STATISTICS IN TRANSITION new series, December 2023
Vol. 24, No. 5, pp. 109–127, https://doi.org/10.59170/stattrans-2023-066
Received – 08.03.2022; accepted – 13.02.2023

A study of a survival data using kernel estimates of hazard rate and aging intensity functions

Magdalena Szymkowiak[1], **Anasuya Roychowdhury**[2], **Satya Kr. Misra**[3], **Rajib Lochan Giri**[4], **Subarna Bhattacharjee**[5]

Abstract

Analyzing survival (life-testing) data and drawing inferences about them is a part of engineering and health sciences. So far, various statistical tools, e.g., survival (reliability) function (sf), probability density function (pdf), and hazard rate function (HR) were available among decision-making scientists to handle time-to-event data (complete or censored). But because functions (pdf) estimators were interval (window) based, they mostly gave qualitative ideas having pictorial representation resembling step functions, ordinate remain constant when abscissa vary over an interval, thereby giving incomplete information. However, it can be sorted out with the use of kernel estimates of the above mentioned functions, resulting into smooth estimators. Moreover, the metric based on aging intensity function (AI) gives an alternative way of studying lifetime or clinical datasets as it is a quantitative measure (not interval-based), thereby depicting a broader view of a given data. In our study, we primarily focus on AI and HR functions estimated using four different kernels. We apply them to a case study of patients with primary malignant tumors of sternum (cf. Daniel and Cross, 2014) with the right-censored data. Our result shows that kernel estimates of HR and AI functions for patients with high grade tumor (HGT) are higher than for patients with low grade tumor (LGT), as expected. Thus, the study opens up a new direction for applying AI and HR functions in health sciences and engineering studies.

Key words: hazard rate, aging intensity function, kernels, survival analysis, cancer statistics, clinical datasets.

1. Introduction

The comparison of two different products for two different brands has high importance in many fields including but not limited to reliability theory, biological sciences, and forensic sciences. In survival analysis, the remaining lifetimes of a component at different times of its life span needs to be compared to determine how the component is aging with time. Various stochastic orders between random variables, viz., classical stochastic (st) order, hazard rate (hr) order, likelihood ratio (lr) order, aging intensity (ai) order, etc. have been studied in the literature (cf. Shaked and Shanthikumar (2007)). In this regard, our article

[1]Institute of Automatic Control and Robotics, Poznan University of Technology, Poznań, Poland. ORCID: https://orcid.org/0000-0002-5066-8629.
[2]Biochemistry and Cell Biology Laboratory, School of Basic Sciences, IIT Bhubaneswar, India.
[3]Department of Mathematics, KIIT University, Bhubaneswar-751024, Odisha, India.
[4]Department of Mathematics, Ravenshaw University, Cuttack-753003, Odisha, India.
[5]Corresponding author :Department of Mathematics, Ravenshaw University, Cuttack-753003, Odisha, India. E-mail: subarna.bhatt@gmail.com. ORCID: https://orcid.org/0000-0002-3697-4216.

analyzes a case study of cancer data cited in Daniel and Cross (2014).

Emmerson and Brown (2021), Rosen et al. (2020) and others discuss the use of Kaplan-Meier survival analysis in evaluating the efficiency of onco-drugs in randomised controlled trials (RCTs). Further, Nayak et al. (2021) apply Kaplan-Meier analysis to assess a potential oncoprotein ATAD2 as a prognostic marker for stomach cancer. Inquisitive readers can further explore on Kaplan-Meier Plotter (KMP, http://kmplot.com/analysis/), which is a web-based meta-analysis biomarker validation tool used in medical research and also used by Nayak et al. (2021) in the analysis of stomach cancer.

The treatment management of cancer starts with the determination of stages (how big the tumor is and how far it is spread) and grades (how fast it grows) of the tumor. A higher stage and/or grade of tumor may grow and spread rapidly and may require immediate treatment action. For example, high-grade tumors (HGT) are more aggressive than the low grade tumors (LGT). Therefore, severity of the disease and treatment management could be more complicated for HGT. Moreover, every cancer sub-type is unique.

A statistical account on patients with the same cancer sub-type often help the public health community to estimate the prognosis better. Therefore, to improve the treatment spectrum of the complex disease like cancer, there is an immense importance of statistical analysis using incidence and survival data from the vast range of original datasets that are generally accumulated in authentic and authoritative public repository databases. However, to dig out meaningful information from those datasets, the statistical analytical tools are required to be robust and bias-free. So, we make an attempt to apply it for a particular data so as to follow its aging pattern.

Although aging intensity function AI, defined as the ratio of the instantaneous HR to its average, has already gathered some familiarity in recent literature of statistics, to the best of our knowledge its application in analysis of survival data is sparse (cf. Misra and Bhattacharjee (2018)). Here, in this study, we take up a strategy of implementing kernels estimates of hazard rate (HR) and aging intensity (AI) functions for the survival analysis of a particular censored data of patients suffering from malignant tumors of sternum (cf. Daniel and Cross (2014)). Censoring of data arises as lifetimes occur only within certain intervals. Censored data are useful when their survival time is truncated at a certain point of time.

The rest of our article is organized as follows. Portfolio of HR and the AI functions are presented in Section 2. A brief survey on kernels used in estimation follows. In Section 3, we cite a dataset which has been taken up in our present study and discuss the results so obtained. The significance of the paper is established in concluding remarks of Section 4. In Section 5, the Appendix speaks about the detailed calculations of this work connected with four presented kernels and gives a short comparison between them based on goodness-of-fit test. The notation *r.v.* is used in place of random variable.

2. Portfolio of HR and AI functions

The keywords of this work are hazard rate HR and aging intensity AI functions, which are being here used along with their kernel estimates for application in medical statistics, especially cancer analysis and in health sciences. To this end, we give a brief note of the

aforementioned concepts in the ensuing discussions.

2.1. Hazard rate interpretation

Let T be a random variable representing any lifetime of a system with a well-defined statistical distribution having probability density function (pdf) denoted by f, and survival function (or popularly known as reliability function) (sf) denoted by $\overline{F}$. Sf of a *r.v.* T at time t is given by $\overline{F}_T(t) = P(T > t)$ which represents the probability of surviving over time t. The cumulative distribution function (cdf) is $F_T(t) = 1 - \overline{F}_T(t)$, $t > 0$. Hazard rate (HR) function, known also in survival analysis as failure rate, is defined for a continuous random variable T as

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} = \frac{f_T(t)}{\overline{F}_T(t)}, \text{ where density is defined.} \quad (2.1)$$

If the hazard rate is high, then it implies that the corresponding unit with life-time T is aging faster. Other functions used in the study of aging analysis are reversed hazard rate, mean residual life (cf. Nanda et al. (2010)), reversed mean residual life (cf. Nanda et al. (2006), Shaked and Shanthikumar (2007)) functions, etc.

2.2. Significance of aging intensity function

Jiang et al. (2003) classifies a unimodal hazard rate as quasi-decreasing (anti-aging), quasi-increasing (aging) or quasi-constant (non-aging) depending on whether its mode t_c, (called critical time) is small, large or moderate, respectively. A distribution is classified as quasi-constant if the hazard rate curve is relatively flat. They claim that the representation of aging of a system by hazard rate is qualitative. Thereby, they introduced a notion, called aging intensity (AI), to quantitatively evaluate the aging property of a system. AI of a random variable T, denoted by $L_T(t)$, is defined as the ratio of the instantaneous hazard rate $h_T(t)$ given by (2.1) to the hazard rate average $\frac{1}{t}H_T(t)$, where $H_T(t) = \int_0^t h_T(u)du$ is the cumulative hazard rate, i.e.,

$$L_T(t) = \frac{h_T(t)}{\frac{1}{t}H_T(t)}. \quad (2.2)$$

It is easy to see that (2.2) can be also presented as

$$L_T(t) = \frac{-t f_T(t)}{\overline{F}_T(t) \ln \overline{F}_T(t)}, \text{ for } t > 0. \quad (2.3)$$

The concept of aging intensity (AI) function is found in Nanda et al. (2007), Bhattacharjee et al. (2013b), Bhattacharjee et al. (2022) for quantitative study of the aging process of a system. Bhattacharjee et al. (2013a), Misra and Bhattacharjee (2016), Swain et al. (2021) illustrate the properties of AI function and its usage on a complete dataset. Misra and Bhattacharjee (2018) gave a comparative role of HR, and AI functions on analysis of data (censored). Szymkowiak (2018, 2019, 2020) gave a detailed literature on AI function. Giri et al. (2023) studied HR, AI functions of different Weibull models and simulated data from Weibull distributions.

For a complete data, the empirical estimates of pdf, cdf and AI functions, denoted by $\hat{f}(t), \hat{F}(t)$ and $\hat{L}(t)$, respectively, are given by (cf. Bhattacharjee et al. (2013a), Szymkowiak (2018))

$$\hat{f}_{emp}(t) = \frac{N_s(t_j) - N_s(t_j + \Delta t_j)}{N \Delta t_j},$$

$$\hat{F}_{emp}(t) = \frac{N - N_s(t_j)}{N}, \tag{2.4}$$

$$\hat{L}_{emp}(t) = -\frac{t\hat{f}(t)}{\left[1 - \hat{F}(t)\right] \ln\left[1 - \hat{F}(t)\right]} = -t\left\{\frac{N_s(t_j) - N_s(t_j + \Delta t_j)}{\Delta t_j N_s(t_j) \ln \frac{N_s(t_j)}{N}}\right\}, \tag{2.5}$$

for $t_j \leq t \leq t_j + \Delta t_j$. Here, $N, N_s(t_j)$ and $N_s(t_j + \Delta t_j)$ refer to the total number of survivors at $t = 0$ (beginning of the life-testing), $t = t_j$ and $t = t_j + \Delta t_j$, respectively.

The above defined estimates $\hat{f}_{emp}(t)$ and $\hat{F}_{emp}(t)$ for HGT and LGT patients are presented in Figure 5.13 and Figure 5.29, respectively, as the blue step functions. One can justify the fact that AI function is a quantitative measure of aging as the factor t is involved in (2.5), which gives rise to a smooth (or not window) estimator.

Refer Klein and Moeschberger (2003) for detailed analysis on estimator of cumulative hazard rate $H(t) = \int_0^t h(u)du$ and hazard rate $h(t)$. $\hat{H}(t)$, is the estimator given by Nelson-Aalen for $H(t)$ and the slope of this estimator gives a rough estimate of the $\hat{h}(t)$. Clearly, the estimator of AI function (2.2) is given by $\hat{L}(t) = \frac{\hat{h}(t)}{\frac{1}{t}\hat{H}(t)}, t > 0$.

The following subsection gives a brief survey of kernels which helps us to give smooth (not window-based) estimators for functions used in statistics.

2.3. Kernels: a brief survey

If probability density function is unknown or difficult to obtain in parametric distributions, we can use kernel estimates of pdf and cdf functions for their applications in statistical inference. One can refer to DiNardo and Tobias (2001) to name a few. These problems are faced primarily by statisticians who are engaged in evaluating reliability. An important aspect of aforementioned kernel estimation is associated with selecting suitable kernel and the choice of its corresponding bandwidth (b). Readers can explore on some well-known literature (cf. Miladinovic (2008)) on ranking of seven crucial kernels on the basis of their (optimal) bandwidth.

The expressions of kernel estimates of pdf, cdf and sf are, respectively, given by the following definition (cf. Miladinovic (2008)).

Definition 2.1 *If $T_1, T_2, \ldots, T_n$ are i.i.d. random variables with the same $f_n(t)$, the kernel estimate of pdf is given by*

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - T_i}{b}\right), \tag{2.6}$$

where b is the bandwidth and $K(u)$ is a kernel smoothing function so chosen. The kernel

estimate of the cdf and survival (reliability) function (sf) are, respectively, given by

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t} K\left(\frac{u-T_i}{b}\right)du, \tag{2.7}$$

and $\widehat{\overline{F}}_n(t) = 1-\hat{F}_n(t)$.

We list four kernels as follows (cf. Bowman and Azzalini (1997), Miladinovic (2008)). Here, $I(A)$ refers to the fact that $I(x)=1$, if $x\in A$ and $I(x)=0$, if $x\notin A$, where $A\neq\emptyset$.

(i) Epanechnikov (EPA) kernel, $K(u)=\frac{3}{4}(1-u^2)I(|u|\leq 1)$

(ii) Normal (Gaussian) kernel, $K(u)=\frac{1}{\sqrt{2\pi}}e^{-0.5u^2}$

(iii) Triangle kernel, $K(u)=(1-|\,u\,|)I(|\,u\,|\leq 1)$

(iv) Box (Uniform) kernel, $K(u)=0.5I(|\,u\,|\leq 1)$

The kernels must satisfy a set of properties as given in the following remark.

Remark 2.1 *The kernels must satisfy the conditions* $\int_{-\infty}^{\infty}K(u)du=1$, $\int_{-\infty}^{\infty}uK(u)du=0$ *and* $\int_{-\infty}^{\infty}u^2K(u)du>0$.

The properties possessed by the kernels are partially the same as that of the kernel density estimates. Epanechnikov introduced the kernel after his name for density estimation in 1956. The bandwidth b plays a crucial role and is assigned a value in such a way that it minimizes mean-squared error or it helps in obtaining the required degree of smoothness. To obtain the aging intensity estimator, we use (2.6) and (2.7). These are available in *MATLAB* as `ksdensity` function.
As usual,

$$\hat{h}_n(t)=\frac{\hat{f}_n(t)}{1-\hat{F}_n(t)}=\frac{\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{b}\right)}{1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du}.$$

and by (2.3) aging intensity estimate is equal to

$$\hat{L}_n(t) \;=\; -\frac{t\,\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{b}\right)}{\left[1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du\right]\ln\left[1-\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{t}K\left(\frac{u-T_i}{b}\right)du\right]}.$$

Here, we apply only four kernels estimates of HR and AI: Box kernels, $\hat{h}_B$, $\hat{L}_B$, Epanechnikov (EPA) kernels, $\hat{h}_E$, $\hat{L}_E$, Normal kernels, $\hat{h}_N$, $\hat{L}_N$, and Triangle kernels, $\hat{h}_T$, $\hat{L}_T$, respectively.

3. Analysis of cancer data: results and discussion

We refer to data given in Martini et al. (1996) and Daniel and Cross (2014), displayed in Table 3.1. They noted primary malignant tumors of the sternum in patients with low-grade

Table 3.1: Data: Malignant Tumors of Sternum (cf. Daniel and Cross (2014))

Subject	Time t [month]	Vital Status	Tumor Grade	Subject	Time t [month]	Vital Status	Tumor Grade
1	29	*dod*	*LGT*	21	155	*ned*	*LGT*
2	129	*ned*	*LGT*	22	102	*dod*	*LGT*
3	79	*dod*	*LGT*	23	34	*ned*	*LGT*
4	138	*ned*	*LGT*	24	109	*ned*	*LGT*
5	21	*dod*	*LGT*	25	15	*dod*	*LGT*
6	95	*ned*	*LGT*	26	122	*ned*	*HGT*
7	137	*ned*	*LGT*	27	27	*dod*	*HGT*
8	6	*ned*	*LGT*	28	6	*dod*	*HGT*
9	212	*dod*	*LGT*	29	7	*dod*	*HGT*
10	11	*dod*	*LGT*	30	2	*dod*	*HGT*
11	15	*dod*	*LGT*	31	9	*dod*	*HGT*
12	337	*ned*	*LGT*	32	17	*dod*	*HGT*
13	82	*ned*	*LGT*	33	16	*dod*	*HGT*
14	33	*dod*	*LGT*	34	23	*dod*	*HGT*
15	75	*ned*	*LGT*	35	9	*dod*	*HGT*
16	109	*ned*	*LGT*	36	12	*dod*	*HGT*
17	26	*ned*	*LGT*	37	4	*dod*	*HGT*
18	117	*ned*	*LGT*	38	0	*dpo*	*HGT*
19	8	*ned*	*LGT*	39	3	*dod*	*HGT*
20	127	*ned*	*LGT*				

tumor *LGT* (25 patients) or high-grade tumor *HGT* (14 patients), respectively (source: data provided courtesy of Dr. Martini).

The notations used in Table 3.1 are depicted as *dod* for 'dead of disease' (treated as uncensored data); *ned* for 'no evidence of disease' (treated as censored data) and *dpo* for 'dead post operation' (treated as uncensored data). Throughout this paper, t is given in months. In this article we aim to study the aging phenomena among the disease groups

Table 3.2: Kernel estimates of *AI* for *HGT*

t	$\hat{L}_B(t)$	$\hat{L}_E(t)$	$\hat{L}_N(t)$	$\hat{L}_T(t)$
0	–	–	–	–
2	0.3306	0.3351	0.3660	0.3526
3	0.4413	0.4687	0.5132	0.4976
4	0.5328	0.5918	0.6392	0.6198
6	0.7542	0.7965	0.8350	0.8139
7	0.8995	0.8774	0.9082	0.8896
9	1.0286	1.0007	1.0137	1.0130
12	1.1289	1.1295	1.0942	1.1022
16	1.1802	1.1711	1.1230	1.1802
17	1.2682	1.1713	1.1257	1.1783
23	0.9968	1.1023	1.1513	1.1454
27	1.0251	1.0771	1.1057	1.1197

Table 3.3: Kernel estimates of *HR* for *HGT*

t	$\hat{h}_B(t)$	$\hat{h}_E(t)$	$\hat{h}_N(t)$	$\hat{h}_T(t)$
0	0.0318	0.0309	0.0305	0.0307
2	0.0383	0.0377	0.0395	0.0388
3	0.0398	0.0413	0.0441	0.0433
4	0.0415	0.0455	0.0486	0.0475
6	0.0504	0.0540	0.0570	0.0555
7	0.0585	0.0580	0.0608	0.0593
9	0.0663	0.0652	0.0673	0.0666
12	0.0744	0.0751	0.0738	0.0737
16	0.0816	0.0814	0.0783	0.0820
17	0.0890	0.0823	0.0791	0.0828
23	0.0734	0.0810	0.0843	0.0840
27	0.0771	0.0805	0.0828	0.0837

HGT and *LGT* implementing four kernels which are most commonly used in statistical estimation (available with *MATLAB* R2016a version). Our primary focus is on two different measures, one qualitative, i.e., *HR* function, and the other, quantitative , i.e., *AI* function. Here, it is worth mentioning that *HR* and *AI* bear two different dimensions, the former's unit is 1 per unit of time (for considered data $\left[\frac{1}{month}\right]$) and the latter is dimensionless. Their role in the study of system aging behaviour has been discussed in introduction section. However, to the best of our knowledge, not much work has been done where *HR* and *AI* functions

Table 3.4: Kernel estimates of *AI* for *LGT*

t	$\hat{L}_B(t)$	$\hat{L}_E(t)$	$\hat{L}_N(t)$	$\hat{L}_T(t)$
6	0.0430	0.0585	0.0623	0.0633
11	0.0765	0.1058	0.1119	0.1148
15	0.1020	0.1427	0.1501	0.1545
21	0.1383	0.1962	0.2049	0.2105
29	0.3133	0.2642	0.2739	0.2794
33	0.3448	0.2967	0.3065	0.3111
79	0.6124	0.6036	0.6087	0.5948
102	0.7100	0.7157	0.7177	0.6960
212	0.8208	0.9222	0.9834	1.0478

Table 3.5: Kernel estimates of *HR* for *LGT*

t	$\hat{h}_B(t)$	$\hat{h}_E(t)$	$\hat{h}_N(t)$	$\hat{h}_T(t)$
6	0.0013	0.0017	0.0018	0.0018
11	0.0013	0.0018	0.0019	0.0019
15	0.0013	0.0018	0.0019	0.0020
21	0.0013	0.0019	0.0020	0.0020
29	0.0023	0.0020	0.0021	0.0021
33	0.0023	0.0020	0.0021	0.0022
79	0.0026	0.0025	0.0026	0.0025
102	0.0027	0.0028	0.0028	0.0027
212	0.0029	0.0032	0.0035	0.0036

are implemented in assessing the survival of cancer patients. We use four different kernel estimators to address the issue. We find that all considered kernels exquisitely corroborated each other to infer the data.

Table 3.2 and Table 3.3, respectively, depict four different kernel estimates of *AI* and

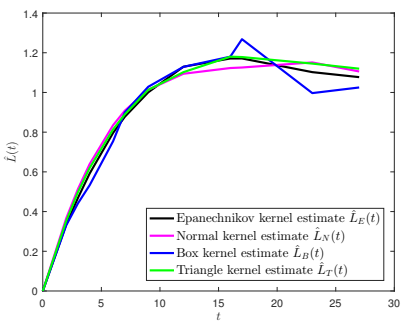

Figure 3.1: *AI* for *HGT*

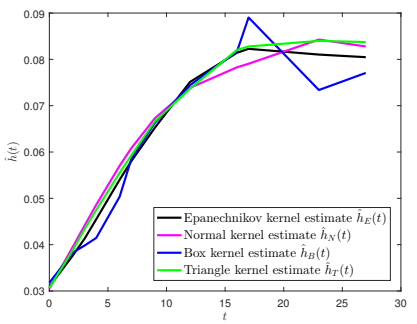

Figure 3.2: *HR* for *HGT*

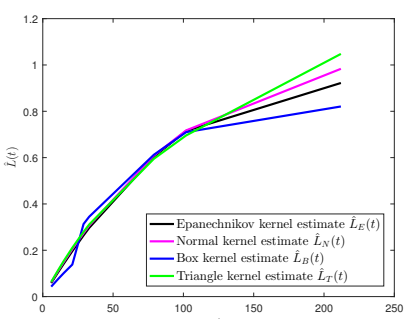

Figure 3.3: *AI* for *LGT*

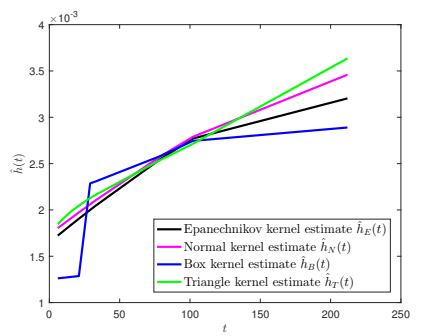

Figure 3.4: *HR* for *LGT*

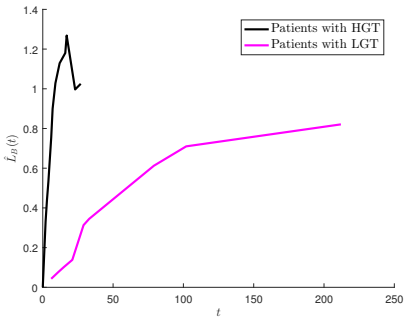

Figure 3.5: *AI*, Box kernel

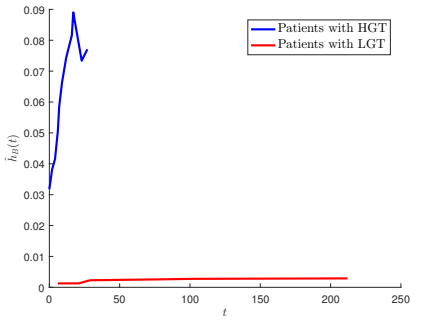

Figure 3.6: *HR*, Box kernel

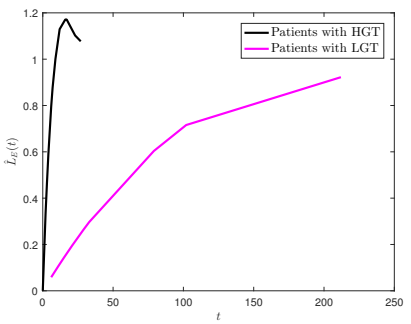

Figure 3.7: *AI*, EPA kernel

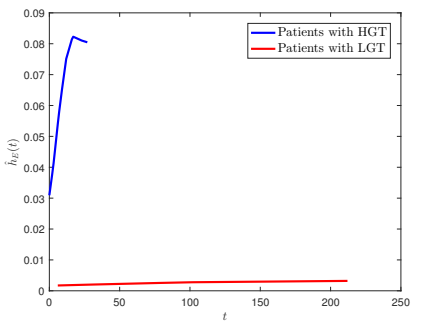

Figure 3.8: *HR*, EPA kernel

HR for patients with *HGT*. The respective four kernel estimates of *AI* and *HR* for patients with *LGT* are given in Table 3.4 and Table 3.5. Note, that for $t = 0$, *AI* estimates are not defined. Figures 3.1–3.12 are plotted for the purpose of drawing inference about the given data (with reference to Table 3.1) from Tables 3.2–3.5. In Figures 3.1–3.4, we have kernel estimates of *AI* for patients with *HGT*, kernel estimates of *HR* for patients with *HGT*, kernel estimates of *AI* for patients with *LGT*, and kernel estimates of *HR* for patients with *LGT*, respectively, which reveal the robustness of the kernels used to evaluate the values of instantaneous hazard rate (given by *HR*) and aging intensity (given by *AI*). While *AI* function is implemented in all four kernel estimators, we find all of them exquisitely corroborates with each other, both for *HGT* (Fig 3.1) and *LGT* (Fig. 3.3) datasets. The same observation is also found while *HR* function highlighting the qualitative aspect of aging is used for *HGT* (Fig 3.2) and *LGT* patients (Fig 3.4). This clearly indicates that both *AI* and *HR* functions could be efficiently implemented to the censored cancer data analysis. Figures 3.5-3.12 represent the differences in the impact of two different tumor grades on patients using *HR* and *AI* functions with respect to each of the four kernels. The sequel of the figures can be found in caption of each figure.

First, we implement *AI* (Fig 3.5) and *HR* (Fig 3.6) in Box kernel estimator for censored *HGT* and *LGT* datasets. As expected, *HGT* shows higher *AI* and *HR*, compared to *LGT*. However, due to the differential nature of these two functions (quantitative vs qualitative)

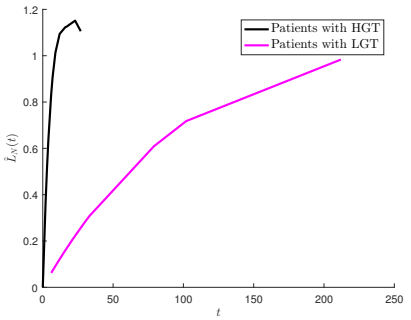

Figure 3.9: *AI*, Normal kernel

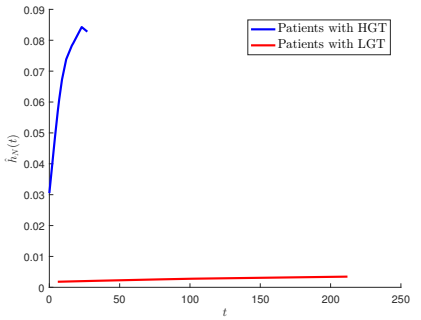

Figure 3.10: *HR*, Normal kernel

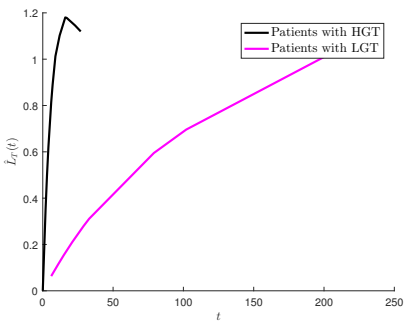

Figure 3.11: *AI*, Triangle kernel

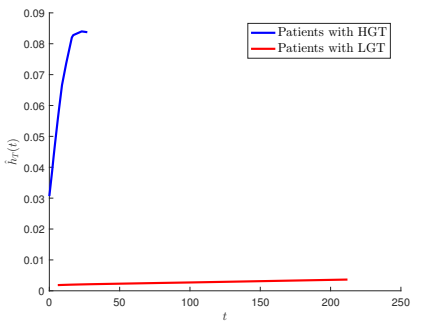

Figure 3.12: *HR*, Triangle kernel

the pattern of curve for *LGT* differs. Since, the representation of aging of a system by hazard rate is qualitative, the curve is relatively flat for *HR* compared to the *AI*. On the other hand, since *AI* quantitatively evaluates the aging property of a system, for *LGT* datasets, lifetime distribution is better represented by *AI* than by *HR* curve. This is worth mentioning here that since *HGT* patients do not survive for longer time period, qualitative (HR) and quantitative (AI) evaluation does not affect much.

Interestingly, similar observation is also found for other three estimators like Epanechnikov (EPA) kernel (Fig 3.7 and 3.8), normal kernel (Fig 3.9 and 3.10) and triangle kernel (Fig 3.11 and 3.12) indicating the importance of implementing *AI* and *HR* functions to the cancer data irrespective of the kernel estimator used. The calculations pertaining to each of the four kernels are given in detail in the Appendix section.

4. Conclusions

The concluding remarks of this paper are compiled as follows.

(i) Cancer statistics is an important domain of cancer treatment management. In this article, we focus on analysing robust statistical methods that can deal with cancer survival data effectively and it can be applied for any survival or life testing data.

(ii) To the best of our knowledge, there are no studies where *HR* and *AI* functions are applied in assessing the survival data from cancer patients. Reports from this work indicate that implementation of *HR* and *AI* functions in human diseases is promising. Therefore, we illustrate the same with detailed analysis using available censored data on cancer-survivals (Martini et al., 1996).

(iii) We use four different kernel estimators to apply *HR* and *AI* functions. Our analysis shows that *HR* and *AI* for patients with HGT are higher than for patients with LGT, as expected, showing a lower survival of HGT patients.

(iv) Since representation of aging in a system by *AI* is more quantitative, *AI* curves are able to provide more information than the *HR* (qualitative) curves (as depicted in their flattened nature) (see LGT curves of Fig 3.5 versus Fig 3.6; Fig 3.7 versus Fig 3.8; Fig 3.9 versus Fig 3.10 and Fig 3.11 versus Fig 3.12). The pattern is particularly prominent for all our LGT curves as we do not have data beyond 50 months for HGT patients (by that time all HGT patients die due to the severity of the disease). On the contrary, as LGT patients survived longer periods of time, we have data until 200 months (approx.). This allows us to visualize the full spectrum of the phenomena of *AI* (with more information) and appreciate the quantitative nature of the function as opposed to less informative flattened pattern for qualitative *HR* function.

(v) Our study strongly indicates that both *AI* and *HR* functions could be efficiently implemented to estimate the survival analysis for cancer patients. We believe, this new avenue of applying *AI* and *HR* functions will be adopted by researchers for implementation in any problem of health sciences or engineering studies.

5. Appendix

In this article, we intend to place the theme directly to the readers and as such we keep at bay the other statistical calculations for discussion in the Appendix section. Here, we give the details of the work done with reference to the aging metrics viz., cdf, pdf, *HR* and *AI* functions. First, we survey HGT patients followed by LGT patients.

5.1. HGT patients

First, we survey HGT patients.

5.1.1 HGT: Box kernel

For HGT patients, Box kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a) is used to determine function estimators. To be precise, we state that the estimates of cumulative distribution function (cdf), probability density function (pdf), hazard rate function (HR) and aging intensity function (AI) for patients with HGT using Box kernel are obtained. Accordingly, we receive plots for the following functions as mentioned here:

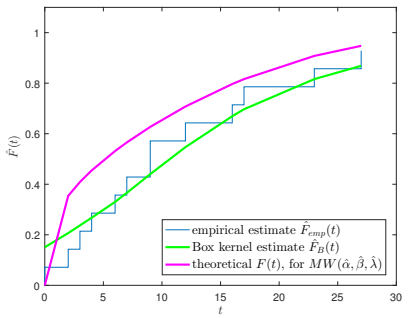

Figure 5.13: cdf for HGT, Box kernel

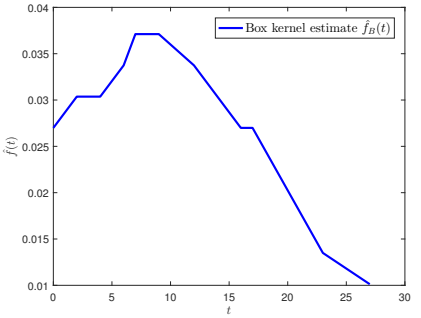

Figure 5.14: pdf for HGT, Box kernel

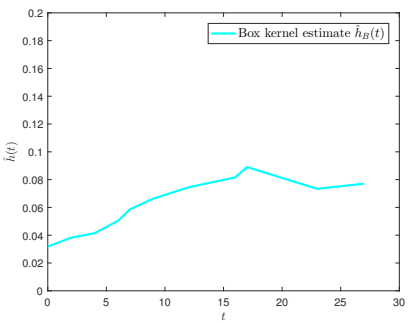

Figure 5.15: HR for HGT, Box kernel

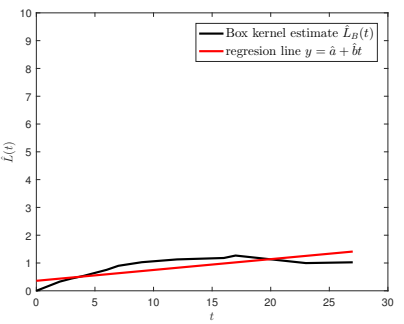

Figure 5.16: AI for HGT, Box kernel

(i) Figure 5.13, empirical cdf (blue step function) and Box kernel estimate of cdf (green function),

(ii) Box kernel estimate of pdf (Figure 5.14),

(iii) Box kernel estimate of HR (Figure 5.15),

(iv) Box kernel estimate of AI (Figure 5.16).

The function $\hat{L}(t)$ presented in Figure 5.16 is seen to oscillate around the linear function $y = a + bt$. So, the AI estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI, see, e.g., Szymkowiak (2020)) are $\hat{\alpha} = \hat{a} = 0.3598$, $\hat{\beta} = \hat{b} = 0.0390$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.3150$. Here, in this section, to obtain the desired value of $\hat{\lambda}$, we make use of the estimator

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i^{\hat{\alpha}} \exp\left(\hat{\beta} T_i\right)}, \tag{5.8}$$

where n is a sample size. The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters received by Box kernel estimates is shown in Figure 5.13 (magenta function).

5.1.2 *HGT*: Epanechnikov (EPA) kernel

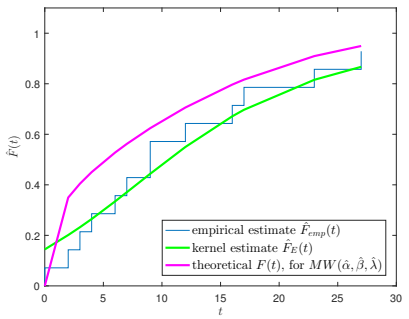

Figure 5.17: cdf for HGT, EPA kernel

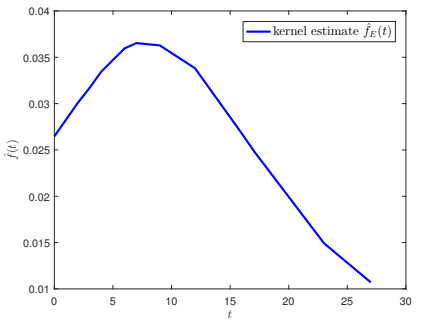

Figure 5.18: pdf for HGT, EPA kernel

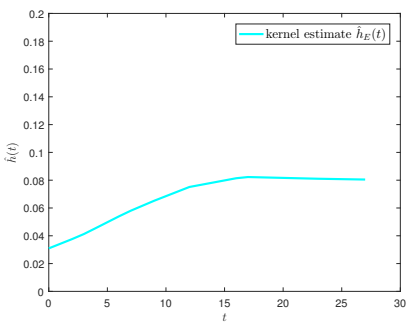

Figure 5.19: HR for HGT, EPA kernel

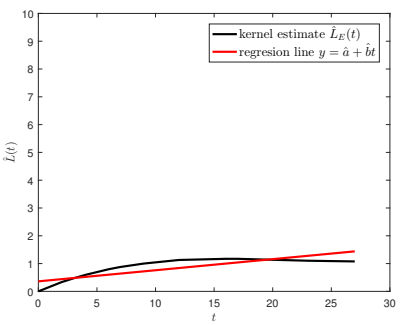

Figure 5.20: AI for HGT, EPA kernel

Next, for HGT patients, we determine the estimates of cdf, pdf, HR and AI by Epanechnikov (EPA) kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a). The corresponding plots are listed:

(i) Figure 5.17, empirical cdf (blue step function) and Epanechnikov kernel estimate of cdf (green function),

(ii) Epanechnikov kernel estimate of pdf (Figure 5.18),

(iii) Epanechnikov kernel estimate of HR (Figure 5.19),

(iv) Epanechnikov kernel estimate of AI (Figure 5.20).

The function $\hat{L}(t)$ in Figure 5.20 oscillates around the linear function $y = a + bt$. Then, the AI estimators of corresponding parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) are $\hat{\alpha} = \hat{a} = 0.3606$, $\hat{\beta} = \hat{b} = 0.0400$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.3093$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters obtained by Epanechnikov kernel estimates is shown in Figure 5.17 (magenta function).

5.1.3 *HGT*: Normal kernel

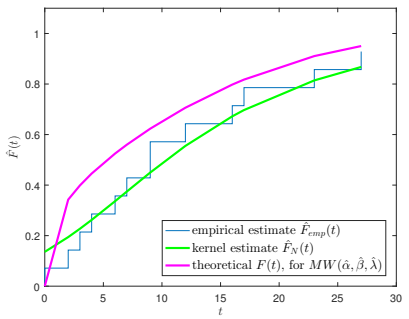

Figure 5.21: *cdf* for *HGT*, Normal kernel

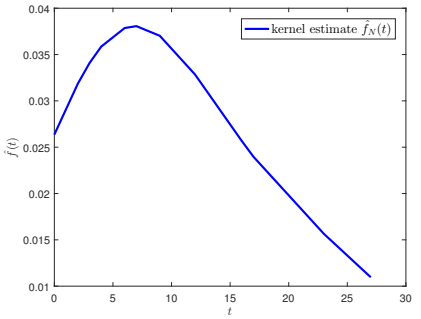

Figure 5.22: *pdf* for *HGT*, Normal kernel

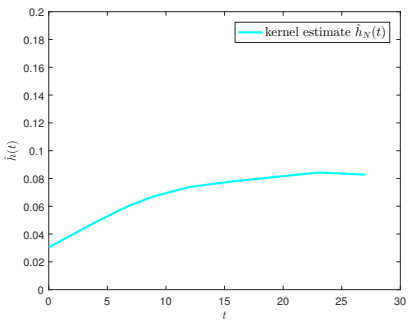

Figure 5.23: *HR* for *HGT*, Normal kernel

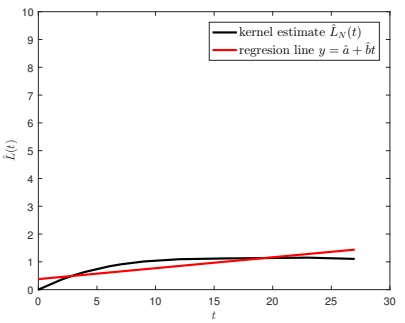

Figure 5.24: *AI* for *HGT*, Normal kernel

For *HGT* patients, Normal kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a) is used to find the estimators of *cdf*, *pdf*, *HR* and *AI*. The corresponding plots are listed:

(i) Figure 5.21, empirical *cdf* (blue step function) and Normal kernel estimate of *cdf* (green function),

(ii) Normal kernel estimate of *pdf* (Figure 5.22),

(iii) Normal kernel estimate of *HR* (Figure 5.23),

(iv) Normal kernel estimate of *AI* (Figure 5.24).

The function $\hat{L}(t)$ presented in Figure 5.24 oscillates around $y = a + bt$. Therefore, we can receive the *AI* estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear *AI*) $\hat{\alpha} = \hat{a} = 0.3789$, $\hat{\beta} = \hat{b} = 0.0393$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.2982$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters determined using Normal kernel estimates is shown in Figure 5.21 (magenta function).

5.1.4 *HGT*: Triangle kernel

Using Triangle kernel with bandwidth $b = 6.1113$ (proposed by *MATLAB* R2016a), we receive the estimators of cdf, pdf, HR and AI for patients with HGT. One can refer to the associated plots as given:

(i) Figure 5.25, empirical cdf (blue step function) and Triangle kernel estimate of cdf (green function),

(ii) Triangle kernel estimate of pdf (Figure 5.26),

(iii) Triangle kernel estimate of HR (Figure 5.27),

(iv) Triangle kernel estimate of AI (Figure 5.28).

The function $\hat{L}(t)$ presented in Figure 5.28 can be considered to oscillate around the linear function $y = a + bt$. So, we can determine the AI estimators of parameters of the Modified Weibull distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) as $\hat{\alpha} = \hat{a} = 0.3672$, $\hat{\beta} = \hat{b} = 0.0408$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 0.2999$ (using (5.8)). The theoretical $F(t)$ for $MW(\hat{\alpha},\hat{\beta},\hat{\lambda})$ with parameters received by Triangle kernel estimates is shown in Figure 5.25 (magenta function).

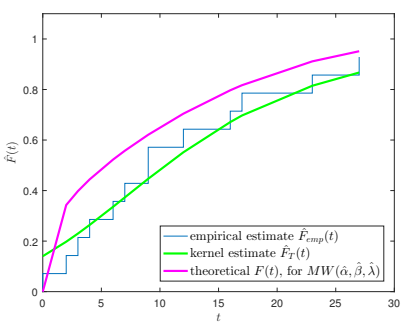

Figure 5.25: cdf for HGT, Triangle kernel

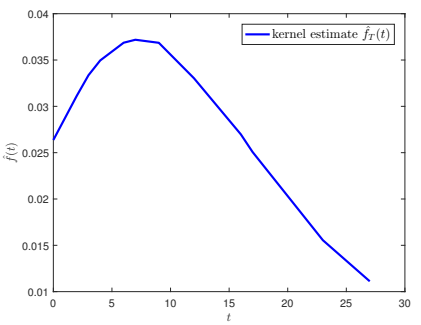

Figure 5.26: pdf for HGT, Triangle kernel

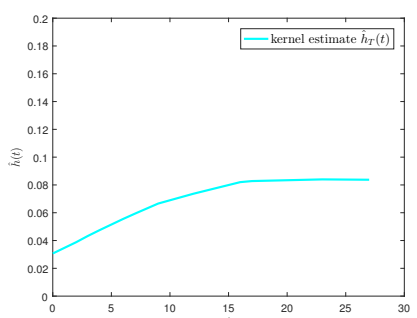

Figure 5.27: HR for HGT, Triangle kernel

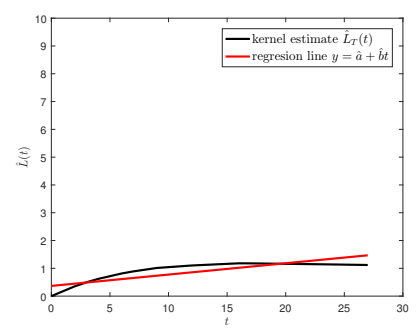

Figure 5.28: AI for HGT, Triangle kernel

5.1.5 Summary for HGT patients

We now summarize our analysis of data for HGT patients by four considered kernels. For each kernel and their AI estimators of parameters of the Modified Weibull distribution, using right type II censored data and Kolmogorov-Smirnov goodness-of-fit test (cf. Agostino and Stephens (1986)) we verify the hypothesis that the data really follow this distribution (with linear AI).

(i) Box kernel: statistics $D^* = 1.1216$ and p-value higher than 0.15

(ii) Epanechnikov kernel: statistics $D^* = 1.1051$ and p-value higher than 0.15

(iii) Normal kernel: statistics $D^* = 1.0787$ and p-value higher than 0.15

(iv) Triangle kernel: statistics $D^* = 1.0795$ and p-value higher than 0.15.

One can note that

$$D^* = \sqrt{n}D + \frac{0.24}{\sqrt{n}}$$

where n is a sample size and

$$D = \max_{1 \le i \le l} \left\{ \frac{i}{n} - F(i), F(i) - \frac{i-1}{n} \right\}.$$

Here, $F(t)$ is the theoretical cdf and l is the number of uncensored data. It means that at the significance level $\alpha < 0.15$, for all considered kernels, we do not reject the hypothesis that data follow the respective Modified Weibull distribution. Moreover, (although the differences are not large) we can notice that the value of statistics D^* is the smallest for Normal kernel. Therefore, we can claim that this kernel function is the best to use in our data analysis and so the Modified Weibull distribution $MW(\alpha, \beta, \lambda)$ with parameters $\hat{\alpha} = 0.3789$, $\hat{\beta} = 0.0393$ and $\hat{\lambda} = 0.2982$ fits the analyzed data best.

5.2. LGT: Normal kernel

For low grade tumor (LGT), we analyze the whole data from Table 3.1 using only Normal kernel with bandwidth $b = 109.2138$ (proposed by *MATLAB* R2016a) to get the function estimators. For patients with LGT we receive the associated plots as:

(i) Figure 5.29, empirical cdf (blue step function) and Normal kernel estimate of cdf (green function),

(ii) Normal kernel estimate of pdf (Figure 5.30),

(iii) Normal kernel estimate of HR (Figure 5.31),

(iv) Normal kernel estimate of AI (Figure 5.32).

The function $\hat{L}(t)$ presented in Figure 5.32 can be considered to oscillate around the linear function $y = a + bt$. Here, the AI estimators of parameters of the Modified Weibull

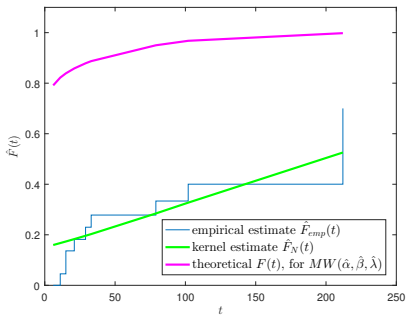

Figure 5.29: cdf for LGT, Normal kernel

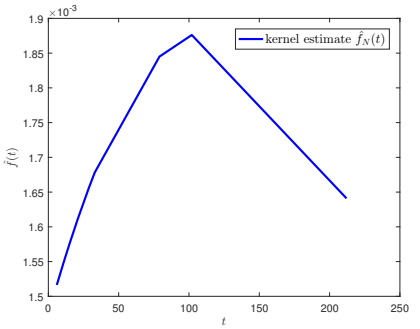

Figure 5.30: pdf for LGT, Normal kernel

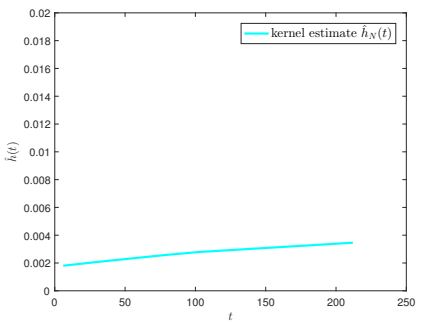

Figure 5.31: HR for LGT, Normal kernel

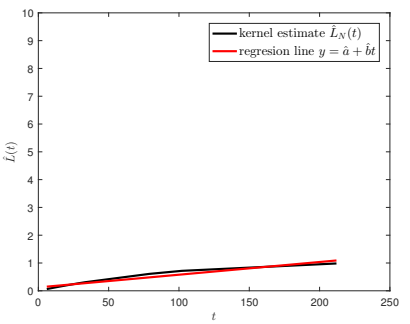

Figure 5.32: AI for LGT, Normal kernel

distribution $MW(\alpha,\beta,\lambda)$ (with linear AI) are $\hat{\alpha} = \hat{a} = 0.1220$, $\hat{\beta} = \hat{b} = 0.0046$, respectively, and the maximum likelihood estimate is $\hat{\lambda} = 1.2252$ (using (5.8)). But we can see in Figure 5.29 that the theoretical distribution function $F(t)$ of the determined (by Normal kernel estimates) Modified Weibull distribution (magenta function) is not close to the empirical estimate $\hat{F}_{emp}(t)$ (blue step function). So, we have to reject the hypothesis that the lifetime of LGT patients follows MW distribution (and this is also the case for the other Box, Epanechnikov, and Triangle kernels under consideration). Moreover, the usage of Kolmogorov-Smirnov goodness-of-fit (see D' Agostino and Stephens (1986)) to verify the hypothesis that data really follow Modified Weibull distribution is controversial because the percent of censored data is higher then 60%.

Acknowledgements

The authors thank the anonymous reviewers and the Editor for their constructive comments which led to the improved version of the manuscript. The first author was partially supported by PUT under grant 0211/SBAD/0123. The corresponding author would like to thank Higher Education Department, Government of Odisha under OHEPEE (Grant No. HE-PTC-WB-02017) and Odisha State Higher Education Council for providing support to carry out the research project under OURIIP, Odisha, India (Grant No. 22-SF-MT-073).

References

D'Agostino, R. B., Stephens, M. A., (1986). Goodness-of-Fit Techniques, New York, NY, USA: Marcel Dekker.

Bhattacharjee, S., Mohanty, I., Szymkowiak, M. and Nanda, A. K., (2022). Properties of aging functions and their means. *Communications in Statistics: Simulation and Computation*, https://doi.org/10.1080/03610918.2022.2141257

Bhattacharjee, S., Nanda, A. K., and Misra, S. K., (2013a). Reliability analysis using aging intensity function. *Statistics and Probability Letters*, 83 (5), pp. 1364–1371.

Bhattacharjee, S., Nanda, A. K., and Misra, S. K., (2013b). Inequalities involving expectations to characterize distributions. *Statistics and Probability Letters*, 83 (9), pp. 2113–2118.

Bowman, A. W., Azzalini, A., (1997). Applied smoothing techniques for data analysis, New York: Oxford University Press Inc.

Daniel, W. W., Cross, C. L., (2014). Biostatistics: Basic Concepts and Methodology for the Health Sciences, Tenth Edition, Wiley India Pvt Ltd.

DiNardo, J., Tobias, J., (2001). Nonparametric density and regression estimation, *The Journal of Economic Perspectives*, 15, pp. 11–28.

Emmerson, J., Brown, J. M., (2021). Understanding Survival Analysis in Clinical Trials, *Clinical Oncology*, 33 (1), pp. 12–14.

Giri, R. L., Nanda, A. K., Dasgupta, M., Misra, S. K., and Bhattacharjee, S., (2023). On aging intensity function of some Weibull models. *Communications in Statistics-Theory and Methods*, 52(01), pp. 227-262. DOI: 10.1080/03610926.2021.1910845

Jiang, R., Ji, P., and Xiao, X., (2003). Aging property of univariate failure rate models. *Reliability Engineering and System Safety*, 79, pp. 113–116.

Klein, J. P., Moeschberger, M. L., (2003). Survival Analysis Techniques for Censored and Truncated Data: Statistics for Biology and Health, Series Editors: Dietz, K., Gail, M., Krickeberg, K., Samet, J,. and Tsiatis, A., 2nd Edition, Springer-Verlag New York, Inc.

Martini, N., Huvos, A. G., Burt, M. E., Heelan, R. T., Bains, M. S., Mccrmack, P. M., Rusch, V. M., Weber, M., Downey, R. J., and Ginsberg, R. J., (1996). Predictions of Survival in Malignant Tumors of the Sternum, *Journal of Thoracic and Cardiovascular Surgery*, 111, pp. 96–106.

Miladinovic, B., (2008). Kernel density estimation of reliability with applications to extreme value distribution, Graduate Theses and Dissertations submitted in University of South Florida, http://scholarcommons.usf.edu/etd/408.

Misra, S. K., Bhattacharjee, S., (2016). Properties of Weibull models, *Far East Journal of Mathematical Sciences*, 100(12), pp. 1965–1979.

Misra, S. K., Bhattacharjee, S., (2018). A case study of aging intensity function on censored data, *Alexendria Engineering Journal*, 57, pp. 3931–3952.

Nanda, A. K., Bhattacharjee, S., and Alam, S. S., (2006). Properties of proportional mean residual life model, *Statistics and Probability Letters*, 76 (9), pp. 880–890.

Nanda, A. K., Bhattacharjee, S., and Alam, S. S., (2007). Properties of aging intensity function, *Statistics and Probability Letters*, 77, pp. 365–373.

Nanda, A. K., Bhattacharjee, S., and Balakrishnan, N., (2010). Mean residual life fuction, associated orderings and properties, *IEEE Transactions on Reliability*, 59(1), pp. 55–65.

Nayak, A., Kumar, S., Singh, S. P., Bhattacharyya A., Dixit A., and Roychowdhury A., (2021). Oncogenic potential of ATAD2 in stomach cancer and insights into the protein-protein interactions at its AAA+ATPase domain and bromodomain, *Journal of Biomolecular Structure and Dynamics*, pp. 1–17.

Rosen, K., Prasad, V., Chen, E. Y., (2020). Censored patients in Kaplan-Meier plots of cancer drugs: An empirical analysis of data sharing, *European Journal of Cancer*, 141, pp. 152-161. doi: 10.1016/j.ejca.2020.09.031.

Shaked, M., Shanthikumar, J. G., (2007). *Stochastic Orders (Springer Series in Statistics)*. San Diego.

Szymkowiak, M., (2018). Characterizations of Distributions Through Aging Intensity, *IEEE Transactions on Reliability*, 67, pp. 446–458.

Swain, P., Bhattacharjee S. and Misra, S. K., (2021). A Case Study to Analyze Ageing Phenomenon in Reliability Theory, *Reliability: Theory and Applications*, 16(4 (65)), pp. 275–285.

Szymkowiak, M., (2019). Measures of aging tendency, *Journal of Applied Probability*, 56(2), pp. 358–383.

Szymkowiak, M., (2020). Lifetime analysis by aging intensity functions studies in systems. In Decision and Control. Series editor. Vol. 196. Switzerland: Springer Nature.

Gamrot, W., (2012). Estimation of Finite Population Kurtosis under Two-Phase Sampling for Nonresponse. *Statistical Papers*, 53, pp. 887–894.

Gamrot, W., (2013). Maximum Likelihood Estimation for Ordered Expectations of Correlated Binary Variables. *Statistical Papers*, 54, pp. 727–739.

Gamrot, W., (2012). Estimation of Finite Population Kurtosis under Two-Phase Sampling for Nonresponse. *Statistical Papers*, 53, pp. 887–894.

Gamrot, W., (2013). Maximum Likelihood Estimation for Ordered Expectations of Correlated Binary Variables. *Statistical Papers*, 54, pp. 727–739.

Särndal, C-E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.