Non-parametric estimation of hazard function with covariates and censoring

Savita Upadhyay

Department of Statistics, Portland State University, Portland, 97201, Oregon.

Abstract

In this paper, we explore the behavior of the kernel density estimator proposed by Zhao et al.[2022]in the paper [ZML22], for estimation of hazard function for censored-survival data with covariates. The kernel density estimator used in this paper is based on the conditional survival model and has two bandwidths, one in the direction of covariates and the other in the direction of event time. We will also discuss the selection of bandwidths which influence the smoothing of the density estimates. We then compare the cumulative hazard from the kernel density estimator with the Nelson-Aalen estimator. We study the two methods on simulated data with one covariate and random uniform censoring. The event time is simulated based on the covariate. We simulated event time from three commonly used distributions exponential, log-normal and Weibull. The study shows that the kernel density estimator very closely approximates the hazard and cumulative hazard curve whereas the Nelson-Aalen estimator performs poorly as it is based solely on event time.

Keywords: Survival analysis, hazard, censored-survival data, covariates, non-parametric, Nelson-Aalen, bandwidth, kernel, kernel density estimator

1 Introduction

In many real life situations we are often interested in analyzing the time to an event. For instance time to an event can be the time from the first diagnosis or treatment to death of a patient suffering from an incurable disease. Study of time to an event is known as survival analysis.

The hazard rate function is conditional probability of occurrence of an event at a given point of time, given that the event of interest has not yet occurred. For example the hazard rate function can be interpreted as rate of deterioration in case of a patient suffering from an incurable disease which worsens the patient's health with time. Hazard rate gives a better understanding of the situation. For instance, changes in failure rate post-treatment or at different stages of diagnosis give critical information when studying the survival analysis for clinical data.

If the event of interest does not occur during the study period the data is said to be censored. When working with real world data often we receive incomplete or censored

data. The factors influencing the occurrence of an event are known as covariates. For the study of hazard function in this paper we have studied simulated data with covariates and random censoring. In real world the distribution of event time is not known, therefore we want a generalized estimator with theses two properties-(i)It has no assumptions about the event time and covariates (ii)It incorporates the covariates.

There are broadly three models in survival analysis to study the hazard function-parametric, semi-parametric and non-parametric. If the distribution of event time and the relationship between the covariates and time is fully known then the parametric models are recommended. Therefore, parametric models is not a right fit for our problem. Semi-parametric models assume that the hazard function can be factored into baseline hazard, where baseline hazard is a function of event time relating to hazard when all covariates are zero, and a function of covariates. Therefore in semi-parametric models distribution of event time is unspecified, however the relationship of covariates with hazard is fully specified. Therefore, semi- is not ideal for our problem as it has restrictions on form of relation between covariates and event time. Non-parametric models have no assumption on event time and also drop any restriction on relation between covariates and event time. However, there are some challenges with the non-parametric models (1) It is not easy to incorporate covariates. (2) The estimated hazard and cumulative hazard functions are not smooth. For instance Nelson-Aalen estimator, which is widely in survival analysis doesn't incorporate covariates. An alternative approach in non-parametric is kernel density estimation which satisfies the requirements of the generalized estimator.

The kernel density estimator from the paper [ZML22] is a good solution to our problem. This estimator is a two-dimensional kernel density estimator which measures the hazard rate change in the direction of time as well as in the direction of covariates. It is based on conditional survival models which frees up the specific functional form between covariates and event time and is asymptotically normal and most efficient in this class of estimators.

In section 2 we provide definitions of relevant terms. In section 3 we describe Nelson-Aalen estimator, and kernel density estimation and kernel properties. In section 4 we describe the bivariate kernel density estimator that we have used in this paper and its properties, and kernel and bandwidth selection. In section 5 experimental setup and results have been described.

2 Background

In this section, we will define key terms used in survival analysis.

Survival analysis is not limited to study of time as a random variable. It provides tools for the study of any positive random variable. For instance, we can study the number of dollars a health insurance company has paid in a particular case. In some cases where the patient has recovered, we know the total amount of money paid by the insurance company. In the other case, if patient is still stick, we have information about the amount paid by the insurance company to date. In the second case, we do not have any information on the total amount paid by health insurance therefore the observation is said to be censored. For this paper, we will study length of time to an event of interest.

Key terms in survival analysis:

Event: An Event is a well-defined endpoint of interest. It is a dichotomous variable with a value of "1" if the event occurs during the study period and "0" if the

event has not occurred during the study.

Event time: Time from the beginning of an observation period to the time of occurrence of an event, end of the study, loss of contact, or withdrawal from the study is known as event time. We denote it as T in this paper.

Covariates: Covariates are the set of factors or predictors which determine the occurrence of an event of interest at a particular time point. We denote covariates with X in this paper.

Survival models: As stated in the previous section, there are mainly three different types of survival models (1) parametric (2) semi-parametric (3) non-parametric These models can be conditional or unconditional. A survival model in which study of event time is conditioned on covariates is known as conditional survival model. We have used conditional survival model in this paper.

Survival Function: Survival function is the probability that event of interest will occur beyond time t. $f(t)$ is the probability density function of T and $F(t)$ is the cumulative distribution function of T. Then the conditional survival function $S(t|x)$ is defined as:

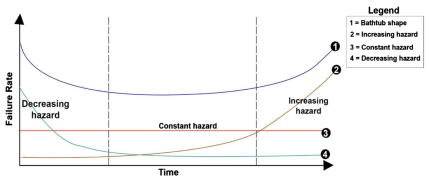
$$S(t|X = x) = 1 - F(t|x) = P(T > t|X = x) \qquad (1)$$

$S(t|x)$ has a unique distribution and is monotonically non-increasing function.

Hazard Function: The hazard rate function is conditional probability of occurrence of an event at a given point of time, given that the event of interest has not yet occurred. Conditional hazard function show how do we incorporate covariates. The conditional hazard function at time time "t" for covariate X, $\lambda(t|x)$ is defined as:

$$\begin{aligned}\lambda(t|X = x) &= lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t, X = x)}{\Delta t} \\ &= \frac{f(t|x)}{S(t|x)} \qquad (2)\end{aligned}$$

There can be multiple points of change in the hazard function. The hazard function is non-negative but have no upper bound. It can be increasing, decreasing, constant, bathtub-shaped, or hump-shaped. Hazard related to wear down of a machine with time is increasing, child mortality hazard decreases with time. Risk of death of humans usually is high at birth and then decrease and remain constant until a certain age and then increase again. This type of hazard is a bathtub-shaped hazard. When hazard increases steadily for some time and then starts declining with time, then it is called a hump-shaped hazard. Hump-shape hazard is common in post-surgery cases.



(a) *Shapes of Hazard function*

Cumulative Hazard function: $\Lambda(t)$ is defined as the total number of events over a periods of time. Cumulative hazard function is always increasing. We can define the conditional cumulative hazard function as :

$$\begin{aligned}\Lambda(t|X=x) &= \int_0^t \lambda(t|x)dt \\ &= -ln\{S(t|x)\} \end{aligned} \tag{3}$$

Censoring: When the researcher has partial information about the event time of a subject, then it is called censoring. There are many types of censoring, with the most common being right censoring. An observation is said to be right-censored when there is no event at the end of the study or if the subject has withdrawn before the end of the study. For instance, if the event of interest is the death of cancer patients, then the observation of all those subjects who are alive at the end of the study is censored. The analysis is done with partial information on censored data as we have no information on the event time for censored data. in this paper we will be studying right-censored data.

Notations: Suppose there are n individuals in a study and the data consists of observations $(T_i, C_i, \Delta_i, X_i)$ where T_i is survival/observation time for the i^{th} individual, Δ_i denotes event and is defined as $\Delta_i = I(T_i \leq C_i)$. C_i is the censoring time(right censored) such that $C_i = T_i$ iff $\Delta_i = 1$ and $C_i < T_i$ iff $\Delta_i = 0$ because all we know is that T_i is longer than C_i. X_i denotes the state of predictor or influencing factor at the time point $Z_i = min(T_i, C_i)$. We assume $T_i \geq 0, C_i \geq 0$ and $Z_i \geq 0$. T_i and C_i are independent. We assume T, Z and C are continuous unless specified otherwise.

3 Non-parametric models for Survival analysis

Non-parametric hazard models provide methodologies for hazard estimation without any assumption on functional form of event time or covariates. Non-parametric methods cannot be defined by a finite parameter space. The term "non-parametric" is little confusing as the non-parametric methods have infinite-dimensional parameter set. Nelson-Aalen model is a frequently used model for cumulative hazard function. Another popular method is kernel density estimators for hazard function. We will discuss both the methods briefly in this section.

3.1 Nelson-Aalen estimator

Nelson-Aalen estimator is a non-parametric estimator which estimates cumulative hazard for censored survival data. It does not require any information on functional

form of distribution of survival data. It is also used to check the fit of parametric models.

The Nelson-Aalen estimator for cumulative hazard function is defined as:

$$\hat{A(t)} = \sum_{t_i \leq t} \frac{\Delta_i}{\sum_{j=1}^{n} I(Z_j \geq Z_i)} \tag{11}$$

Refer to the "Background" section for notations.

3.2 Kernel density estimation

Kernel density estimation is a non-parametric method to estimate the probability density function of a random variable based on kernels and weights using a finite sample from univariate data. In this section will define briefly a , bandwidth, kernel and its properties and kernel density estimator

Let U be a random variable with unknown density function. $u_1, u_2...u_n$ is a random sample drawn from U such that $u_i \overset{iid}{\sim} f_U(u)$.

Kernel is a real-valued integral function K(.) defined on $\mathbb{R} \Rightarrow \mathbb{R}$ which usually satisfies the properties of a probability density function.

Regularity conditions and asymptotic properties of kernels and bandwidths:

There are some regularity conditions which kernel estimator should follow to be asymptotically normal, consistent and achieve optimal efficiency. Basic requirements and conditions on kernels are:

1. Kernels are functions defined on $\mathbb{R} \Rightarrow \mathbb{R}$ which integrates to one.
   $\int_{\infty}^{-\infty} K(u)\, du = 1$

2. Kernels are usually non-negative. Higher order moments of kernel function can be negative. $K(u) \geq 0$.

3. Kernels are symmetric functions.
   $K(u) = K(-u)$

4. Kernel function K(u) is monotonically decreasing for $u > 0$ and are differentiable with bounded derivatives.
   $\int_{\infty}^{-\infty} K'^2(u)du < \infty$,$\int_{\infty}^{-\infty} u^2 K'^2(u)du < \infty$, $\int_{\infty}^{-\infty} K''^2(u)du < \infty$,
   $\int_{\infty}^{-\infty} u^2 K''^2(u)du < \infty$

5. Multi-dimensional kernel can be written as a product of univariate kernels. The d-dimensional multivariate kernel can be written as:
   $K(\mathbf{u}) = \prod_{j=1}^{d} K(u_j)$, for u $= (u_1, u_2...u_d)^T$

6. Order(ν) of a kernel is defined as the first-nonzero moment.
   $K_j(u) = \int_{\infty}^{-\infty} u^j K(u)du$ for $1 \leq j < \nu$, $0 < \int_{\infty}^{-\infty} u^\nu K(u)\, du < \infty$

7. Measure or spread of variance of the kernel estimation is given by the second moment of kernel assuming the first moment is zero. Variance of kernel estimation is defined as :
   $K_2(K) = \int_{\infty}^{-\infty} u^2 K(u)\, du < \infty$

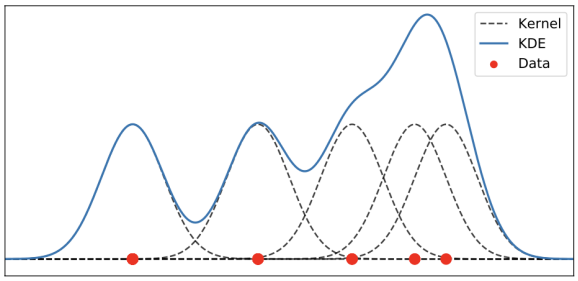
8. Wiggliness or roughness of a kernel estimate is given as :
   $R(K) = \int_{\infty}^{-\infty} K(u)^2\, du$

Bandwidth determines the size of the neighborhood for a kernel density estimator. It controls the smoothness and appearance of density estimate.
Let U be a random variable with unknown density function. $u_1, u_2 ... u_n$ is a random sample drawn from U such that $u_i \overset{iid}{\sim} f_U(u)$.

Kernel density estimator: A kernel density estimator to estimate the density function for the random variable U based on sample data $u_1, u_2 ... u_n$ is defined as:

$$\widehat{f}_h(u) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - u_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{u - u_i}{h}\Big). \tag{12}$$

A kernel is centred at each data point u_i. To get the final kernel density estimate, kernels at all data points are combined and then normalized to ensure that the estimated kernel integrates to one.



(a) *kernel density estimation*

Kernel function K(.) is the weight or measure of nearness of point u_i to u. Neighborhood is determined by the smoothing parameter or bandwidth h. A kernel with subscript h is called the scaled kernel and defined as:

$$K_h(u) = \frac{1}{h} K\big(\frac{u}{h}\big) \tag{13}$$

4 Proposed Kernel Estimator

In real world the distribution of event time is not known, therefore we want a generalized estimator with theses two properties-(i)It has no assumptions about the event time and covariates (ii)It incorporates the covariates. As discussed in the Introduction section we propose the kernel density estimators $\widehat{\Lambda}(Z, X)$ and $\widehat{\lambda}(Z, X)$ from the paper

[ZML22] for cumulative hazard and hazard estimation respectively as a solution to our problem . $\widehat{\lambda}(Z, X)$ is a two dimensional kernel density estimator which captures information in direction of time and covariates as well and assumes no restrictions on form of event time and covariates. The proposed method is asymptotically normal and has lowest variability estimator among the construction of different consistent estimator. The efficiency, consistency and asymptotically normality of the estimator has also been shown in the paper[ZML22]. Hence, we recommend the use of this estimator. The proposed kernel estimator has two bandwidths "h" in the direction of covariates and "b" in the direction of time. Smoothness of the kernel estimate is controlled by careful selection of bandwidths h and b.

4.1 Cumulative hazard and hazard estimator

The kernel density estimator for $\widehat{\Lambda}(Z, X)$ estimated cumulative hazard function and $\widehat{\lambda}(Z,\ X)$ estimated hazard function are:

$$\widehat{\Lambda}(Z, X) = \sum_{Z_i \leq Z} \frac{\Delta_i K_h(X_i - X)}{\sum_{j=1}^{n} I(Z_j \geq Z_i) K_h(X_j - X)} \tag{14}$$

$$\begin{aligned} \widehat{\lambda}(Z,\ X) &= \int_0^{\infty} K_b(t - Z) d\hat{\Lambda}(t|X) \\ &= \sum_{i=1}^{n} K_b(Z_i - Z) \frac{\Delta_i K_h(X_i - X)}{\sum_{j=1}^{n} I(Z_j \geq Z_i) K_h(X_j - X)} \end{aligned} \tag{15}$$

This method is very flexible because (1) It is a non-parametric method. Non-parametric method needs no information about functional form of data. (2)There is no specified form to link covariates with event time.

Properties of estimated kernels

1. **Bandwidths**: For a d-dimensional kernel, $2\nu > d + 1$, if h $\to$ 0, b $\to$ 0, $nh^{d+2}b \to \infty$ and $nh^{2\nu} \to 0$

2. Asymptotic unbiasedness of kernel estimator depends on bandwidth and properties of K(.):
   $\widehat{\lambda}(Z, X) = \lambda(Z, X) + O_p((nhb)^{-\frac{1}{2}} + h^2 + b^2$. (Lemma 1 [**zhao2017efficient**])

3. Kernel estimator has a larger variance than sample variance. This means different K(.) gives us different variance.
   $\widehat{\lambda}(Z, X) - E[\widehat{\lambda}(Z, X)] = O_p(\sqrt{(}Var(\widehat{\lambda}(Z, X)))$. (Lemma 1 [**zhao2017efficient**])

4.2 Kernel and bandwidth Selection

Kernel selection

Kernel estimator has two parameters, i.e., the kernel function K(.) and bandwidth h. As discussed, earlier choice of kernel function is not very crucial for kernel density estimation. The selection of kernel does not impact the accuracy of estimation.

Common Kernel Functions [LOO20]				
Kernel	Equation	R(K)	$K_2(K)$	Efficiency(K)
Uniform	$K(u) = \frac{1}{2} I[\|u\| \leq 1]$	$\frac{1}{2}$	$\frac{1}{3}$	0.9295
Epanechnikov	$K(u) = \frac{3}{4}(1-u^2) I[\|u\| \leq 1]$	$\frac{3}{5}$	$\frac{1}{5}$	1.000
Biweight	$K(u) = \frac{15}{16}(1-u^2)^2 I[\|u\| \leq 1]$	$\frac{5}{7}$	$\frac{1}{7}$	0.9939
Triweight	$K(u) = \frac{35}{32}(1-u^2)^3 I[\|u\| \leq 1]$	$\frac{350}{429}$	$\frac{1}{9}$	0.9867
Gausssian	$K(u) = \frac{1}{\sqrt{2\pi}} exp(\frac{-1}{2}u^2)$	$\frac{1}{2*\sqrt{\pi}}$	1	0.9512
Triangular	$K(u) = (\|u\| - 1) I[\|u\| \leq 1]$	$\frac{2}{3}$	$\frac{1}{6}$	0.9859

Efficiency(K) is the measure of efficiency (minimize AMISE) relative to Epanechnikov kernel. Efficiency of Uniform kernel is 92.95%. Therefore, if a sample size of 100 is needed to obtain an optimal AIMSE using Uniform kernel then we need a sample size of 93 with Epanechnikov kernel to obtain the same optimal AIMSE. As we can see that there is not much difference in the efficiency of kernels, choice of kernels has very little impact on the kernel estimate.

Bandwidth selection

Bandwidth selection is very crucial in kernel density estimation. A small bandwidth leads to undersmoothing and the density estimate plot will be very peaky. A huge bandwidth leads to oversmoothing and can hide modes in the estimate. Density estimate of a multimodal plot may be unimodal incase of oversmoothing.

In conditional hazard we need to estimate two bandwidths where each bandwidth has its own role to play in the quality of estimate. $\widehat{\lambda}(Z, X)$ is a conditional hazard function which uses two bandwidths h and b. Bandwidth h influences smoothness of density estimate in direction of covariates whereas b influences the smoothing of density estimate in direction of event time.

However, the choice of the two bandwidths is not independent. This makes the selection of bandwidths in conditional case even more difficult than the unconditional case where we just have to estimate one single parameter/bandwidth. The computational complexity of estimating bandwidth in conditional kernel density estimation is very high. Computational time of conditional kernel density estimation increases significantly with increase of sample size. Sometimes some assumptions are relaxed to deal with high computational complexity such as h=b[SKH21]. However, this might give a poor estimate owing to the fact that each bandwidth controls the estimates in two different directions.

The two well-known bandwidth selection methods are
1.Plug-in method :replacing unknown parts by their estimators .
2.Cross-validation : Minimize AMSE with respect to bandwidth .

$$MSE = E[\widehat{f}(x) - f(x)]^2$$
$$= bias^2(\widehat{f}(x)) + Var(\widehat{f}(x))$$

$$= \Big[\frac{h^2}{2}\int_{\infty}^{-\infty} f_{''}(\zeta)K(u)u^2, du\Big]^2 + \frac{1}{nh}\int_{\infty}^{-\infty} K(u)^2, du(\text{From Lemma1 [}\textbf{zhao2017efficient}\text{]})$$
$$= \frac{h^4}{4}\Big[\int_{\infty}^{-\infty} f_{''}(\zeta)K(u)u^2, du\Big]^2 + \frac{1}{nh}\int_{\infty}^{-\infty} K(u)^2, du$$
$$= AMSE \qquad (16)$$

Therefore, under weak assumptions MSE equals AMSE. MSE is minimized for bandwidth proportional to $n^{-\frac{1}{5}}$. For this paper we start with a bandwidth proportional to Silverman's bandwidth i.e., equal to $min(IQR, stdev(X)) * (n^{-\frac{1}{5}}) * K$, where K is the proportionality constant and is selected on the basis of MSE value.

5 Numerical Simulation

We will perform three simulation studies to evaluate the performance of the estimator $\widehat{\lambda}(Z, X)$ and compare it against the well-known Nelson-Aalen estimator. We will simulate event time from three most common distributions in survival analysis-exponential, Weibull and log-normal. For each simulation study 1000 samples of 500 triples(X_i, T_i, C_i) of data were simulated. Then the observed data $(X_i, Z_i, C_i, \Delta_i)$ were obtained as $Z_i = min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ for i=1,2,..., n.

X_i is the covariate, T_i is the event time, C_i is the censoring time for the i^{th} observation in a sample. $_i$ is the event for the i^{th} observation. If the observation is censored i.e $T_i > C_i$ then $_i$ is 0 and the observation is said to be censored. If the event is observed during the study i.e $T_i < C_i$ then $_i$ is 1 and is called the complete observation.

For all the three simulation studies we have used Epanechnikov kernel. The bandwidths h, b used in this paper are defined as:

$$h = min(IQR, stdev(X)) * n^{-\frac{1}{5}} * \alpha \qquad (17.a)$$
$$b = min(IQR, stdev(Z)) * n^{-\frac{1}{5}} * \alpha \qquad (17.b)$$

Where IQR is inter-quartile range, stdev is standard deviation and α is a constant multiple. We select bandwidths h and b by minimizing MSE.

Results of three simulation studies are based on 1000 samples of size 500. Mean-square error is calculated as a numerical metric for all three simulations. We plot the actual hazard derived from the distribution of simulated event time and estimated hazard to compare the two. We also compare the actual cumulative hazard, estimated cumulative hazard from the proposed kernel estimator and Nelson-Aalen cumulative hazard by plotting all three with same time and cumulative hazard axes. We optimized the bandwidths h and b to estimate hazard density estimate in two dimensions.

5.1 Simulation 1

In first simulations we will evaluate the hazard estimator $\widehat{\lambda}(Z, X)$ for exponential event times T. We generated covariate X from uniform distribution using $X = Uniform(0, 1)$. Exponential event time for the study was generated using $\mathrm{T} = \exp(\mathrm{X} + \epsilon)$, where ϵ was generated from exponential distribution as $\epsilon = \text{exponential}(1)$. Censoring time was generated from uniform distribution using $C = Uniform(0, c_1)$ where c_1 denotes the censoring proportion. We kept $c_1 = 100$

which censored approximately 8% of observations in data.

Actual hazard and cumulative hazard function for the simulated T

Survival Function of T:

$$S(T = t|X = x) = = \frac{e^x}{t} 1_{(t>=e^x)} \tag{18.a}$$

Hazard Function of T:

$$\lambda(T = t|X = x) = \frac{1}{t} 1_{(t>=e^x)} \tag{18.b}$$

Cumulative hazard $\Lambda(t|X = x)$ can be calculated using (3)

Results

Hazard for different T at fixed X and estimated hazard for different X at fixed T were estimated with a mean square error of 0.23 and 0.06 respectively. Average bandwidth b and h are 0.41 and 0.12 respectively.

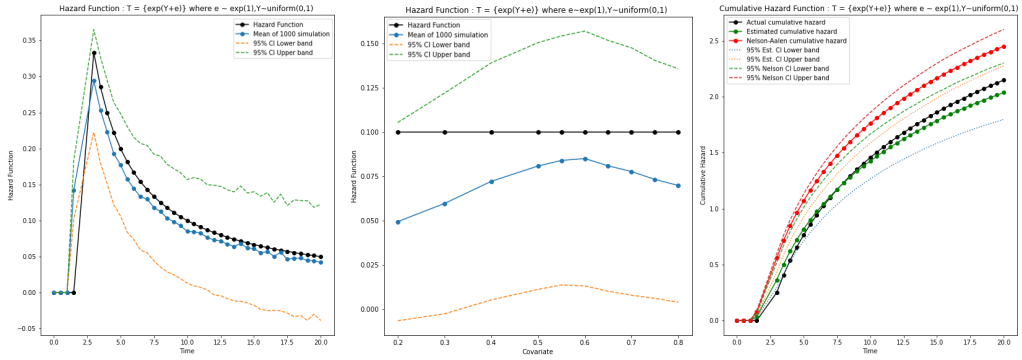


Fig. 3: Simulation1:(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.6.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.6

5.2 Simulation 2

For the second simulation log-normal event time was generated using T = exp(5 - 10$(1 - X)^2$ + ϵ), where ϵ = Normal(0,1) and the covariate X was generated from uniform distribution using $X = Uniform(0, 1)$. Censoring time was generated from uniform distribution using $C = Uniform(0, c_1)$ where c_1 denotes the censoring proportion. c_1 was set at 100 which censored approximately 30% of observations in data

Actual hazard and cumulative hazard function for the simulated T

Survival Function of T:

$$S(t|x) = 1 - \frac{1}{2}\left[1 + erf\left(- \frac{(ln(t) - (5 - 10(1-x)^2))^2}{2}\right)\right] \tag{19.a}$$

Hazard Function of T:

$$\lambda(t|x) = \frac{\frac{1}{t\sqrt{(2*\pi)}} exp\left(\frac{-(ln(t)-(5-10(1-x)^2))^2}{2}\right) 1_{(t>0)}}{1 - \frac{1}{2}\left[1 + erf\left(- \frac{(ln(t)-(5-10(1-x)^2))^2}{2}\right)\right]} \tag{19.b}$$

Result

Hazard for different T at fixed X and estimated hazard for different X at fixed T were estimated with a mean square error of 0.001 and 0.001 respectively. Average bandwidth b and h are 0.12 and 0.1 respectively.

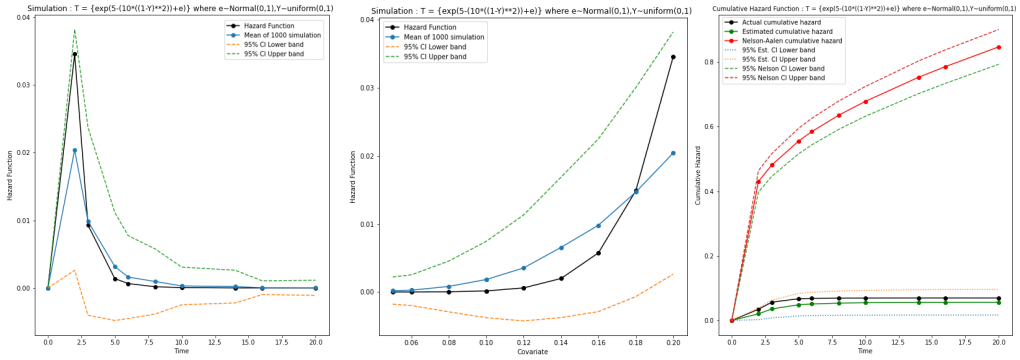


Fig. 4: Simulation2:(i)Comparison of estimated hazard with actual hazard for different T at fixed covariate X=0.2.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T=2.(iii)Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X=0.2

5.3 Simulation 3

For the third simulation we generated covariate X from normal distribution using $X = Normal(0, 1)$. Weibull event time for the study was generated using $\mathrm{T} = 25 * exp(Y) * (-(log\epsilon))^{\frac{1}{3}}$, where ϵ was generated from exponential distribution as $\epsilon =$ exponential(1). Censoring time was generated from uniform distribution using $C = Uniform(0, c_1)$ where c_1 denotes the censoring proportion. We kept $c_1 = 100$ which censored approximately 30% of observations in data.

Actual hazard and cumulative hazard function for the simulated T

Survival Function of T:

$$S(t|x) = e^{(-\frac{t}{25*exp(x))})^3} \tag{20.a}$$

Hazard Function of T:

$$\lambda(t|x) = 3(25 * exp(x))^{-3} t^2 \tag{20.b}$$

Result

Hazard for different T at fixed Y and estimated hazard for different Y at fixed T were estimated with a mean square error of 0.0003 and 3.4e-05 respectively. Average bandwidth b and h are 0.59 and 1.13 respectively.

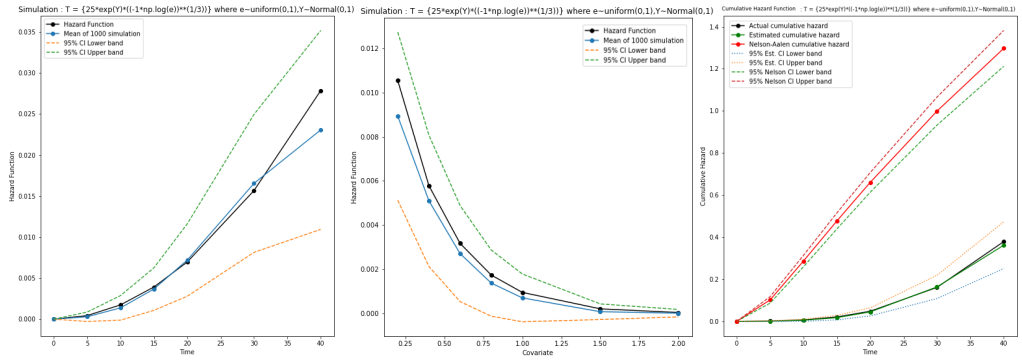


Fig. 5: Simulation3: (i) Comparison of estimated hazard with actual hazard for different T at fixed covariate Y =0.8.(ii)Comparison of estimated hazard with actual hazard for different Y at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate Y =0.8

6 Conclusion

We proposed using kernel density estimator $\widehat{\lambda}(Z, X)$ (15) from the paper [ZML22] and evaluated its performance using numerical simulations. We tested the method only on simulated data for three most common event time distributions in survival data-log-normal and Weibull. Using the kernel density estimator $\widehat{\lambda}(Z, X)$ we obtained estimates of hazard and cumulative hazard which closely resembles the shape of the actual hazard and cumulative hazard for all the three simulations. The cumulative hazard estimated from this method was compared with those of Nelson-Aalen estimator. Nelson-Aalen estimator performed poorly in all three simulations as it does not account for the covariate. The proposed kernel captures changes in the hazard function in the direction of observed time as well as the covariate in the simulated data. However, the

simulated surface can be irregular at some time points leading to higher estimation error. Boundary effect correction method may be explored further to rectify this issue.

There are automatic procedures such as MSE and cross validation methods for bandwidth selection, but they cannot be relied whole self. The estimates can be bumpier with a very low MSE because MSE is prone to outliers. We used mean square error and visualization to select bandwidths. Selection of bandwidth can be subjective based on the degree of smoothness. Complexity of bandwidth selection increase with increase in dimension of kernel and sample size. The estimator goes through n(sample size) data points and assign weights to m equidistant grid points. The algorithm runs in O(N2d) time in d dimensions. As the number of covariates increase we need a more advanced and sophisticated tool for visualization. However, this doesn't mean that we should discard the kernel estimation method.

Acknowledgments

I would like to thank my advisor Prof. Ge Zhao for his guidance throughout this work.

References

[LOO20] Reuben Lang'at, George Orwa, and Odhiambo Otieno. "Kernel function and nonparametric regression estimation: which function is appropriate?" In: *African Journal of Mathematics and Statistics Studies* 3 (2020), pp. 51–59.

[SKH21] Iveta Selingerova, Stanislav Katina, and Ivanka Horova. "Comparison of parametric and semiparametric survival regression models with kernel estimation". In: *Journal of Statistical Computation and Simulation* 91 (2021), pp. 2721–2723.

[ZML22] Ge Zhao, Yanyuan Ma, and Wenbin Lu. "Efficient Estimation for Dimension Reduction with Censored Data". In: *Statistica Sinica* 32 (2022), pp. 2359–2380.

7 Appendix

The estimated curved surface can be irregular at some points, and this could result in increased estimation error. Here are some additional results for the three simulations.

7.0.1 Additional results for simulation 1:

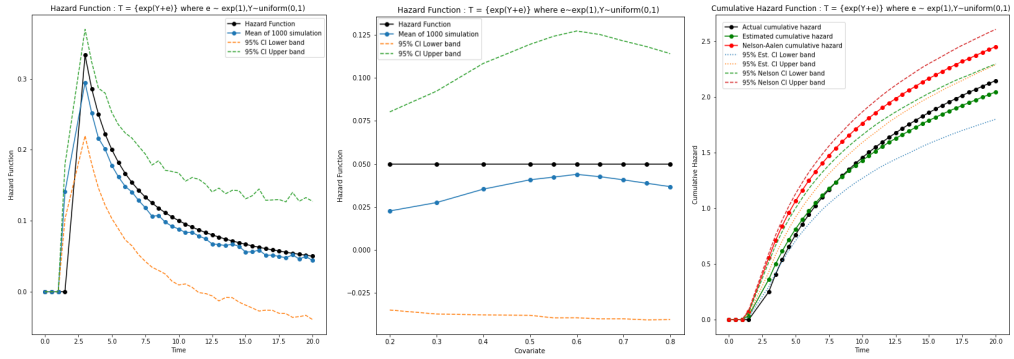


Fig. 6: Simulation1 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.6.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =20.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.6

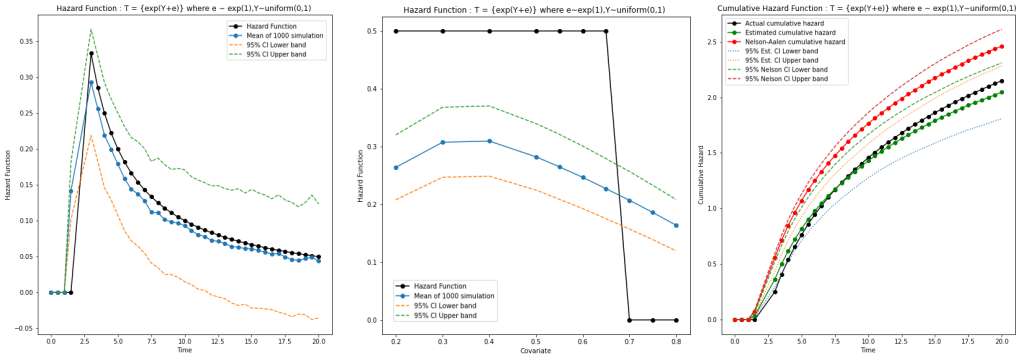


Fig. 7: Simulation1 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X=0.6.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =2.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.6

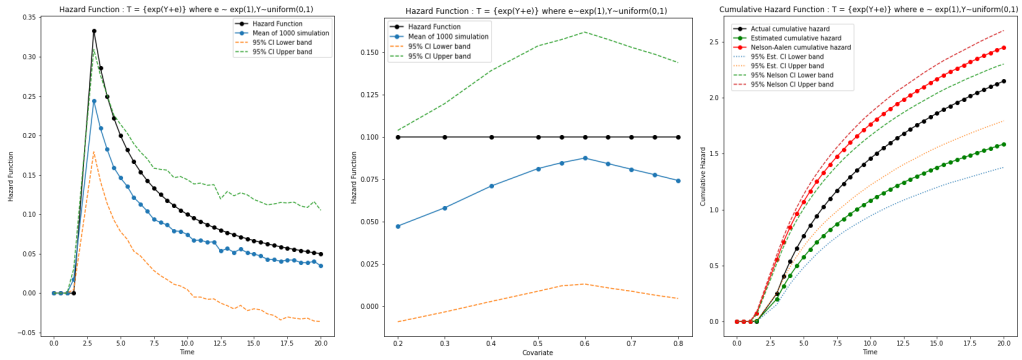


Fig. 8: Simulation1 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.8.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.8

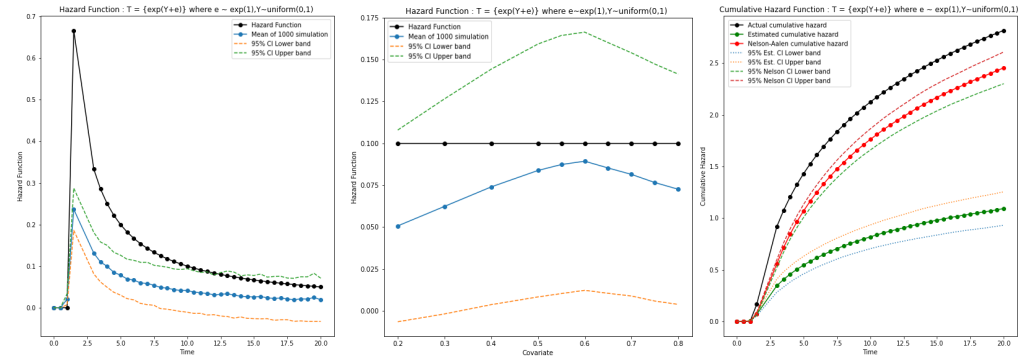


Fig. 9: Simulation1 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.1.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.1

7.0.2 Additional results for simulation 2:

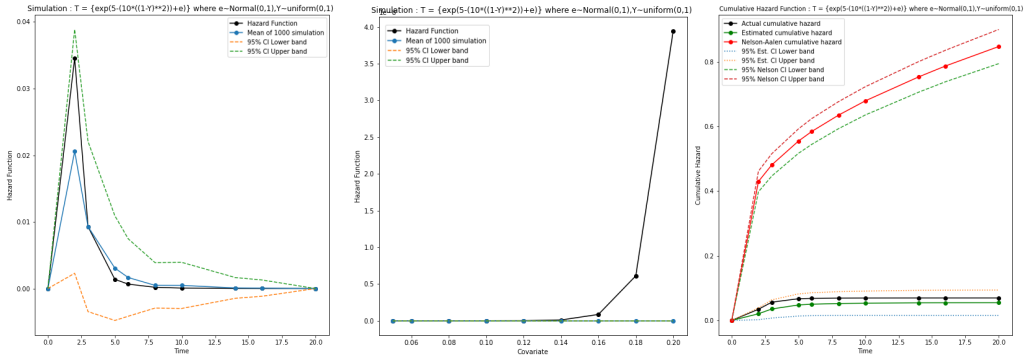


Fig. 10: Simulation2 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.2.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =40.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.2

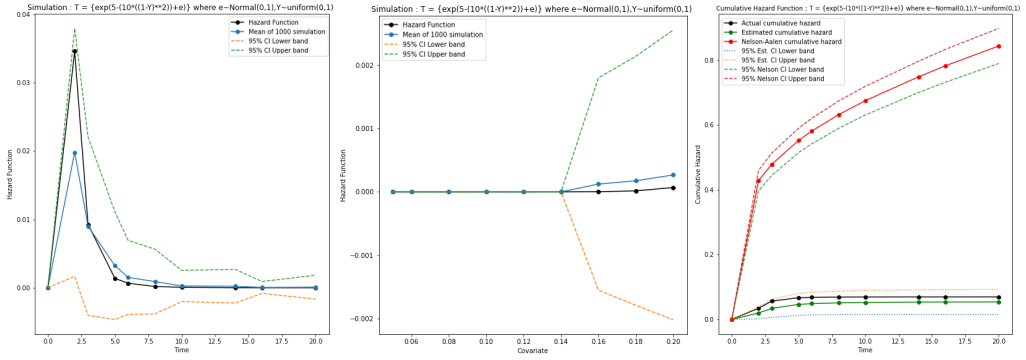


Fig. 11: Simulation2 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.2.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.2

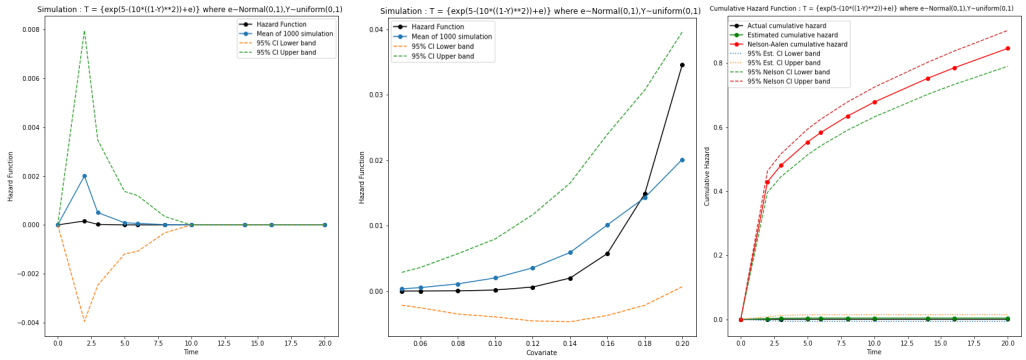


Fig. 12: Simulation2 : (i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.1.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =2.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.1

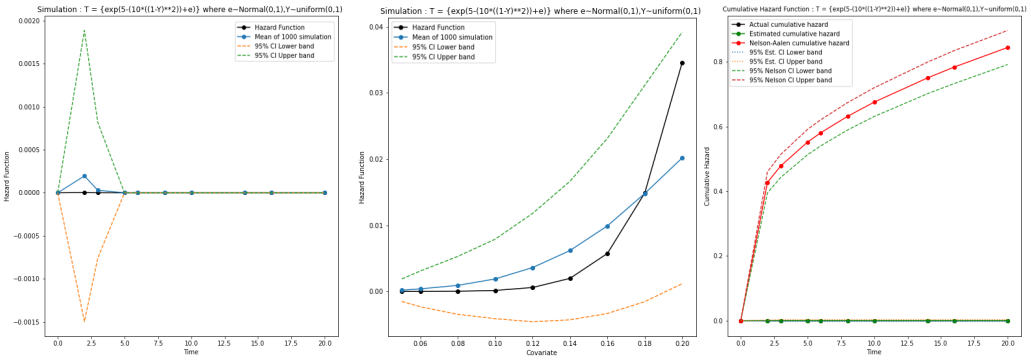


Fig. 13: Simulation2 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.05.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =2.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.05

7.0.3 Additional results for simulation 3:

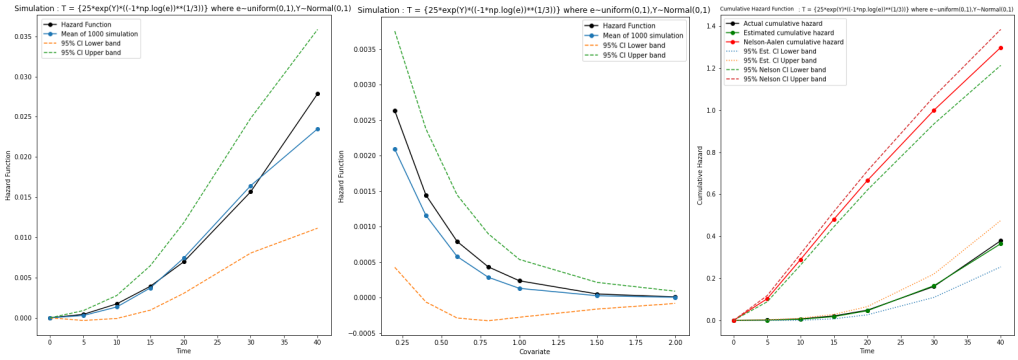


Fig. 14: Simulation3 :Estimated hazard and cumulative hazard from simulated data using proposed method. (i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.8.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =5.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.8

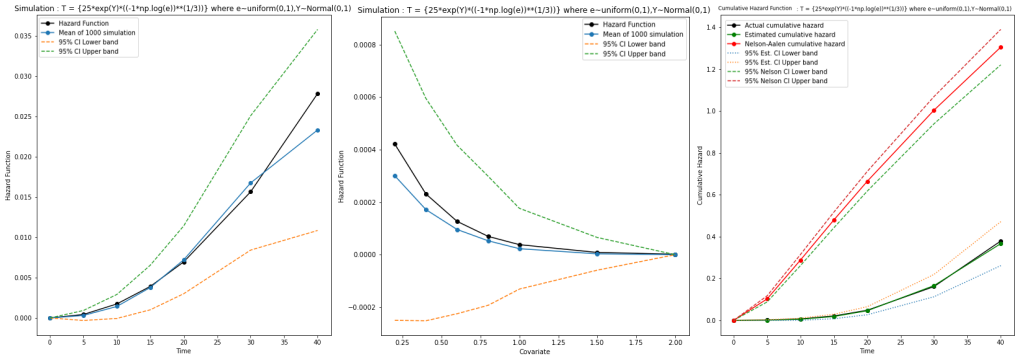


Fig. 15: Simulation3 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.8.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =2.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.8

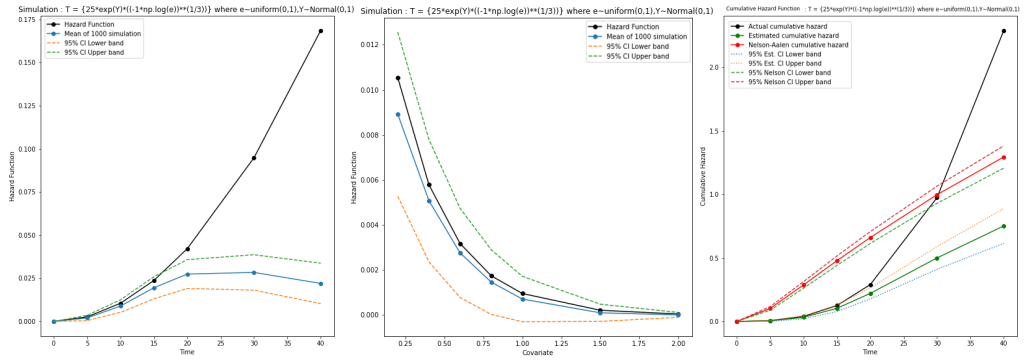


Fig. 16: Simulation3 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.2.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.2

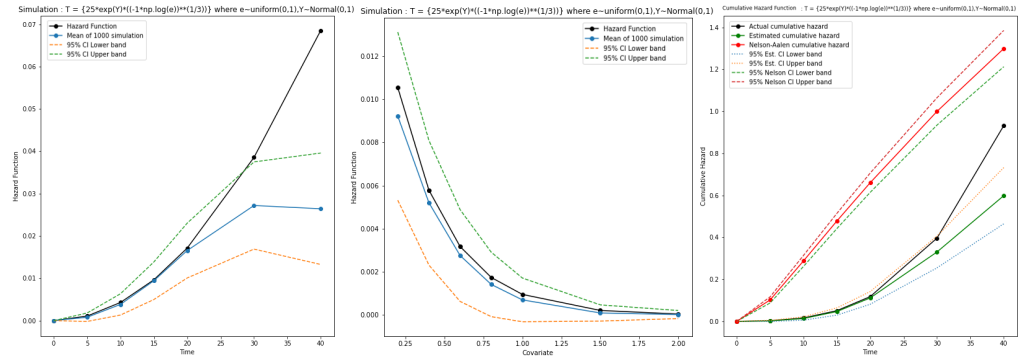


Fig. 17: Simulation3 :(i) Comparison of estimated hazard with actual hazard for different T at fixed covariate X =0.5.(ii)Comparison of estimated hazard with actual hazard for different X at fixed event time T =10.(iii) Comparison of estimated cumulative hazard with actual cumulative hazard for different T at fixed covariate X =0.5