JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original Article

A Non-parametric Method for Hazard Rate Estimation in Acute Myocardial Infarction Patients: Kernel Smoothing Approach

Ali Reza Soltanian (PhD)[a], **Hossein Mahjub (PhD)**[a]*

[a] Department of Biostatistics & Epidemiology and Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

ARTICLE INFORMATION

Article history:
Received: 08 January 2012
Revised: 16 February 2012
Accepted: 10 March 2012
Available online: 25 March 202

Keywords:
Acute myocardial infarction
Curve estimation
Non-parametric method
Kernel smoothing

*** Correspondence**
Hossein Mahjub (PhD)
Tel: +98 811 8380025
E-mail: mahjub@umsha.ac.ir

ABSTRACT

Background: Kernel smoothing method is a non-parametric or graphical method for statistical estimation. In the present study was used a kernel smoothing method for finding the death hazard rates of patients with acute myocardial infarction.

Methods: By employing non-parametric regression methods, the curve estimation, may have some complexity. In this article, four indices of Epanechnikov, Biquadratic, Triquadratic and Rectangle kernels were used under local and k-nearest neighbors' bandwidth. For comparing the models, were employed mean integrated squared error. To illustrate in the study, was used the dataset of acute myocardial infraction patients in Bushehr port, in the south of Iran. To obtain proper bandwidth, was used generalized cross-validation method.

Results: Corresponding to a low bandwidth value, the curve is unreadable and the regression curve is so roughly. In the event of increasing bandwidth value, the distribution has more readable and smooth. In this study, estimate of death hazard rate for the patients based on Epanechnikov kernel under local bandwidth was 1.011×10-11, which had the lowest mean square error compared to k-nearest neighbors bandwidth. We obtained the death hazard rate in 10 and 30 months after the first acute myocardial infraction using Epanechnikov kernelas were 0.0031 and 0.0012, respectively.

Conclusion: The Epanechnikov kernel for obtaining death hazard rate of patients with acute myocardial infraction has minimum mean integrated squared error compared to the other kernels. In addition, the mortality hazard rate of acute myocardial infraction in the study was low.

Citation: Soltanian AR, Mahjub H. A Non-parametric Method for Hazard Rate Estimation in Acute Myocardial Infarction Patients: Kernel Smoothing Approach. JRHS. 2012;12(1):19-24.

Introduction

Currently incidence of acute myocardial infarction is leading cause of death in developing countries[1]. The most appropriate method for calculating hazard rates from acute myocardial infarction is survival models[1].

Because of difficulty implementing cohort studies, most researchers estimate mortality rates caused by myocardial diseases using historical cohort studies[2].

In analysis of survival data, the most common methods to study the risk factors effect on patients' survival time, are Cox regression and the cumulative hazard function graph, which show cumulative hazard rate in observed times. These methods cannot show hazard rates in arbitrary time with low variance and bias mutually[3].

In survival studies, knowing the hazard rates, such as survival rates are important[4].If considering an appropriate distribution for survival time, can be calculated the related survival indices precisely. However, in most cases, it is impossible to fit a parametric distribution[4].

Moreover, since many possible models can fit to data, the proper parametric model may not exist as well as the computational cost may be high[5]. In addition, performing analysis based on censored data is more difficult than the complete data[6].

Conventional approaches for analyzing censored data and estimating of hazard function are computationally complicated and often difficult to explain to practitioners. Moreover, alternative methods such as partial likelihood and full likelihood estimations for semi-parametric and parametric models respectively have several limitations[7,8].

Non-parametric methods such as kernel smoothing do not require a previous knowledge of how the variables are distributed. In addition, these methods are understood more easily so are implemented more than other approaches[9,10]. In survival analysis, hazard rates are important for estimating transition rates. In smoothing methods, a data adaptive bandwidth improves the bias-variance trade off and reduces the boundary bias near the origin. Kernel smoothing methods used in cardiac disease especially in magnetic resonance imaging (MRI) for assessment of ventricular function in acute myocardial infarction[11].

Soltanian et al[2] estimated only cumulative hazard rates of first acute myocardial infarction on Bushehr inhabitants, in the south of Iran. A non-parametric method like kernel smoothing could apply for analyzing right censored data in an experiment.

In this study, is considered only right censored data, in which the values of the observations in one of the distribution tails were not known. Moreover, a non-parametric method for estimation of a smooth curve for hazard rates on the acute myocardial infarction people has not performed yet.

Therefore, the aim of the present study is using a set of kernels smoothing for estimation of death hazard rate in the patients with acute myocardial infarction to select a kernel as the most proper between the set of kernels.

Methods

In this study, we used the Bushehr's dataset[2], which examines the factors affecting the survival of patients with first acute myocardial infarction (AMI) on the resident population of Bushehr City, in the southern of Iran. In which, was considered the extracted baseline information of 197 patients diagnosed with AMI retrospectively.

In a checklist, were collected demographic and baseline information of patients which were admitted and hospitalized in Bushehr's hospitals. The follow up time was up to five years in terms of mortality status and cause of death. Cause of death was collected based on death certificate and burial certificates deceased. The first acute myocardial infarction considered as the starting point for each patient.

Initially, the survival rate of individuals was estimated using Kaplan-Meier estimator. Then, were estimated the hazard rates. In the next step, we have drawn smoothed graph of the hazard rates measures using non-parametric kernel smoothing. The kernel smoothed hazard rate estimator defined for all time points;

$0 = t_0 < t_1 < ... < t_D$. For time points t, $b \le t \le t_D - b$, the kernel smoothed estimator of hazard rate of based on the kernel $K(.)$ is given by

$$\hat{f}(t,b) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t-t_i}{b}\right) \qquad (1)$$

with bandwidth b and number of data points of n. Under mild conditions (bandwidth must decrease with increasing n), the kernel estimate converges in probability to the true density. Kernel function has alteration according to distribution type[12]. Choosing of optimal bandwidth can be obtained by two methods, Plug-in and Cross-Validation[5, 13-14], in which Cross-Validation method was used in the study. For choosing the best bandwidth cross-validation method is a common method. For more sensitive examination was developed a Generalized Cross-Validation using equation 2 that suggested byHardle[15],

$$GCV(b) = \frac{RRS(b)}{n\left\{1 - \frac{tr(B)}{n}\right\}^2} \qquad (2)$$

Where;, *RRS* denotes the sum of error, *tr (B)* denotes trace smoothing matrix or hat matrix and *n* shows the sample size. The *RRS* calculates as equation (3),

$$RRS = n\sum_{i=1}\left(t_i - m(u_i)\right)^2 \qquad (3)$$

where, $m(u_i)$ shows the smoothing function and t_i means the observed value. In this article, we compared between Mean Integrated Squared Error (MISE) indices of Epanechnikov, Biquadratic, Triquadratic and Rectangle kernels under local and k-nearest neighbors bandwidth[12-15] selection method. We used KernelSmooth, MASS and mvtnorm under R 2.12.1 software for data analysis.

Results

The totally 197 patients were enrolled in the present study. All patients followed to the end of the study. Twenty-seven (13.7%) subjects were died and two subjects were withdrawal. In the study, 144 subjects were man and 53 subjects were woman. There was no significant difference between survival rate of men and women (Table 1, *P*=0.438).

Table 1: Kaplan-Meier estimate of survival time among patient with acute myocardial infarction (AMI)

				Kaplan-Meier estimate		95% CI		
Gender	Event of AMI	Number of death	Number of censors (%)	Mean	SE	Lower	Upper	*P* value
Male	144	20	124 (86.1)	53.514	1.415	50.741	56.287	0.438
Female	53	7	46 (68.8)	54.377	2.088	50.285	58.470	
Total	197	27	170 (86.3)	53.746	1.177	51.439	56.054	

In this study, we obtained interpretable regression curves using different kernels, which estimated based on generalized cross-validation, and different kernel density estimation. Considering distribution type, cross-validation algorithm can determine optimal bandwidth value. According to the distribution type, such as Epanechnikov, Bi-quadratic, Rectangle and Tri-quadraticcan obtain different regression curves (Figure 1).

Each distribution has different curve as well as mathematical equation. As in Figure 1 is observed, the obtained mortality hazard curves using Bi-quadratic and Tri-quadratic kernels are huge smooth, but the mortality hazard curve based on Rectangle is very rough and jagged.

The hazard rate curve based on Epanechnikov kernel does not such defects. MISE values of each kernel in this study indicate that the obtained hazard rate of the first acute myocardial infarction under Epanechnikov kernel is more precise than the other kernels (Table 2).

In addition, Table 1 shows the MISE measurements based on different kernels and bandwidth selection methods. In which, the study showed that Epanechnikov kernel using bandwidth selection method of local has minimum MISE to estimate hazard rate in acute myocardial infarction (Table 2).

In this study, we obtained the death hazard rate in 10 months after the first acute myocardial infarction using Epanechnikov, Bi-quadratic, Tri-quadratic and Rectangle kernels, were 0.0031, 0.0028, 0.0036 and 0.0027, respectively. Also, the hazard rate in 30 months after the first acute myocardial infarction using Epanechnikov, Bi-quadratic, Tri-quadratic and Rectangle kernels were 0.0012, 0.0011, 0.0013 and 0.001, respectively. Figure 1 shows that Epanechnikov kernel estimates death hazard rates through 30th to 40th month are less than the others kernels.

Discussion

An effective procedure based on non-parametric method, kernel smoothing, is proposed well in this work for hazard rate estimation of patients with acute myocardial infarction when exist singly censored data. In the study, we converted the cross-sectional design to a hypothetical cohort design by Lexis method for estimate of death hazard rate.

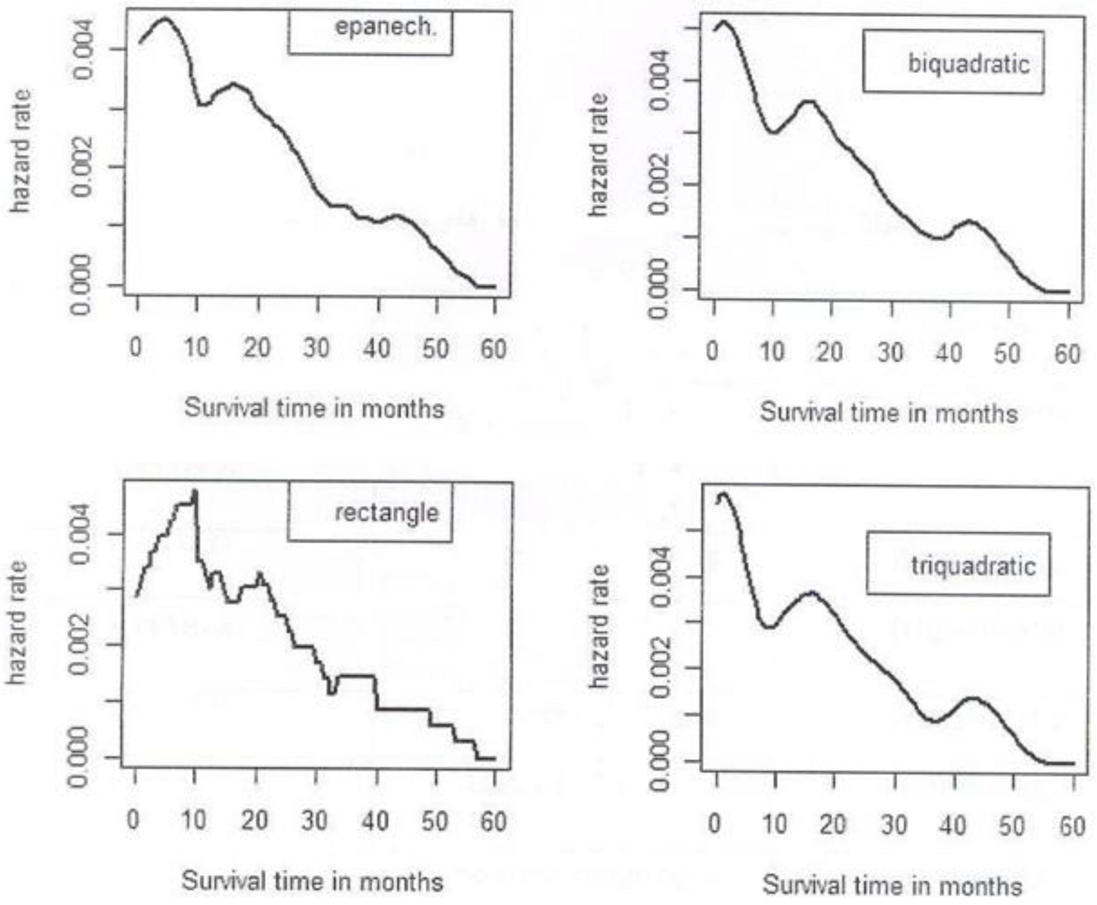

Figure 1: Smoothed plot hazard rate across the time to event among patient with acute myocardial infarction (AMI) for Epanechnikov, Biquadratic, Rectangle and Triquadratickernel and bandwidth (*h*=2)

Table 2: Comparing between mean integrated square error (MISE) indices based on different kernel

Kernels	Bandwidth selection method	MISE
Epanechnikov	Local	1.011×10^{-11}
$\frac{3}{4}(1-u^2)I(\lvert u\rvert\le1)^a$	K nearest neighbors	8.535×10^{-9}
Biquadratic	Local	1.112×10^{-8}
$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert\le1)$	K nearest neighbors	1.021×10^{-4}
Triquadratic	Local	1.231×10^{-8}
$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert\le1)$	K nearest neighbors	1.192×10^{-4}
Rectangle	Local	2.231×10^{-8}
$\frac{1}{2}I(\lvert u\rvert\le1)$	K nearest neighbors	7.759×10^{-5}

[a] $-1\le u=(t-t_i)/b\le1$ and *I* is an identity matrix

If bandwidth has a low value, the curve is unreadable and the regression curve is so roughly. In the event of increasing bandwidth value, the distribution has more readable and smooth. However, choosing inappropriate high bandwidth value, will be hidden the variance of population. Contrast to high-level value of bandwidth, the small value can produce curse of dimensionality[15]. Therefore, choosing the bandwidth value is too crucial. We measured performance of kernel by MISE like Oden et al. study[16], which compared the distribution of cystatin C for kidney disease. They used a smoothing bandwidth of 1.5 times of the standard deviation. In addition, Roger et al[17] used kernel smoothing in order to compare post processing techniques for perfusion defects in patients having myocardial infarction. They reported that bandwidth of 200 was equivalent to bandwidth of 150 under Epanechnikov kernel method. In addition, Jonset et al[18] used kernel smoothing method for study of incidence and prognostic factors on fibrillation in patients with acute myocardial infarction and left ventricular systolic dysfunction. They found a bandwidth 4.8 months using MISE. In the previous study, was used kernel-smoothing plot, but in our study, four kernels used for smoothing hazard rate curves. The estimate hazard rates based on the four kernels in the first 10 month were different. In other word, hazard rate estimation from starting time to 5th month under Epanechnikov and Rectangle kernels were increasing but under the other kernels were decreasing. Therefore, as we mentioned before, a kernel function must be used with the lowest MISE, i.e., Epanechnikov kernel, for estimating the hazard rate of myocardial infarction.

The proposed procedure for estimating the hazard rate is simpler than the parametric model. Kernel function is capable of creating optimal curves according to appropriate distribution types. In this study, curse dimensionality and kernel function was scrutinized to fix the problem. In this work, based on MISE values, the hazard rates estimation under Bi-quadratic, Tri-quadratic and Rectangle kernels are overestimate or underestimate.

Conclusion

The Epanechnikov kernel to obtain death hazard rate of patients with acute myocardial infarction has minimum mean integrated squared error compared to the other kernels. In addition, the mortality hazard rate of acute myocardial infarction in the study was low.

Acknowledgments

This article is a part of MSc thesis supported by Hamadan University of Medical Sciences. We thank all trial site investigators for their dedication towards data collection, registry and patient's follow-up.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Funding

Research Center for Health Sciences in Hamadan University of Medical Sciences supported this research.

References

1. Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, Ali F, Sholley C, Worrall C, et al. Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with Rosiglitazone or Pioglitazone. *JAMA*. 2010;304(4):411-418.
2. Soltanian AR, Mahjub H, Goodarzi S, Nabipour I, Jamali M. Five years survival rate in patients with myocardial infarction in Bushehr. *Scientific Journal of Hamadan University of Medical Sciences*. 2009;16(3):33-37.
3. Klein JP,Moeschberger ML. *Survival analysis: Techniques for censored and truncated data*.2th ed.New York: Springer; 2006.
4. Scheike TH. A generalized additive regression model for survival times. *The Annals of Statistics*. 2001;29(5):1344-1360.
5. Kayri M, Zırhlıoglu G. Kernel smoothing function and choosing banwidth for non-parametric regression methods. *Ozean Journal of Applied Sciences*. 2009;2(1):49-54.
6. Tong LI, Su C. A non-parametric method for experimental analysis with censored data. *International Journal of Quality & Reliability Management*. 1996;14(5):456-463.
7. Nelson W, Hahn G. Linear estimation of a regression relationship from censored data, part I – simple methods and their application. *Technometrics*. 1972;14:247-269.
8. Muller M. Estimation and testing in generalized partial linear model-A comparative study. *Statistics and Computing*. 2001;11:299-309.
9. Verwij P, Van-Houwelingen J. Cross validation in survival analysis. *Statistics in Medicine*. 1993;12:2305-2314.

10. Duong T, Hazelton M. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*. 2005;32:485-506.

11. Mahnken AH, Koos R, Katoh M, Spuentrup M, Busch P, Wildberger JE, et al. Sixteen-slice spiral CT versus MR imaging for the assessment of left ventricular function in acute myocardial infarction. *Eur-Radiol*. 2005;15:714-720.

12. Rice J, Silverman B. Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *Journal of the Roya Statistica Society, Series B*. 1991;53(1):233-243.

13. Doksum K, Peterson D, Samarov A. On variable bandwidth selection in local polynomial regression. *Journal of the Royal Statistical Society, Series B*. 2000;62:432-448.

14. Jones M, Marron J, Sheather S. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*. 1996;11(3):337-381.

15. Härdle W, Müler M, Sperlich S, Werwatz A. *Non-parametric and semi-parametric models*. Heidelberg: Springer; 2004.

16. Odden MC, Tager IB, Gansevoort RT, Bakker SJL, Katz R, Friad LF, et al. Age and cystatin C in healthy adults: a collaborative study. *Nephrol Dial Transplant*. 2010;25:463-469.

17. Rogers IS, Cury RC, Blankstein R, Shapiro MD, Nieman K, Hoffmann U, et al. Comparison of post-processing techniques for the detection of perfusion defects by cardiac computed tomography in patients presenting with acute ST-segment elevation myocardial infarction. *Journal of Cardiovascular Computed Tomography*. 2010;4:258-266.

18. Jons C, Jacobsen UG, Joergensen RM, Olsen NT, Dixen U, Johannessen A, et al. The incidence and prognostic significance of new-onset atrial fibrillation in patients with acute myocardial infarction and left ventricular systolic dysfunction: A CARISMA sub study. *Heart Rhythm*. 2011;8(3):342-348.