UDC: 334.72:005.41]:303.092.9(496.5)
Conference Paper

Kernel smoothing method for hazard rate estimation: an application to Albanian firm survival

Lule Basha[1], Markela Muça[2]

[1,2] *Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania*
e-mail: lule.hallaci@fshn.edu.al

Abstract

The nonparametric approach to estimate hazard rates for lifetime data is flexible, model-free and data-driven. No shape assumption is imposed other than that the hazard function is a smooth function. Such an approach typically involves smoothing of an initial hazard estimate, with arbitrary choice of smoother. In this paper we demonstrate how we can obtain hazard estimators, by smoothing the increments of the Nelson-Aalen estimator for the cumulative hazard function, using kernel-type nonparametric method.
This paper analyses the duration of the life for new entrant Albanian firms. We estimate firms' hazards of failure and density function using some kernel estimators, based on a sample retrieved from the database of National Business Center. This sample contains 1000 firms, which were newly-established over the period January 2000 – December 2017. In this study censored data refers to those firms which were still alive at the time when the data was last updated. We also make a comparison between two classes of firms: Class I firms, Natural Person (PP), constituting 42.9% of the firms and Class II firms, Limited Liability Corporation (LLC), constituting 57.1% of the firms. All analysis were performed using R.

Keywords: Duration models, kernel-type methods, hazard function, firm exit.

1. Introduction

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. The occurrence of the event times (called survival) of individuals may be prevented by the previous occurrence of another competing event (called censoring) (Kleinbaum and Klein *et al*, 2005). The distribution of survival times is usually described or characterized by three functions: the survivorship function, the probability density function and the hazard function. The estimation of the probability density and hazard function has received considerable attention, as it allows visualizing and exploring the distribution of data. The hazard function of survival time gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval. (Lee and Wang *et al*, 2003)
Kaplan–Meier estimator provides an estimate of the survival function and the Nelson–Aalan estimator provides an estimate of the cumulative hazard rate. Although these two statistics provide an investigator with important information about the eventual death time of an individual, they provide only limited information about the mechanism of the process under study, as summarized by the hazard rate. The slope of the Nelson– Aalan estimator provides a crude estimate of the hazard rate, but this estimate is often hard to interpret (Klein and Moeschberger *et al*, 2003). These crude estimates of the hazard rate can be smoothed to provide a better estimator of the hazard rate by using a kernel-smoothing technique. Kernel estimation is one of the most important data analytical tools, if we consider the non parametric approach in the estimation of the probability density function. In parallel, it is used to estimate the hazard rate function, which is one of the most important ways for representing the life time distribution in the survival analysis. No shape restriction on the hazard rate is assumed except for smoothness. Such a model-free approach is data driven and can be used for parametric model checking. The nonparametric approach of hazard rate estimation typically involves the smoothing of an initial hazard estimate. In this paper we demonstrate how we can obtain hazard estimators, by smoothing the increments of the Nelson-Aalen estimator for the cumulative hazard function. We used

three different kernel estimators for estimating hazard function, Epanechnikov kernel, Gaussian kernel and Gamma kernel.

As an application, this paper analyses the duration of the life for new entrant Albanian firms. We estimate firms' hazards of failure using some kernel estimators and survival function, based on a sample retrieved from the database of National Business Center. This sample contains 1000 firms, which were newly-established over the period January 2000 – December 2017. The advantage of this sample is that the date of firm entry is observed so left-censoring is not an issue. In this study censored data refers to those firms which were still alive at the time when the data was last updated. We also make a comparison between two classes of firms: Class I firms, Natural Person (NP), constituting 42.9% of the firms and Class II firms, Limited Liability Corporation (LLC), constituting 57.1% of the firms.

The paper is organized as follows. Section 2 demonstrates kernel smoothing method for hazard rate estimation. In Section 3 we show the survival and hazard model of firm survival and the comparison between Natural Person (NP) and Limited Liability Corporation (LLC) and Section 4 concludes.

2. Methodology

Let $Y_1,.....Y_n$ be a variable of interest with density *f* and distribution function *F*, and let $C_1,.....C_n$ be a censoring variable with continuous distribution function *G*. We assume that *Y* is independent of *C*. Under random right censoring, the variable is not completely observed. One can only observe (T_i,δ_i) where $T=\min(Y_i,C_i)$ and $\delta_i = I(Y_i \le C_i)$ with $I(\cdot)$ being the indicator function. We denote by $h = f/(1-F)$ the corresponding hazard function and by $H = -\log(1-F)$ the cumulative hazard function also known as the Nelson–Aalen estimator. Based on Kaplan-Meier estimator proposed by Kaplan and Meier (Kaplan and Meier *et al*, 1958) several nonparametric density estimators have been proposed.

$$1-\hat{F}(t)=\prod_{i:X_i\le x}\left(1-\frac{1}{\sum_{j=1}^{n}1_{\{X_j\ge X_i\}}}\right)^{\delta_i} \tag{1}$$

The slope of the Nelson–Aalen estimator provides a crude estimate of the hazard rate. The kernel-smoothed estimator of hazard rate is a weighted average of these crude estimates over event times close to a specific time. Closeness is determined by a bandwidth *b*, which is chosen either to minimize some measure of the mean-squared error or to give a desired degree of smoothness. The weights are controlled by the choice of a kernel function, which determines how much weight is given to points at a distance from a specific time.

The kernel-smoothed hazard rate estimator is defined for all positive time points. The kernel smoothed estimator of *h(t)*, proposed by Muller and Wang (Muller and Wang *et al*, 1994) with boundary correction, by smoothing the increments of the Nelson-Aalen estimator for the cumulative hazard function, based on the kernel *K( )* is given by

$$\hat{h}(t)=\int K\left(\frac{t-t_i}{b}\right)d\tilde{H}(t_i)=\frac{1}{b}\sum_{i=1}^{n}K\left(\frac{t-t_i}{b}\right)\Delta\tilde{H}(t_i) \tag{2}$$

where *K* is a kernel function generally chosen to be a symmetric probability density function, $0<b\equiv b_n$ is a bandwidth sequence and $\Delta\tilde{H}(t)$ is the Nelson-Aalan estimator. This method is totally nonparametric and admirably impartial to special types of shapes of the underlying density.

We used three different kernel estimators for estimating hazard function. One is epanechnikov kernel, proposed by Muller and Wang (MW) with boundary correction. For the Muller estimator, we use the data-driven global optimal bandwidth as proposed by the authors. To calculate the MW estimator and its bandwidth parameter, we make use of the R package muhaz; see Hess and Gentleman. The second estimator is Gaussian kernel

$$K(t,b)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}*\left(\frac{t-t_i}{b}\right)\right) \tag{3}$$

The third estimator is gamma kernel, proposed by Bouezmarni (Bouezmarni *et al*, 2011) that corrects for the boundary effects.

$$K(x,\mathrm{b})(t)=\frac{t^{\rho_b(x)-1}\exp(-t/\mathrm{b})}{b^{\rho_b(x)}\Gamma(\rho_b(x))},\qquad \rho_b(x)=\begin{cases}\dfrac{x}{b} & \text{if } \mathrm{x}\geq 2\mathrm{b}\\ \dfrac{1}{4}\left(\dfrac{x}{b}\right)^2+1 & \text{if } \mathrm{x}\in[0,2\mathrm{b})\end{cases} \tag{4}$$

To calculate the Gaussian and Gamma estimator and their bandwidth parameters, we used the programming in R.

The variance of $\hat{h}(t)$ is estimated by the quantity

$$\sigma^2[\hat{\mathrm{h}}(\mathrm{t})]=\frac{1}{b^2}\sum_{i=1}^{n}K\left(\frac{t-t_i}{b}\right)^2\Delta\hat{V}[\tilde{H}(t_i)] \tag{5}$$

The kernel estimator has a rate of convergence of $\sqrt{nh}$, (Lo *et al*, 1989) which is slower compared with the $\sqrt{n}$ rate of convergence established in the parametric approach. However, a major complication that is emphasized in parametric modeling is the risk of biased and inconsistent parameter estimation due to misspecification problem.

3. Survival function and hazard model of firm survival

This paper analyses the duration of the life for new entrant Albanian firms. Traditionally, economists view entry and exit as serving an equilibrating role for industry dynamics in the absence of any barriers. A major problem encountered when analyzing duration data is that of censored data. In this study censored data refers to those firms which were still alive at the time when the data was last updated. We estimate firms' hazards of failure using some kernel estimators and survival function. The hazard rate can be thought of as the rate at which firms die after duration t, given that they are alive at least until time t.

The analysis is made based on a sample retrieved from the database of National Business Center (NBC). This sample contains 1000 firms, which were newly-established over the period January 2000 – December 2017. In this study censored data refers to those firms which were still alive at the time when the data was last updated. We also make a comparison between two classes of firms: Class I firms, Natural Person (NP), constituting 42.9% of the firms and Class II firms, Limited Liability Corporation (LLC), constituting 57.1% of the firms.

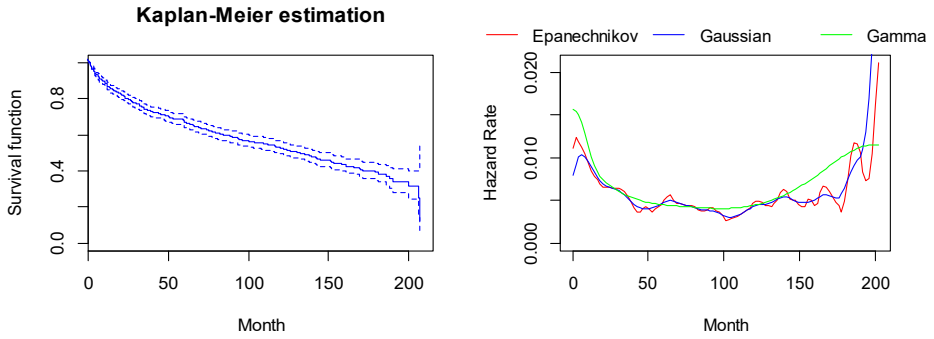
All firms own a Unique Identification Number (NUIS). Completion of registration and issuance of registration certificate are the evidence of the registration of a business registration, registration of social insurance, registration of health insurance and registration with labor inspectorate authorities. The birth time is measured as the time in which a firm is registered in (NBC). A firm's exit occurs when it is deregister. The unique NUIS each firm receives ensures against any potential over false exit and re-entry. For example, a name change by a firm will not be recorded as an exit and entry because the NUIS does not change for this firm.

Of the 1000 firms, 470 (47%) were still in business (were 'alive') in April 2018 (the end of the sample period), while 530 (53%) had ceased trading (were 'dead'). Table 1 provides a summary of the number of firms in the sample by the sub-period in which they were established and also provides a breakdown by sub-period of those which were alive or dead in April 2018.

Table 1. Number of firms established by date of establishment and status in April 2018

Sub-period	Number of firms established	Status in April 2018	
		Alive	**Dead**
2000-2004	149	84	65
2005-2008	360	206	154
2009-2012	203	94	109
2013-2017	288	146	142
2000-2017	1000	530	470

It is obvious from Table 1 that there was a large influx of firms in the period 2005-2008.



Figu 1. (a) Survival rates for firms; **(b)** Epanechnikov, Gaussian and Gamma kernels for firms' hazards of failure

The probability of continuing business activity in the first year since the moment of registering amounted to 86.6%. The first quartile life of firms is 34 months; in other words, 25% will close within 34 months of their registration. The median life of firms is 132 months; half will close within 132 months of registration form Figure 1 (a). The third quartile life of firms is 206 months; 75% will close within 206 months of registration. The mean time of survival is 120 month.

Figure 1 (b) shows epanechnikov, gaussian and gamma kernel estimators, with boundary correction, for hazard rate estimation. The hazard appears to be greatest at the beginning of follow-up time and then decreases until it levels off at around 25 months and stays low and mostly constant till 181 months. Indeed the hazard rate right at the beginning is more than 1.6 times larger than the hazard 22 months later. Thus, at the beginning of the study, we would expect around 0.0111 failures per month, while 22 months later, for those who survived we would expect 0.0066 failures per month. The sudden upticks at the end of follow-up time are not to be trusted, as they are likely due to the few number of subjects at risk at the end.

We also make a comparison between two classes of firms Figure 2: Class I firms, Natural Person (NP), constituting 42.9% of the firms and Class II firms, Limited Liability Corporation (LLC), constituting 57.1% of the firms.

For Class I firms, the probability of continuing business activity in the first year since the moment of registering amounted to 78.8%. The first quartile life of firms is 18 months; in other words, 25% will close within 18 months of their registration. The median life of firms is 68 months; half will close within 68 months of registration. The third quartile life of firms is 145 months; 75% will close within 145 months of registration. The mean time of survival is 84 month. For Class II firms, the probability of continuing business activity in the first year since the moment of registering amounted to 92.8%. The first quartile life of firms is 65 months; in other words, 25% will close within 65 months of their registration. The median life of firms is 188 months; half will close within 188 months of registration. The third quartile life of firms is 206 months; 75% will close within 206 months of registration. The mean time of survival is 144 month.

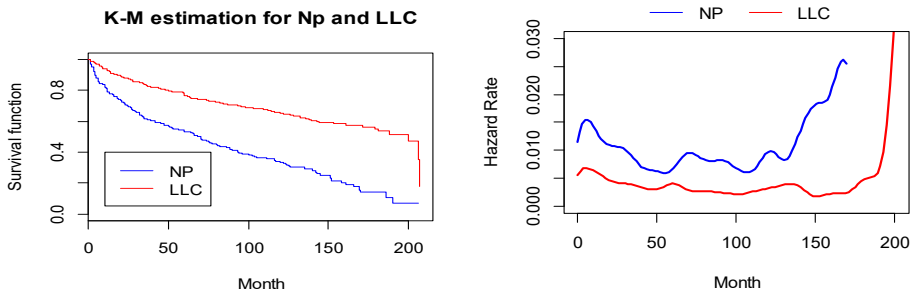


Fig. 2. (a) Survival rates for NP and LLC firms; **(b)** Hazard rate for NP and LLC firms

The estimated survival function for LLC firms lies above the survival function for NP firms throughout the entire time analyzed. For NP firms, the figure shows that risk of failure increases rapidly different from LLC firms. The hazard function is generally higher for the NP firms.

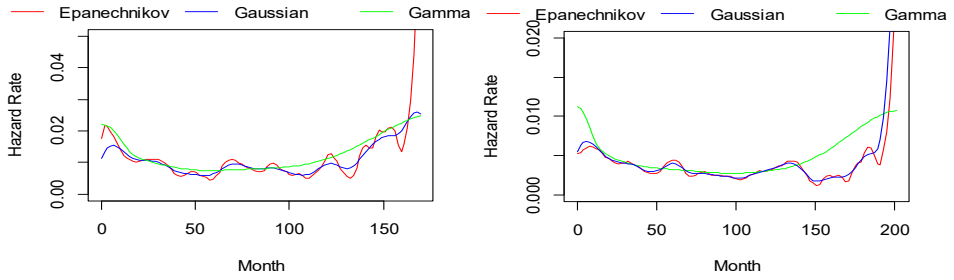


Fig. 3. (a) Epanechnikov, Gaussian and Gamma kernels for NP firms' hazards of failure; **(b)** Epanechnikov, Gaussian and Gamma kernels for LLC firms' hazards of failure

For NP firms, the hazard appears to be greatest at the beginning of follow-up time and then decreases until it levels off at around 8.4 months and stays low and mostly constant till 130 months, when it increases Figure 3 (a). For LLC firms, the hazard appears to be constant from the beginning of follow-up time and stays constant till 193 months, when it increases Figure 3 (b).

Conclusion

In this paper, we investigated kernel smoothing method for hazard rate estimation. The kernel-smoothed estimator of hazard rate is a weighted average of the crude estimates taken by Nelson–Aalen estimator, over event times close to a specific time.

We also analyzed the duration of the life for new entrant Albanian firms. The probability of continuing business activity in the first year since the moment of registering amounted to 86.6%. The mean time of survival for firms is 120 months. The hazard rate for NP firms lies above the hazard rate for LLC firms throughout the entire time analyzed. For NP firms half will close within 68 months of their registration. For LLC firms half will close within 188 months of their registration.

References

[1]. Bouezmarni, T., El Ghouch, A. and Mesfioui, M. 2011. Gamma kernel estimators for density and hazard rate of right-censored data. *Journal of Probability and Statatistics*, Vol.2011. 16 pages.

[2]. Kaplan, E. and Meier, *P. 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association*, Vol. 53, pp. 457–481.

[3]. Klein, J. P. and Moeschberger, M. L. 2003. Survival Analysis. Techniques for Censored and Truncated Data, *Vol.15 of Statistics for Biology and Health, Springer*, New York, NY, USA, 2nd edition.

[4]. Kleinbaum, D. and Klein, M. 2005. Survival Analysis. *Statistics for Biology and Health, Springer*, New York, NY, USA, 2nd edition.

[5]. Lee, E. and Wang, J. 2003. Statistical Methods for survival data analysis, *Wiley*, New Jersey, USA, third edition.

[6]. Lo, S. H., Mack, Y.P. and Wang, J. L. 1989. Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Relat. Fields* 80, 461-473.

[7]. Muller, H. and Wang, J. 1994. Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics*, Vol. 50, No. 1, pp.61–76.