balicki [2006, p. 33]; Jackowska [2013, p. 25]; Landmesser [2013, p. 42], and the equivalents of the basic probabilistic functions for time treated as discrete random variable T were included in the publication by Bieszk-Stolorz [2013, p. 29].

The conditional probability of (experiencing en event of) liquidation $q(t^*)$ is described as the function whose values define conditional probability of experiencing an event of liquidation i.e. exiting the cohort of objects in unit interval of time duration $\left\langle t^*, t^*+1\right)$ on condition that until time t^* the enterprise does not experience an event of liquidation. It can be written as:

$$q(t^*) = 1 - \exp\left[-\int_{t^*}^{t^*+1} h(u)du\right]. \tag{3}$$

The relation does not mean that function $h(t^*)$ is a continuous equivalent of discrete function $q(t^*)$ but rather that $h(t^*)\Delta t^*$ can be understood as a limit version $q(t^*)$, when interval $\left\langle t^*, t^*+1\right)$ becomes very small (see: formula (1)).

The conditional probability of survival $p(t^*)$ is the function whose variables define the conditional probability that an enterprise does not experience an event of liquidation in unit interval of time duration $\left\langle t^*, t^*+1\right)$ on condition that, until time t^*, the examined entity does not experience the event of liquidation:

$$p\left(t^*\right) = 1 - q\left(t^*\right),\ p(t^*) = \exp\left[-\int_{t^*}^{t^*+1} h(u)du\right]. \tag{4a-b}$$

Despite the fact that conditional probabilities $p(t^*)$ and $q(t^*)$ are not continuous functions, their relations with the intensity function of liquidation $h(t^*)$ and the survival function $S(t^*)$, for instance for $t^*_1 < t^*_2$ we have that: $p\left(t_1^*, t_2^*\right) = \exp\left[-\int_{t_1^*}^{t_2^*} h\left(u\right)du\right] = \frac{\mathrm{S}\left(t_2^*\right)}{\mathrm{S}\left(t_1^*\right)}$ and $q\left(t_1^*, t_2^*\right) = 1 - \exp\left[-\int_{t_1^*}^{t_2^*} h\left(u\right)du\right] =$

$= \frac{S(t_1^*) - S(t_2^*)}{S(t_1^*)}$. In general terms, these relations can be written as:

$$p(t^*) = \frac{S(t^*+1)}{S(t^*)},\ q(t^*) = \frac{S(t^*) - S(t^*+1)}{S(t^*)}. \tag{5a-b}$$

For data grouped in tables – each of these functions can be determined on the basis of estimators of table functions taking into account: the number of persisting enterprises (n_t), liquidated (z_t) and censored (c_t) entities [Mikulec 2018] – however, the access to primary data is not always possible. Theoretically, knowing even one of them produces an unambiguous empirical distribution and allows to determine the remaining functions related to the time duration of enterprises. **In practice, in a non-parametric approach, to determine the values of all the remaining functions on the basis of the relations between them it is also sufficient to know at least one of them, i.e. $f(t^*)$, $F(t^*)$, $S(t^*)$, $h(t^*)$, $H(t^*)$, $p(t^*)$ or $q(t^*)$.** Knowing any two functions, and in particular both the hazard functions $h(t^*)$, $H(t^*)$ (see: formulas (11), (17a-b), (18) and (20), (23)), allows us to calculate the values of all the other functions without any problem but with a different degree of accuracy:

$$f(t^*) = h(t^*)\exp\left[-H(t^*+1)\right],$$
$$S(t^*) = \exp\left[-H(t^*)\right],\ F(t^*) = 1 - \exp\left[-H(t^*)\right], \tag{6a-e}$$
$$p(t^*) = \frac{\exp[-H(t^*+1)]}{\exp[-H(t^*)]},\ q(t^*) = \frac{\exp\left[-H(t^*)\right] - \exp\left[-H(t^*+1)\right]}{\exp\left[-H(t^*)\right]}.$$

In duration analysis an exceptional popularity was gained by: survival function $S(t^*)$, hazard function $h(t^*)$ and cumulative hazard function $H(t^*)$.

3. Estimation of the value of the hazard function

Balicki [2006] discussed two methods of discrete estimation of the values of intensity function $\hat{h}(t^*)$ for data grouped in the table.

Method I: Let us assume that $h(t^*)$ is a constant value within interval $\langle t^*, t^*+1)$, approximating the continuous intensity function in this interval, and $x(t^*) = \langle t^*, t^*+1)$ is a span of the interval. A sequence of values $h(t^*)$ represented graphically by a step curve will constitute an approximation of the intensity function which is equal to the area under continuous curve $h(t^*)$ in the analysed interval of time duration [Balicki 2006, pp. 57–59]:

$$h\left(t^*\right)\cdot x\left(t^*\right)=\int_{t^*}^{t^*+1}h\left(u\right)du. \tag{7}$$

Taking into account equation $h\ t^*\right)=\frac{-d\ln S(t^*)}{dt^*}$ it can be proved that:

$$\int_{t^*}^{t^*+1}h\left(u\right)du=-\int_{t^*}^{t^*+1}d\ln S\left(u\right)=\left[-\ln S\left(t^*\right)\right]_{t^*}^{t^*+1}=\ln S\left(t^*\right)-\ln S\left(t^*+1\right). \tag{8}$$

If the result is used in equation (7) we get an estimator of the hazard function for data grouped in the table into time intervals $x\left(t^*\right)=t^*,t^*+1\right)$ – when values are estimated at the end of the interval[1]:

$$\hat{h}\left(t^*\right)=\frac{\ln\hat{S}\left(t^*\right)-\ln\hat{S}\left(t^*+1\right)}{x\left(t^*\right)}, \tag{9}$$

or

$$\hat{h}_t=\frac{\ln\hat{S}_t-\ln\hat{S}_{t+1}}{x_t}, \tag{10}$$

where: $t = 0, 1, \ldots, w$ is the number of period in the table; $\hat{S}_t$ and $\hat{S}_{t+1}$ are the values of the survival function; and x_t is a span of interval t in the table.

Moreover, when the values of survival function $\hat{S}_t$ and $\hat{S}_{t+1}$ are substituted by their estimations on the basis of the data from the table (complete data) we get an estimator of the density function in the following form:

$$\hat{h}_t=\frac{\ln\left(n_t\right)-\ln\left(n_{t+1}\right)}{x_t}, \tag{11}$$

where: n_t, n_{t+1} is the number of entities which persisted until the beginning of period number t and $t + 1$; and x_t is a span of interval t in the table.

Method II: When we use the above relation (dependence) between the intensity, density and survival functions $h(t^*)=f(t^*)/S(t^*)$ and substitute the continuous functions with their discrete equivalents, and finally assume that the survival function is in the whole interval represented – in the most basic case – by the arithmetic mean of its value in points t^* and $t^* + 1$, it can be written that [Balicki 2006, pp. 59-62]:

$$h\left(t^*\right)\cdot x\left(t^*\right)=\frac{f\left(t^*\right)\cdot x\left(t^*\right)}{0.5\cdot\left(S\left(t^*\right)+S\left(t^*+1\right)\right)}. \tag{12}$$

[1] In the analysis, if survival function S in the table is assigned to the beginning of particular intervals of time duration, then the values of function S from previous intervals should be taken for calculations, and consequently time intervals t^*, t^*+1 in formulas (5a-b), (8), (9), (12)-(14) ought to be decreased by 1 i.e. $S(t-1)$ and $S(t^*)$ should be accepted, and the numbers of periods t in formulas (10), (15), (16) for the tables should be changed into S_{t-1} and S_t, respectively.

The estimator of the hazard function, together with the included estimator of the density function are given by the following formulas:

$$\hat{h}\left(t^*\right) = \frac{\hat{f}(t^*)}{0.5 \cdot (\hat{S}\left(t^*\right) + \hat{S}\left(t^* + 1\right))}, \tag{13}$$

$$\hat{f}(t^*) = \frac{\hat{S}(t^*) - \hat{S}(t^* + 1)}{x\left(t^*\right)}, \tag{14}$$

or

$$\hat{h}_t = \frac{\hat{f}_t}{0.5 \cdot (\hat{S}_t + \hat{S}_{t+1})}, \tag{15}$$

$$\hat{f}_t = \frac{\hat{S}_t - \hat{S}_{t+1}}{x_t}, \tag{16}$$

where: $t = 0, 1, \ldots, w$ is the number of period in the table; $\hat{S}_t$ and $\hat{S}_{t+1}$ are the values of the survival function; and x_t is a span of interval t in the table.

By substituting $\hat{f}_t$, $\hat{S}_t$ and $\hat{S}_{t+1}$ with their estimations based on data from the table (complete data) we get an estimator of the density function in the form [Balicki 2006, p. 60; Landmesser 2013, p. 46]:

$$\hat{h}_t = \frac{z_t}{n_t x_t - \frac{1}{2} z_t x_t}, \; \hat{h}_t = \frac{z_t}{n_{t+1} x_t + \frac{1}{2} z_t x_t}, \tag{17a-b}$$

where: z_t is the number of enterprises liquidated in period number t; n_t, n_{t+1} is the number of entities which persisted until the beginning of period number t and $t + 1$; and x_t is a span of interval t in the table.

The estimator calculated with the use of method I is a maximum likelihood estimator and effective estimator, and the one calculated with the use of method II is also a maximum likelihood estimator. In practice, both estimators give very accurate estimations of the hazard function for the grouped data.

In the case of the censored data, as a rule we use the actuarial method of estimating the survival function and the remaining table functions. The first step is to determine, for each interval of the table, the number of entities likely to experience the event of liquidation $n_t^{'} = n_t - 0.5c_t$, where c_t is the number of censored enterprises in this interval. Moreover, it is assumed that the enterprises were exposed to liquidation, on average, for a period equal to a half of this interval (censored observations are distributed uniformly in a single period of time duration).

Method III: The actuarial method for the grouped data leads to estimating the constant intensity function with the use of intervals. Its estimator is the quotient of the number of liquidated enterprises and the mean time of survival in this interval[2]. If we assume uniformity of liquidation of entities in each interval we get [Balicki 2006, p. 85]:

$$\hat{h}_t = \frac{z_t}{\left(n_t' - \frac{1}{2} z_t\right) x_t}. \tag{18}$$

Method IV: In the work of Jackowska we can find one more formula of actuarial estimator for the table hazard function (in the case of censored data) in the form [Jackowska 2015, p. 135]:

$$\hat{h}_t = \frac{2\hat{q}_t}{x_t(1+\hat{p}_t)}, \tag{19}$$

where: $\hat{q}_t = \frac{z_t}{n_t'}$ and $\hat{p}_t = 1 - \hat{q}_t$ are estimations of conditional probability of liquidation $q(t^*)$and survival $p(t^*)$ for individual periods of numbers $t = 0, 1, \ldots, w$; and x_t is a span of interval in the table.

Method V: The author of the paper states that there are no theoretical obstacles for one more method to be added to the mentioned-above methods of estimation of the hazard function (ordinary) for the censored data. This method shall be based on method I of estimating the hazard function for the complete data. In formula (10), for the estimator of intensity function $\hat{h}_t$ we can substitute the Kaplan-Meier survival curve ($\hat{S}_t$), given by formula (25) in a further part of the paper.

4. Estimation of the value of the cumulative hazard function

Method I: The estimation of the cumulative hazard function for small samples – on the basis of detailed data – is usually based on the Nelson-Aalen estimator [Jackowska 2013, pp. 96-100]:

$$\hat{H}(t) = \begin{cases} 0 & \text{for} \quad t = 0, \\ \sum_{k=1}^{t} \frac{z_k}{n_k} & \text{for} \quad t = 1, 2, \ldots, w-1, \\ \textit{indefinite} & \text{for} \quad t = w, \end{cases} \tag{20}$$

[2] See: method II of estimating the hazard function for complete data.

where: $t = 0, 1, \ldots, w$ denotes the random points determined by complete observations where jumps of function $\hat{H}(t)$ occur (piecewise constant function); and z_k and n_k correspond to the number of enterprises liquidated in k-th time interval and persisted until the beginning of period .

Method II: The values of the cumulative intensity function can also be determined on the basis of the Kaplan-Meier survival curve[3], using the previously given property so that:

$$\hat{H}(t) = -\ln \hat{S}(t), \tag{21}$$

$$\hat{S}(t) = \begin{cases} 1 & \text{for} \quad t = 0, \\ \prod_{k=1}^{t} \left(\frac{n_k - z_k}{n_k} \right) & \text{for} \quad t = 1, 2, \ldots, w-1. \\ 0 \, or \, indefinite^{*} & \text{for} \quad t = w, \end{cases} \tag{22}$$

(*) the estimator takes an indefinite value when censored data appear in the last interval (0 when there are no censored data inthe last interval).

The estimator of the cumulative hazard function given by formula (20) is negatively biased i.e. it gives values on average lower in comparison with the Nelson-Aalen estimator (21), based on the survival function given by formula (22).

Method III: The estimation of the value of the cumulative hazard function – on the basis of data grouped in the table, including the censored data – can also be based on the Nelson-Aalen estimator in the form [Landmesser 2013, pp. 44-47]:

$$\hat{H}_t = \begin{cases} 0 & \text{for} \quad t = 0, \\ \sum_{k=0}^{t-1} \frac{z_k}{n'_k} & \text{for} \quad t = 1, 2, \ldots, w, \end{cases} \tag{23}$$

where: $t = 0, 1, \ldots, w$ is the number of period (time interval) in the table, while z_k is the number of enterprises liquidated in the interval k; n'_k is the number of entities in a given interval exposed to experiencing the event of liquidation: $n'_k = n_k - 0.5c_k$, and c_k is the number of censored enterprises in this interval.

Method IV: However the values of the cumulative intensity function based on the Kaplan-Meier survival curve – on the basis of the data grouped in the table, and including the censored data – are calculated with the use of formula [Jackowska 2013, p. 91]:

[3] Using reverse formula $\hat{S}(t) = \exp\left[-\hat{H}(t)\right]$ we get an estimator of the survival function asymptotically equivalent to the Kaplan-Meier estimator. It can be proved that the estimator is always bigger than or equal to the Kaplan-Meier estimator of the survival function [Ptak-Chmielewska 2016, p. 75].

$$\hat{H}_t = -\ln \hat{S}_t, \tag{24}$$

$$\hat{S}_t = \begin{cases} 1 & \text{for} \quad t = 0, \\ \prod_{k=0}^{t-1} \left(\frac{n_k' - z_k}{n_k'} \right) & \text{for} \quad t = 1, 2, \ldots, w. \end{cases} \tag{25}$$

Method V: It is worth mentioning that in the literature there are other well-known ways of determining ordinary hazard function $h(t^*)$ from cumulative hazard function $H(t^*)$ – which use the non-parametric methods of the kernel estimation [Ptak-Chmielewska 2016, pp. 75-76]. An overview of this kind of estimation, together with the problems of the selection of a smoothing parameter in the kernel estimation is given by Baszczyńska [2016].

Bieszk-Stolorz presents a relation between the Kaplan-Meier and the Nelson-Aalen estimators, which allows to determine the values of the survival function on the basis of the hazard function for discrete random variable [Bieszk-Stolorz 2013, pp. 92-94].

5. Statistical calculations – guidelines for analysis

An analysis of the discussed formulas and relations was followed by an examination of the correctness and accuracy of estimating the values of individual table functions in the duration analysis of enterprises. Data on enterprises established in the Łódzkie Voivodeship in 2001-2015 (including those which went into liquidation), were used for the purpose of the analysis, and for the purpose of presentation we used data for a cohort of enterprises established in 2001 and observed until the end of 2015. The number of enterprises established in the Łódzkie Voivodeship in 2001 reached 14.9 thousand; out of this number 10.3 thousand enterprises went into liquidation, and 4.6 thousand continued to function after the end of the observation (censored observation). Calculations were made in *STATISTICA 13.1* program and in *Excel 2013* program.

STATISTICA program supports calculations in the area of estimating values of individual probability functions in survival analysis (menu: *Statistics / Advanced linear and non-linear models / Survival analysis / Survival tables and distributions*), while special attention should be paid to the way data for analysis are introduced: one variable – survival time (e.g. years, months, days); two variables – dates of beginning and ending an observation (foundation and liquidation); and six variables which are used to give precise (day-month-year) dates of foundation and liquidation of the enterprise[4]. Moreover, the way of coding of the complete and the censored

[4] If we want to group enterprises in the survival table into annual intervals of time duration and data related to survival time, are given with the use of two variables i.e. date of foundation and date of liquidation, then *de facto* the basic time unit (for dates) for this type of analysis is equal to 1 day. To obtain annual intervals of time duration we should determine the size of step at the level of 365 or 366

data ought to be defined, together with the number of intervals or size of step (time) for intervals in the table.

The results of the analysis presented in Figure1 (*STATISTICA* program) are exhaustive and almost complete when we focus on the characteristics of the time duration of enterprises in the Łódzkie Voivodeship founded in 2001 – yet they do not include the results of the estimation of distribution function and the estimation of cumulative hazard function . A detailed description of this kind of results can be found in the work by Stanisz [2007, pp. 365-370].

Tabela (Dane 2001 STAT)
Log wiarygodności danych:

Przedz.	Przedz. Szerok.	Liczba wchodz.	Liczba uciąt.	Liczba zagroż.	Liczba zgonów	Proporc. zgonów	Proporc. przeżyw.	Skum.pro przeżyw.	Gęstość prawdop.	Stopa hazardu	Błąd std Skum.prz	Błąd std Gęst.pr.	Błąd std Stopa h.	Mediana czas ocz	**Błąd std czas ocz**
L.pocz1	366,0000	14896	0	14896,00	825	0,055384	0,944616	1,000000	0,000151	0,000156	0,000000	0,000005	0,000005	3146,944	16,68036
L.pocz2	366,0000	14071	0	14071,00	825	0,058631	0,941369	0,944616	0,000151	0,000165	0,001874	0,000005	0,000006	2893,696	16,21187
L.pocz3	366,0000	13246	0	13246,00	599	0,045221	0,954779	0,889232	0,000110	0,000126	0,002571	0,000004	0,000005	2779,478	43,60601
L.pocz4	366,0000	12647	0	12647,00	623	0,049261	0,950739	0,849020	0,000114	0,000138	0,002933	0,000004	0,000006	2630,750	37,35023
L.pocz5	366,0000	12024	0	12024,00	1006	0,083666	0,916334	0,807197	0,000185	0,000239	0,003232	0,000006	0,000008	2471,662	36,41866
L.pocz6	366,0000	11018	0	11018,00	1097	0,099564	0,900436	0,739662	0,000201	0,000286	0,003595	0,000006	0,000009	2495,158	42,78151
L.pocz7	366,0000	9921	0	9921,00	674	0,067937	0,932063	0,666018	0,000124	0,000192	0,003864	0,000005	0,000007	2640,713	56,40406
L.pocz8	366,0000	9247	0	9247,00	998	0,107927	0,892073	0,620771	0,000183	0,000312	0,003975	0,000006	0,000010	2562,000	0,00000
L.pocz9	366,0000	8249	0	8249,00	1339	0,162323	0,837677	0,553773	0,000246	0,000483	0,004073	0,000006	0,000013	2196,000	0,00000
L.pocz10	366,0000	6910	0	6910,00	483	0,069899	0,930101	0,463883	0,000089	0,000198	0,004086	0,000004	0,000009	1830,000	0,00000
L.pocz11	366,0000	6427	0	6427,00	551	0,085732	0,914268	0,431458	0,000101	0,000245	0,004058	0,000004	0,000010	1464,000	0,00000
L.pocz12	366,0000	5876	0	5876,00	449	0,076413	0,923588	0,394468	0,000082	0,000217	0,004004	0,000004	0,000010	1098,000	0,00000
L.pocz13	366,0000	5427	0	5427,00	397	0,073153	0,926847	0,364326	0,000073	0,000207	0,003943	0,000004	0,000010	732,000	0,00000
L.pocz14	366,0000	5030	5	5027,50	323	0,064247	0,935753	0,337675	0,000059	0,000181	0,003875	0,000003	0,000010	366,000	0,00000
L.pocz15		4702	4609	2397,50	93	0,038790	0,961210	0,315980			0,003809				

Fig. 1. Table of survival of enterprises founded in the Łódzkie Voivodeship in 2001

Source: author's own calculations *STATISTICA 13.1*.

A module related to survival analysis for the data grouped in the table is also available in *SAS* program (*Survival Tables ...*) and *SPSS* program (*Mortality Tables ...*).

It is worth mentioning that in `R` environment we can find a number of packages supporting the estimation of the hazard function, e.g. on the basis of the Kaplan-Meier survival function – `epiR`. However, as a rule they use advanced methods of statistical analysis, e.g. regression with the use of non-parametric smoothing of the hazard function through splines or B-splines – `polspline`, `gss`, `logspline`, `bshazard`; kernel estimators – `muhaz`, and simultaneously, use detailed data instead of grouped in the tables.

The analysis carried out in *Excel* program (the author's own calculations) was based on constructing a spreadsheet which enables to calculate the values of the functions discussed above on the basis of primary data as well as on the basis of different formulas – making use of the relations (dependencies) between these functions – and checking the accuracy of the estimations.

Analysis I – only hazard function $\hat{h}_t$ is known:

If we had access to the results of only one table function in the area of survival analysis – or the function was calculated on the basis of primary data (taking into account the type of data – complete, censored) – of hazard function $\hat{h}_t$ (see: formulas

days. When value $x_t = 366$ is given (see: formulas (10), (11), (16), (17ab), (18) and (19)), it results in a proper grouping of the examined units according to intervals 01.01-31.12, i.e. according to calendar years, and so correct estimations of individual functions.

(11), (17a-b), (18)), it would be possible to determine all the other table functions in the following order: $\hat{q}_t$ on the basis of formulas (3), (7) and (18); $\hat{p}_t$ (4a); $\hat{S}_t$ (25); $\hat{H}_t$ (24); $\hat{F}_t$ (definition) and $\hat{f}_t$ (14). A comparison of the results for the estimation of values of these additional functions for tables on the basis of primary data (*STATISTICA*) and on the basis of hazard function $\hat{h}_t$ is presented in Figure 2.

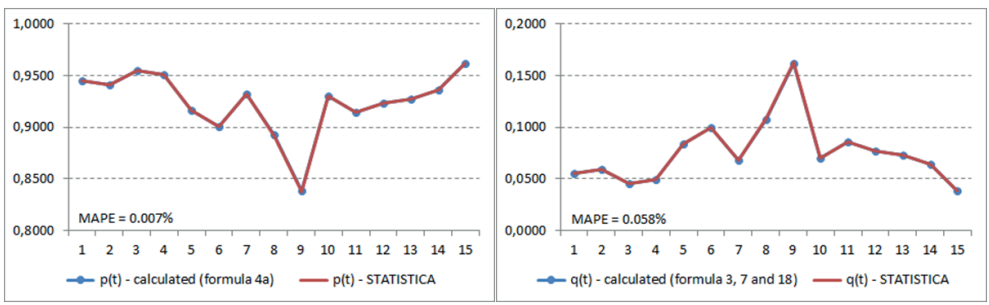

Fig. 2. Estimations of selected table functions based on survival function and on primary data

Source: author's own calculations.

Theoretical relations between the functions in the duration analysis allowed us to obtain an excellent approximation of the values of the remaining table functions on the basis of the results of survival function. The mean absolute percentage error *MAPE* in the values of individual functions ranged from 0.007% (conditional survival function) to 0.058% (conditional function of experiencing the event of liquidation).

Analysis II – only cumulative hazard function $\hat{H}_t$ is known:

If we had access to the results of only one table function from the area of analysis of survival, or if the function was calculated on the basis of primary data (taking into account the type of data, complete or censored) of cumulative hazard function $\hat{H}_t$ (see: formulas (20), (23)), there would be a possibility for all the other table functions to be determined in the following order: $\hat{S}_t$ on the basis of formulas (6b) and (23);

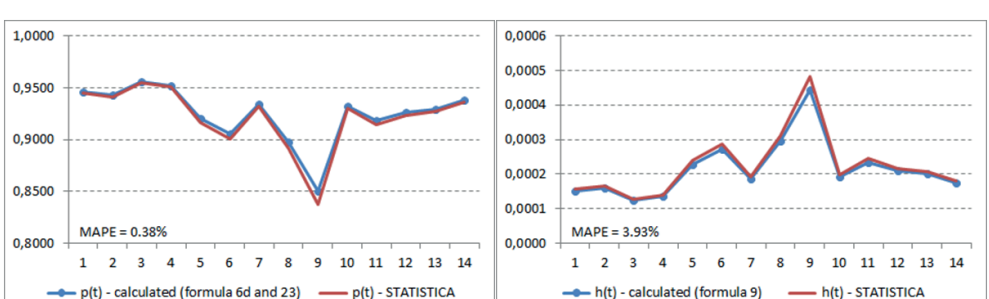

Fig. 3. Estimations of selected table functions based on survival function and on primary data

Source: author's own calculations.

$\hat{F}_t$ (definition); $\hat{p}_t$ (6d) and (23); $\hat{q}_t$ (6e) and (23); $\hat{f}_t$ (14) and $\hat{h}_t$ (9). A comparison of the results for the estimations of the values of these additional functions for tables on the basis of primary data (*STATISTICA*) and on the basis of cumulative hazard function $\hat{H}_t$ is presented in Figure 3.

The theoretical relations between the functions in the duration analysis allowed to obtain a good approximation of the values of the remaining table functions on the basis of the results for the survival function. The mean absolute percentage error *MAPE* in the values of the individual functions ranged from 0.38% (conditional survival function) to 3.93% (ordinary hazard function).

Analysis III – only survival function $\hat{S}_t$ is known:

If the results were accessible for only one table function in the area of survival analysis – or the function was calculated on the basis of primary (taking into account the type of data complete or censored), then the best option would be survival function $\hat{S}_t$ (see: formulas (22), (25)). It has been well-examined theoretically and has direct relations with all the other functions: $\hat{f}_t$ on the basis of formulas (16) and (25); $\hat{F}_t$ (definition); $\hat{h}_t$ (10) and (25); $\hat{H}_t$ (24) and (25); $\hat{p}_t$ (5a) and (25), and finally $\hat{q}_t$ (5b) and (25). A comparison of the results for the estimation of the values of these additional functions for tables based on primary data (*STATISTICA*) and on survival function $\hat{S}_t$ is presented in Figure 4.

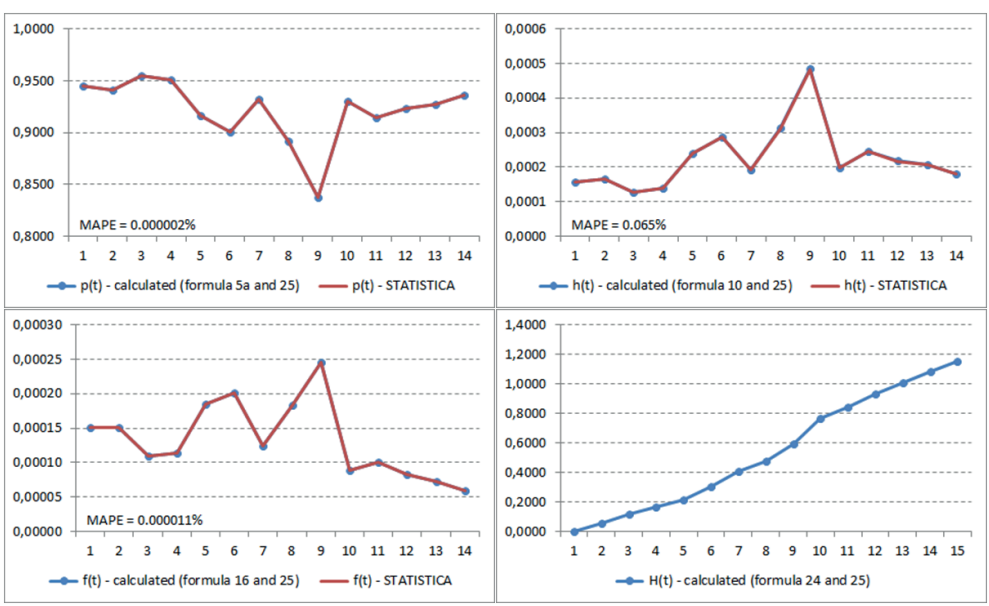

Fig. 4. Estimations of selected table functions based on survival function and on primary data

Source: author's own calculations.

The theoretical relations between the functions in the duration analysis allowed us to obtain a very good approximation of the remaining table functions on the basis of the results of the survival function. The mean absolute percentage error *MAPE* in the values of the individual functions was equal to a maximum 0.065% (ordinary hazard function).

Analysis IV – only hazard function $\hat{h}_t$ and cumulative hazard function $\hat{H}_t$ are known:

If the results for both hazard functions in the table were accessible or if they were calculated on the basis of primary data (see: formulas (11), (17a-b), (18) and (20), (23)) – taking into account the type of data (complete or censored), there would be a possibility of calculating the values of the remaining five table functions: $\hat{f}_t$, $\hat{S}_t$, $\hat{F}_t$, $\hat{p}_t$ and $\hat{q}_t$ given by formulas (6a-e) on the basis of $\hat{h}_t$ (18) and $\hat{H}_t$ (23). A comparison of the results for the estimation of the values of these additional functions for tables on the basis of primary data (*STATISTICA*) and on the basis of hazard function $\hat{h}_t$ (ordinary) and $\hat{H}_t$ (cumulative) are presented in Figure 5.

The theoretical relations between the functions in the duration analysis allowed to obtain a good approximation of the values of the remaining table functions based on the results of both hazard functions. The mean absolute percentage error *MAPE* in the values of the individual functions ranged from 0.38% (conditional survival function) to 3.81% (conditional function of experiencing the event of liquidation).

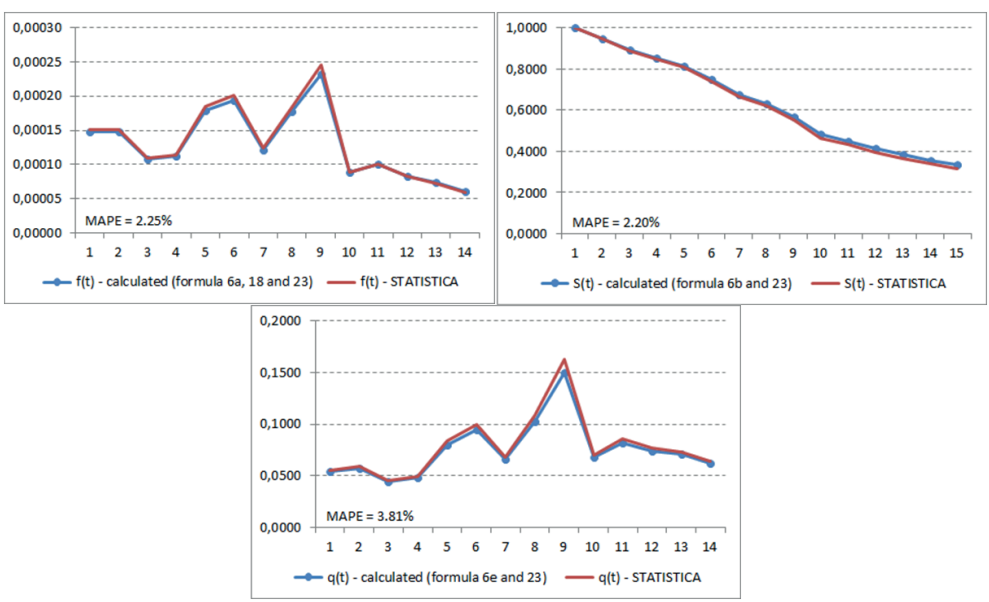

Fig. 5. Estimations of selected table functions based on hazard functions and and on primary data

Source: author's own calculations.

A compilation of the errors in the estimation of the values of the individual table functions for all the analyses discussed here is presented in the table below – the minimum level of error is given in italics and the maximum level of error is given in bold.

Table 1. Errors in the estimation of the values of the individual table functions

Function	Mean absolute percentage error (*MAPE*)*			
	analysis I	analysis II	analysis III	analysis IV
$\hat{f}_t$	0.052%	2.28%	0.000011%	2.25%
$\hat{F}_t$	0.043%	2.80%	0.000009%	2.82%
$\hat{S}_t$	0.044%	2.41%	X	2.20%
$\hat{h}_t$	x	**3.93%**	**0.065%**	x
$\hat{H}_t$	not available	x	not available	x
$\hat{q}_t$	**0.058%**	3.81%	*0.000002%*	**3.81%**
$\hat{p}_t$	*0.007%*	*0.38%*	0.000005%	*0.38%*

* The 'x'-sign indicates the function given in the analysis.

Source: author's own calculations.

According to the author, the discrepancies in the estimations of the values of particular table functions result from the type of base function used in the analysis and the (direct or indirect) method of calculation of the remaining table functions.

The best results were obtained in analysis I on the basis of hazard function $\hat{h}_t$, where the calculations of the values of the subsequent functions are based on the previous results for other functions. Similar estimations were obtained in analysis III, in which the values of all the table functions were calculated only on the basis of survival function $\hat{S}_t$.

In analysis IV, in which the values of the remaining functions were calculated only on the basis of hazard function $\hat{h}_t$ and cumulative hazard function $\hat{H}_t$ and in analysis II, using cumulative hazard function $\hat{H}_t$ – errors in the estimation of the remaining table functions were also very similar but higher than in the above-mentioned analyses.

6. Conclusions

The paper gives an overview of the theoretical properties of the individual functions applied in the duration analysis of enterprises and the relations between them. Both in theory and in practice it proves sufficient to know one of the functions discussed above, i.e. $f(t^*)$, $F(t^*)$, $S(t^*)$, $h(t^*)$, $H(t^*)$, $p(t^*)$ or $q(t^*)$, and on the basis of the relations between them it is possible to calculate the values of all the remaining functions.

The results of the theoretical considerations and calculations give grounds to state that survival function $S(t^*)$ and hazard function $h(t^*)$ are the most universal

functions in survival analysis. They make the basis for a highly accurate estimation of values of all the other table functions in the non-parametric duration analysis of enterprises.

The most accurate estimations of the values of the remaining table functions were obtained on the basis of hazard function $h(t^*)$, yet in analysis I for further calculations the survival function was also used. The mean absolute percentage errors did not exceed 0.058%. Thus it can be stated that the hazard function (ordinary) plays an essential role in survival analysis.

In analysis III, the results obtained for table functions on the basis of survival function $S(t^*)$ were very good, and better than the results based on both hazard functions $h(t^*)$ and $H(t^*)$ combined, or on cumulative hazard function – analyses IV and II, respectively. The mean absolute percentage errors did not exceed 0.065%. The results of analysis IV showed that with the use of both hazard functions and we obtained smaller errors in the estimation of the values of individual table functions than the ones obtained in analysis II.

However, calculations of the values of table functions based on cumulative intensity function $H(t^*)$ – analysis II, present a good alternative in the case when functions $h(t^*)$, or $S(t^*)$ are unknown. The mean absolute percentage errors did not exceed 3.93%. Moreover, knowledge of the cumulative hazard function (the Nelson-Aalen estimator diagram) may be applied to construct the parametric model – the diagram approximately linear is a proof of exponential distribution.

Bibliography

Balicki A., 2006, *Analiza przeżycia i tablice wymieralności*, PWE, Warszawa.

Baszczyńska A., 2016, *Parametr wygładzania w estymacji jądrowej funkcji gęstości dla zmiennych losowych w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Bieszk-Stolorz B., 2013, *Analiza historii zdarzeń w badaniu bezrobocia*, Wydawnictwo volumina.pl Daniel Krzanowski, Szczecin.

Frątczak E., Sienkiewicz U., Babiker H., 2014, *Analiza historii zdarzeń. Elementy teorii, wybrane przykłady zastosowań*, Oficyna Wydawnicza SGH, Warszawa.

Jackowska B., 2013, *Modele dalszego trwania życia oraz ich zastosowania w przypadku osób starszych*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Jackowska B., 2015, *Analiza kohortowa czasu istnienia mikroprzedsiębiorstw w Gdańsku*, Zarządzanie i finanse. Journal of Management and Finance, 13(4), pp. 127-145.

Landmesser J.M., 2013, *Wykorzystanie metod analizy czasu trwania do badania aktywności ekonomicznej ludności w Polsce*, Wydawnictwo SGGW, Warszawa.

Markowicz I., Mikulec A., 2018, *Trwanie przedsiębiorstw w Łodzi i Szczecinie – analiza porównawcza*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 507, Taksonomia 30, Klasyfikacja i analiza danych – teoria i zastosowania, K. Jajuga, M. Walesiak (eds.), UE, Wrocław, pp. 151-160.

Mikulec A., 2018, *Kohortowe tablice trwania przedsiębiorstw w województwie łódzkim*, Wiadomości Statystyczne, 5(684), pp. 56-77.

Ptak-Chmielewska A., 2016, *Determinanty przeżywalności mikro- i małych przedsiębiorstw w Polsce*, Oficyna Wydawnicza SGH, Warszawa.

Stanisz A., 2007, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*, StatSoft, Kraków.

FUNKCJA HAZARDU I JEJ ZNACZENIE W NIEPARAMETRYCZNEJ ANALIZIE TRWANIA PRZEDSIĘBIORSTW W WOJEWÓDZTWIE ŁÓDZKIM

Streszczenie: W analizie trwania przedsiębiorstw z reguły wyznacza się cztery podstawowe funkcje związane z czasem ich trwania, tj.: funkcję gęstości; dystrybuantę; funkcję przetrwania oraz hazardu. Okazuje się jednak, że to funkcja hazardu oraz jej wersja skumulowana są kluczem do zrozumienia nowoczesnej analizy przeżycia. Celem artykułu jest wskazanie najlepszego sposobu szacowania wartości poszczególnych funkcji w analizie przeżycia na bazie innych funkcji. W artykule zostały zebrane i usystematyzowane informacje o poszczególnych funkcjach wykorzystywanych w nieparametrycznej analizie czasu trwania przedsiębiorstw. Przedstawiono teoretyczne i praktyczne zagadnienia związane z wyznaczaniem m.in. funkcji hazardu oraz skumulowanej funkcji hazardu na podstawie danych z tablic kohortowych oraz wyników estymacji funkcji przeżycia metodą Kaplana-Meiera. Rozważania przedstawione w pracy zilustrowane zostały wynikami analiz dla przedsiębiorstw powstałych (w tym zlikwidowanych) w województwie łódzkim w latach 2001-2015.

Słowa kluczowe: przedsiębiorstwa, analiza trwania, funkcja hazardu, skumulowana funkcja hazardu, średni absolutny błąd procentowy.