International Journal of Business Quantitative Economics and Applied Management Research
Volume-3, Issue-3, August-2016 ISSN No: 2349-5677

SOME ASPECTS OF SURVIVAL ANALYSIS AND ITS USE IN ECONOMICS

Daniela-Emanuela Dănăcică
Department of Finance and Accounting
Faculty of Economics, Constantin Brâncuși University of Târgu-Jiu
Târgu-Jiu, Romania
danadde@yahoo.com

Abstract

The aim of this article is to present some methodological aspects of survival analysis and its use in economics. Survival analysis can be used in the economic research to investigate complex phenomena such as unemployment, employment, inflation, bank loans supply and demand, life expectancy of products, consumer behavior, migration, etc. Due to their particularities, survival data requires a different statistical approach than quantitative data; methodological and applicative problems of survival analysis and its use are presented in the paper.

Key words: survival, curves, hazard, tests

I. INTRODUCTION

Old studies using mortality table can be considered as the origin of survival analysis (Lee, 1997). Survival analysis is actually a concept that brings together a series of techniques and statistical models used for modeling duration of time from the origin until an event occurs (time-to event data). Such events are named in the literature "failure" and can be: death or failure of a therapy, time to first occurrence of a tumor in medical research, or may take other forms depending on the area of study (e.g. operating time of a machine, survival time until death, survival time until reemployment, etc).
Although survival analysis was initially used in studying death as an event in biostatistics (Armitage, 1971; Pike, 1966; Peto & Lee, 1973), and in demographic studies (Berkson and Gage, 1950; Cutler and Ederer 1958, Gehan, 1969), since the 70s has become increasingly used in economics and social sciences. Due to their particularities, survival data requires a different statistical approach than quantitative data. Survival data are not normally distributed, most often having an asymmetric distribution. Usually the histogram of durations for a group of subjects is positive asymetrical, having a longer tail at right intervals containing the largest number of observation. Another characteristic of survival data is that often have incomplete

information, subjects who do not realize the default event at the end of study. Incomplete information requires right or left censoring. A detailed presentation of survival analysis is presented by Altman (1991), Hosmer and Lemeshow (1999), Therneau and Grambsch (2001), Collett, (2003), Lee and Wang, (2003) and Klein and Moeschberger, (2005).

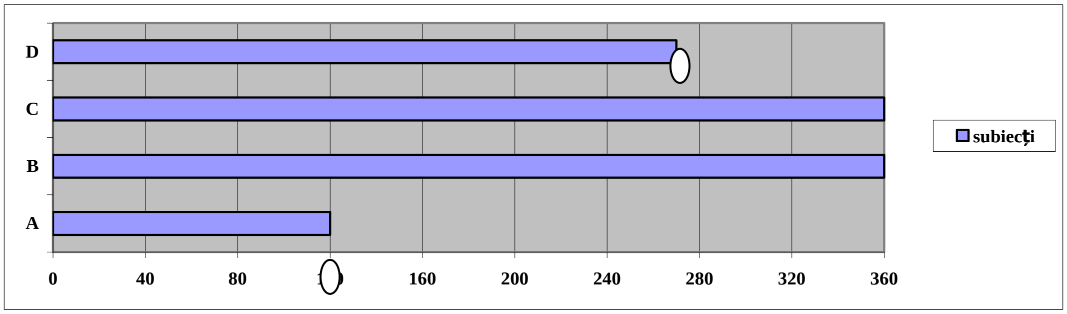
Survival analysis can be used in the economic research to investigate complex phenomena such as unemployment, employment, inflation, bank loans supply and demand, life expectancy of products, consumer behavior, etc.

In order to estimate survival time we need three main elements: the initial time (time at the beginning of study), the occurrence of the event and the measurement scale for time. If a subject does not accomplish the default event, then the survival time is censored. In the literature (Collett, 2003; Klein and Moeschberger, 2005) there are described three types of censoring:

a) Right censoring, that occur when the subject does not realize the default event (e.g. death or failure) until the end of study. In this situation we cannot determine the time passed until event occurs. The subjects that are lost from the observation are right-censored too. For example, after being selected to participate in a clinical trial, a subject changes his residence to another city or even to another country and cannot be observed anymore in the study for that was selected. Another case of right-censorship appears when "death", the default event, occurs from other causes, unrelated to the procedures fixed in the study (e.g. when the default event is re-employment, we can have subjects who exit from unemployment due to other causes, like expiry of legal period for getting unemployment allowance, or retirement, or maternity leave, etc.). Knowing and including in the analysis of censored subjects is important. However, as Greene (2003) underlines, a large number of censored subjects can affect the accuracy of the statistical tests.

In figure 1 we have a graphical representation of survival data with right-censoring. On the abscissa we have the observation period (time), measured in days, and on the ordinate we have the subjects observed during the study. Horizontal segments represent periods of observation for the subjects; with a white circle is marked the occurrence of event. Let suppose that the study has a period of observation of 360 days. *A* subject has the event after 120 days and *D* subject after 270 days. We can observe that *B* and *C* subjects are right-censored; they do not realize the event during the 360 days observation period.

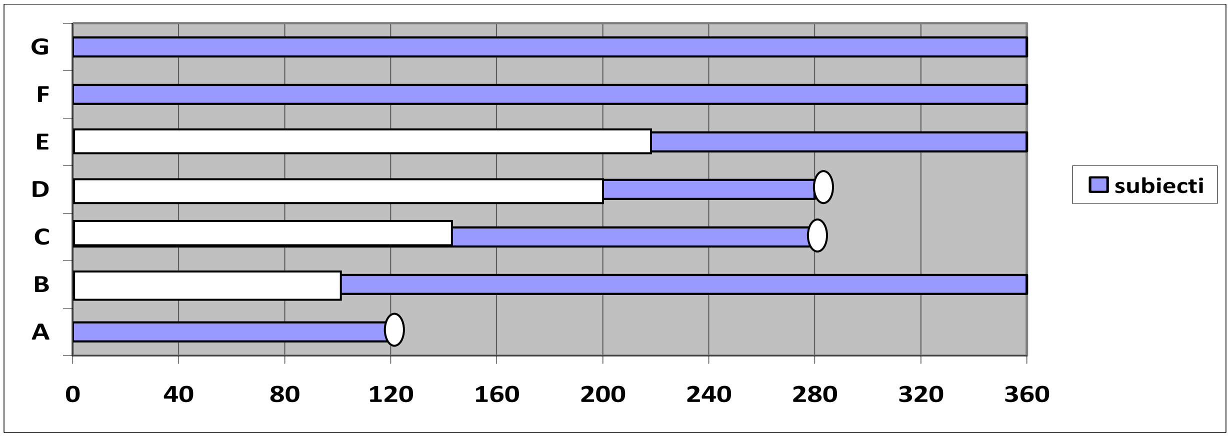
Figure 1. Graphs of survival data



Survival time (days)

b) Left censoring is necessary when the subjects do not enter at the same time in the study. Frecvently in practice we have left and right-censored subjects simultaneous. Sometimes we have to deal with progressively censored data (figure 2). The observed period is, again, 360 days. The study began at the 0 moment with the subjects *A*, *F* and *G*. After 100 days from the start of observation *B* subject enter in the sample, followed by the *C* subject (after 150 days from the start) and *D* and *E* subjects (at 200 days, respectively 220 days from the start). *A* – subject has the event after 120 days from the start, *C* and *D* subjects at 280 days from the start; the rest of subjects do not realize the event until the end of the study, beging right-censored and left-censored too.

Figure 2. Graph of survival data



Survival time (days)

c) interval censoring, necessary when a subject perform the event not at a moment of time, but during an interval.

II. SURVIVAL FUNCTION AND HAZARD FUNCTION

Two functions are of major interest for survival analysis, survival function and hazard function (Collett, 2003). We consider T a random with values different than 0, representing survival time of the sample subjects ($T \geq 0$). T variable cant take different values; all the values of T have a probability distribution and we name T as the random variable describing the survival time. The distribution function of T is given by

$$F(t) = P(T < t) = \int_0^t f(u)du \quad (1)$$

and describe the probability that survival time is less than *t*.

Let $S(t)$ be the survival function, defined as the probability that the survival time is greater than or equal with *t*. We will have:

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

Another meaning for $S(t)$ is of survival rate. For example, if we analyze unemployment duration, and we have an observation time of 1200 days, $S(10)$ is the 10-day survival rate in unemployment, $S(20)$ is the survival time estimated for the 20th day and $S(1000)$ is the 1000 days survival rate. The graph of survival function $S(T)$ versus t is called the survival curve. The survival function gives the posibility to estimate important parameters for the analysis, like median survival time and mean survival time. The $T_{0.5}$ median is a timp point when $S(T_{0.5}) = 0.5$. The mean survival time is given by the area under the survival curve. Frecvently the survival curves are positively skeewed due to the anormality of survival data distribution; therefore the value of median survival time is in these cases lower than the value of mean survival time.

The hazard function, denoted as $\lambda(t)$ is described by the following formula:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \tag{3}$$

Where $\lambda(t)$ is the risk, or death failure (the occurence of the event) at the t moment of time. $\lambda(t)$ function is referend in the literature as the hazard rate, the instantaneous death rate, the the risk rate and measure probability that a subject fail at a time *t*, conditional on the subject survived until that time *t*. Higher values of hazard function can be interpreted as an increase of the risk that the event occurs. According to Le (1997), the hazard function can increase, deacrese, or remain constant with time for long-term and short term risks. If the hazard remain constant we have an exponential model that allows us the estimate the hazard rate for different groups of analyzed subjects.

From the equation (3) we can notice that $\lambda(t)\delta t$ is the probability that a subject to "die" (the event occurs) during the time interval ($t, t + \delta t$) conditional of subject survival until the moment *t*. From the equation (3) we can obtain useful relations between survival function and hazard function (Collett, 2003). From the probability theory we know that the probability of an event noted with *A*, conditional of the occurence of an event *B* is given by the formula $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ where $P(A \cap B)$ is the joint probability of the occurence of event *A* and *B*. Therefore, the conditional probability from the definition of the hazard function presented in the equation (3) is $\frac{P(t \le T < t+\delta t)}{P(T \ge t)}$, which is equal to $\frac{F(t+\delta t)-F(t)}{S(t)}$, where $F(t)$ is the distribution function of T.

We will have:

$$\lambda(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t)-F(t)}{\delta t} \right\} \frac{1}{S(t)} \qquad (4)$$

$\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t)-F(t)}{\delta t} \right\}$ is the first order derivate of the function *F(t)* with respect to *t* which is *f(t)* and we will have:

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad (5)$$

Therefore $\lambda(t) = \frac{-d}{d(t)}\{\log S(t)\}$ (6)

and $S(t) = \exp\{-\Lambda(t)\}$ where $\Lambda(t) = \int_0^t \Lambda(u)du$ (7)

In survival analysis $\Lambda(t)$ function is called cumulative hazard. From the equation (7) we can notice that the cumulative hazard function can be obtianed from survival function, because:

$$\Lambda(t) = -\log S(t) \qquad (8)$$

III. ESTIMATING THE SURVIVAL FUNCTION

A first step in using survival analysis is to present numerical or graphical summarizations for survival time of observed subjects.

The most used techiques to estimate survival function are life table estimator and Kaplan-Meier product limit estimator. We will present here the Kaplan-Meier product limit estimator (Kaplan and Meier, 1985). Kaplan-Meier method is a non-parametric techique to estimate survival.

Let be $S(t)$ the probability as a subject from a given observed population, to have a survival time longer than *t*. For a sample from this population, with a size *N*, the survival time until the "event" occur is: $t_1 \le t_2 \le t_3 \le .... \le t_N$. For each t_i we have an n_i, the number of subjects with the death risk occurence right before t_i and the number of death subjects at the moment t_i. The time intervals between each moment of event occurence are not necessary equal. For example, if we have a sample with 10 subjects, a subject can have the event in the second day, another one can be censored after 7 days from the begining of study, and another one may have the event after 15 days. Thus we will have $t_1 = 2,\ t_2 = 15,\ n_1 = 10,\ n_2 = 8$ şi $d_1 = 10,\ d_2 = 1$. The Kaplan-Meier estimator has the following formula:

$$\hat{S}(t) = \prod_{t_i<t}^{k} \frac{n_i - d_i}{n_i} \qquad (9)$$

If we do not have censored time intervals in the sample, we will have $n_i - d_i = n_{i+1},\ i = 1,2,3.....k$ and the Kaplan-Meier estimator has the expression:

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times .... \times \frac{n_{k+1}}{n_k} \qquad (10)$$

The Kaplan-Meier estimator is given by the product of a series of estimated probabilities. To facilitate understanding the differences between analyzed subpopulations, we can use, as a graphical visualization, the Kaplan-Meier curve. We will emphasize the above presented aspects with a small example. We will use a small random sample of 10 unemployed subjects registered at National Agency for Employment Romania, from the total registered individuals in between 2008-2010. The 10 subjects sample is presented in table 1. The preestablished event is exit from unemployment due to (re)employment.

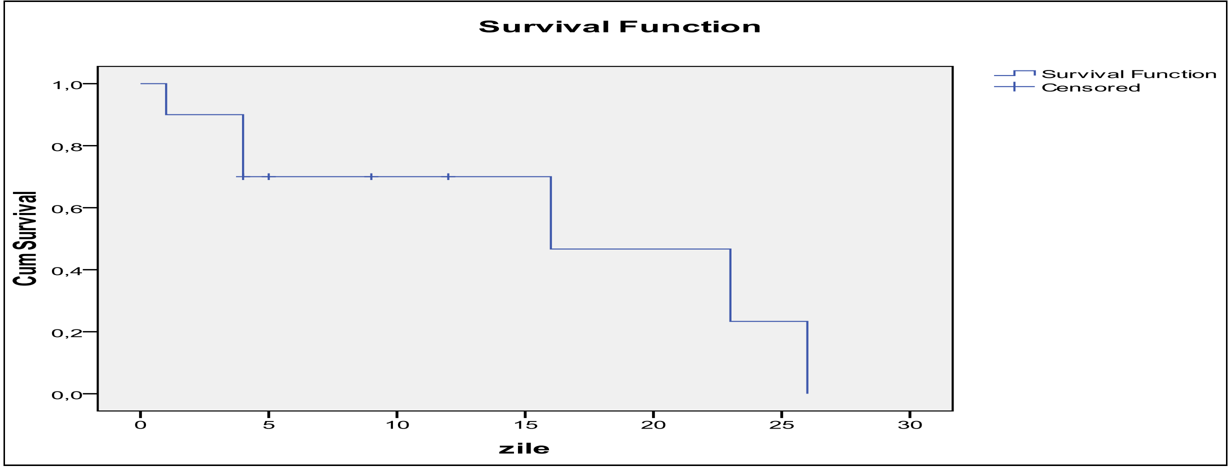
TABLE 1. Statistical data about 10 unemployed subjects

Subjects	Gen	Age	Unemployment duration	Censored/ Notcensored
1	Feminin	19	1	NO
2	Masculin	20	4	YES
3	Feminin	21	4	NO
4	Masculin	24	4	NO
5	Feminin	45	5	YES
6	Feminin	34	9	YES
7	Masculin	35	12	YES
8	Feminin	19	16	NO
9	Masculin	55	23	NO
10	Feminin	20	26	NO

*Source of data: ANOFM, 2008

From all 10 subjects, 6 has accomplish the event, re-employment, and the rest are right - censored subjects (lost in traking or exit from unemployment due to other causes then re-employment). In figure is presented the survival curve from the above analyzed data.

Figure 3. Survival curve when the event is (re)employment



As we can notice from figure 3, the Kaplan-Meier survival curve is scalariform; the frecvency of subjects with probability of survival without the event occurs is changing each moment when the event occur. The survival rate is 100% at the curve origine until the moment of a first occurence of the event, (re)employment in our case; after this point the curve colapse until a new estimated value of survival, making a new area on which duration the survival rate is constant (Cadariu, 2004, pp. 66-67).

As we already mentionated, we can estimare the mean, median and other survival parameters of analyzed subjects. In our case mean survival time of the analyzed subjects is 16.067 days. Mean survival time can be estimated also as a sum of the trapezoidal area obtained if we fall down perpendiculars from certain points of the curves on the abscissa.

Because mean survival time can be estimated only if we have a complet survival curve, which run down to zero, it is not a frecvently used statistic in survival analysis, compared with median survival time. (Le, 1997, pp. 62]). For many studies it is easy to use instead of mean survival time another item of central tendency, namely median survial time. In our small example median survival time is 16 days. If there are exogenouse variables that influemce the survival duration, the survival curves can be used to realize comparison between analyzed subjects.

We can not use the t–statistical test for testing the differences between survival curves of two or more groups, one one hand because the abnormality of survival data distribution, and on pther hand because *t* evalates the statistical significance of a difference between mean of a two or more population, while in survival analysis, the time distribution is the one analyzed.

If we do not have censored subjects in our study, then, in order to check the significance of the observed differences between two or nmore survival curves we can use the standard tests, like Wilcoxon, Mann-Whitney test, or Kruskal-Wallis test. Wilcoxon test involves sequencing ascending survival time, without taking into account of the subjects those data belong, and then applying the statistical test *t*. However, the absence of censored subjects suppose an expectation of occurrence of the event for all the subjects from the sudy, and ttherefore a longer reseach time. Because of that, in practice, comparing a two or more survival curve is realized with the Longrank test, Breslow test and Tarone-Ware test.

The Log-Rank test, also named in the literature the Mantel-Cox test, is a non-parametric test used especially when the sample has right-censored subjects. The Log-Rank test was proposed by Nathan Mantel and named log-rank by Richard and Julian Peto. The Lon-Rank test can be easy interpreted if the difference between survival probability of a two subpopulations is always of the same sign (Cadariu, 2004, pp. 70]). When the survival curves cross, the interpretation of them is difficult.

IV. COX PROPORTIONAL HAZARD FUNCTION

The standard econometric model used in survival data analysis is the model indroduced by Cox (1972) and known in the literature as the Cox proportional hazard model.

Cox proportional hazard model is a semiparametric method that allows us to estimate the effect of different exogenouse variable on the event hazard. Cox proportional hazard model is a very popular model within duration models, because the model can be functional even in the presence of censored spells. However, in order to use Cox proportional hazard model, the data observations must be independent and the hazard rate must be constant in time.

Suppose we have „n” individual units under observation, then the model has the following form:

$$\lambda_i(t)=e^{x_i\beta}\cdot\lambda_0(t)=c_i\cdot\lambda_0(t)\quad i=1,2,\ldots,n \tag{11}$$

where $x_i=(x_{i1},x_{i2},\ldots x_{ik})$ represents the k variable vector for i unit, $\beta=(\beta_1,\beta_2,\ldots\beta_k)$ is the regression vector, $\lambda_i(t)$ is the hazard estimated for each individual unit *i*, and $\lambda_0(t)$ is the este baseline hazard. The baseline hazard coresponde to an observation for which $x_i=0$. $\lambda_0(t)$ is the time-dependent element from the Cox model, but it is independent related to exogenous variables of the model; $e^{x_i\,\beta}$ is the element of the model that dependes on the exogenous variables, but not also of time.

The hazard is the risk of event occurences (death, failure, employment in our small example..,etc) for a subject that survived until that moment of study.

The hazard rate of a group of subjects compared with another group of subjects is the difference of the hazard of the two groups. The hazard rate can be estimated using the formula: $\frac{\lambda_i(t)}{\lambda_j(t)}$.

We will have: $\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{e^{x_i\beta} \cdot \lambda_0(t)}{e^{x_j\beta} \cdot \lambda_0(t)} = e^{(x_i - x_j)\beta}$ (12)

The hazard rate show us the more likely a subject 1 is to achive the event compared with subject 2. For example, if the hazard rate of (re)employment for men is 3, and women are the reference category, we say that men have a (re)employment hazard three times higher, in any moment of the study, than women.

We will have the following interpretation of Cox regression coefficients: a negative coefficient indicates a decreased hazard of event occurence due to the exogenous variable, while a pozitive coefficient indicates an increased hazard of default event. If we have $\beta = 0$, then the hazard rate for exogenous variable is equal with $e^0 = 1$, and the conclusion is that the exogenous variable doesn't have any effect on the survival.

The interpretation of the hazard rate is similar to the interpretation of the odd ration for logistic regression. A hazard rate higher than 1 increases the risk of default event occurs due to the influence of the exogenous variable; a hazard rate lower than 1 decreases the risk of event occurrence.

In order to test the null hypothesis that the exogenous variables have no effect on survival, we use the Wald test and the likelihood ratio test. Hypotheses are verified:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$

Maximum likelihood estimation of regression coefficients β and the hazard rate is achieved by solving a set of simultaneous linear equations using Newton-Raphson technique or different iterative methods (Persson, 2002).

Cox proportional hazard model is based on the assumption that the hazard ration does not depend on the time. Exogenous variables can be stationary or dependent with time (Le, 1997). An exogenous variable is dependent on the time if the difference between the values of the exogenous variable for two different subjects varies with time. Sometimes it happens that the the asumption of hazard proportionality is not satisfied; in this case the Cox proportional hazard model does not fit. Thus, testing the proportional hazard asumption of the Cox model is vital for the acuracy of results. The literature presents different aproaches to test the assumption of proportional hazards using tests of proportionality like partitioning the time of failure, categorization of exogenous variables, the use of spline function test (Hosmer and Lemenshow, 2003), or graphical check of harzard asumption. An often procedure is the use of log-minus-log curve. Another graphical method is the analysis of partial residuals. Partial residuals are defined only for censored cases.

If the hazard assumption is not verified for some exogenous variable, the literature suggests several approaches. Thus, a first opinion is to buid a not proportional hazard model, specifying in it the interaction between time and the exogenous variable. Such a model is called Cox model with time-dependent covariates. Another option is to get a non-proportional hazard model obtained by stratification of the categorial explanatory variables. There are situations when we do not have only one possible event (e.g. death, failure), but several events that may occur (like exit from unemployment due to: reemployment, expiry of the legal period for receiving unemployment benefits, transition into inactivity, etc). In this case we are not dealing with a single risk model, but a model with multiple risks, referred in to literature as competing-risks model. (Jensen & Westergaard – Nielsen, 1990) underlines that the use of a competing-risks model increase the useful information compared with a single-risk model, thefere, a competing-risks model is a better option.

REFERENCES

1. Altman, D.G. "Practical Statistics for Medical Research", Chapman and Hall, London, 1991
2. Armitage, P., "Statistical Methods in Medical Research", Oxford, Blackwell, 1971
3. Berkson, J. and R.P. Gage, "Calculation of Survival Rates for Cancer", Proceedings of Staff Meetings of the Mayo Clinic, 25, 1950, pp. 270-286.
4. Borsic D., et. all, "Cox Regression models for unemployment duration in Romania, Austria, Slovenia, Croatia and Macedonia", Romanian Journal of Economic Forecasting, (2), 2009, pp. 81-104.
5. Cadariu, A. A., "Methodology of Research in Medical Science", 2004, disponibil la: http://www.info.umfcluj.ro/resurse/Laborator/Metodologie/LabStoma/Materiale/CursMetodologie.pdf
6. Collett, D., "Modeling Survival Data in Medical Research", 2nd edition, 2003, Taylor & Francis.
7. Cox, D.R., "Regression Models and Life Tables", Journal of Royal Statistical Society B34, 1972, pp. 187-220.
8. *Cutler*, S. J. and *Ederer* F., "Maximum Utilization of the Life Table Method in Analyzing Survival", Journal of Chronic Diseases, 8, 1958, pp. 699-712.
9. Dănăcică D., "Cercetări privind impactul factorilor ce influențează durata șomajului și probabilitatea (re)angajării în România", Editura Academiei Române, 2013, Bucureşti.
10. Gehan, E.A., Estimating Survival Function for the Life Table, Journal of Chronic Diseases, 21, 1969, pp. 629-644.
11. Greene, W. H., "Econometric Analysis", 2003, New York: Prentice-Hall.
12. Han, A. and J. Hausman, "Flexible Parametric Estimation of Duration and Competing Risks Models", Journal of Applied Econometrics, 5, 1990 pp.1-28.

13. Hosmer, D. H. and S. Lemeshow, "Applied Survival Analysis: Regression Modelling of Time to Event Data", 2003, New York:Wiley-Interscience.
14. Jensen P. and Westergaard-Nielsen N., "Temporary Layoffs. In Panel Data and Labour Market Studies" (eds.) J. Hartog, G. Ridder & J. Theeuwes. North- Holland, 1990, Amsterdam.
15. Kaplan, E. L. and Meier, P., "Nonparametric Estimation from Incomplete Observations", Journal of American Statistical Association, 53, , 1958, pp. 457–81.
16. Kiefer N.M., "Economic Duration Data and Hazard Functions", Journal of Economic Literature, 26, 1988, pp. 646-679.
17. Klein, J. P., and M. L. Moeschberger, "Survival Analysis: Techniques for Censored and Truncated Data", 2005, New York: Springer Verlag.
18. Le, C.T., "Applied Survival Analysis", John Wiley & Sons, 1997, New York.
19. Lee, E.T. and J. Wang, "Statistical Methods for Survival Data Analysis", 3rd edition, NewYork, 2003, John Wiley & Sons.
20. Nickell, Stephen J., "Estimating the Probability of Leaving Unemployment", Econometrica, 47 (5), 1979, pp. 1249-1266.
21. Peto, R. and Lee, P., "Weilbull Distributions for Countinous Carcinogenesis Experiments". Biometrics 29, 1973, pp. 457–470.
22. Pike, M.C., "A Method of Analysis of a Certain Class of Experiments in Carcinogenesis", Biometrics 22, 1966, pp. 142-161.
23. Therneau, T. M. and P. M. Grambsch, "Modeling Survival Data: Extending the Cox Model", 2001, New York: Springer Verlag.