Some Results

2020-02-11

Abstract

It is a continuing challenge to handling censoring in survival analysis. The most commonly used models are defined as the general random censorship model from the right (GRCM), which assumes that the censoring time is independent of the survival time. However, this assumption is strong and hard to achieve in practice. We propose a new weaker assumption of independence and extend a semi-parameter model developed by Dikta (1998) under this condition. We provide examples and simulation studies to illustrate the validation of the model under our new assumption. The consistency of the extended model is also studied.

Introduction

In lifetime data or failure time data, it is almost impossible to collect the outcome variable of all participants. Missingness, which is recognized as censorship in survival study, frequently happens due to various reasons. For example, a participant may move to another city and then lost connection, or a participant may drop off the trial since some negative effect of the interventions. Based on different mechanisms, censorship can be categorised into non-informative and informative censoring, depend on whether the survival time and the censoring time are dependent or not. However, dealing with informative censoring is challenging, and lots of existing methods, such as Kaplan Meier estimator or Koziol-Green model, need the assumption of independent censoring. Kaplan Meier estimator (KME), also known as a product-limit estimator (PLE), is one of the most commonly used non-parametric methods in survival analysis. Several semi-parametric random censorship models (SRCM) were proposed by Dikta et al., which generalized the product-limit estimator. The asymptotic properties of the SRCM were also provided by Dikta. Besides, he also showed that the SRCMs achieve better performances than KME in terms of asymptotic variance. Although the above methods are widely applied, the independence assumption is necessary and usually too strong to meet in practice. Therefore, we demonstrate a weaker assumption of independence, which generalizes the consistency of Dikta's SRCMs. Besides, a new semi-parametric estimator is also proposed in this paper.

The rest of the article is organized as follows. In section 2, we propose the generalized assumptions and prove its association with the condition proposed by Slud. In addition, we provide a new semi-parametric estimator based on the ones proposed by Dikta. In section 3, we construct examples and simulation studies to show that the semi-parameter models perform well under the weak assumption. A discussion is contained in section 4.

Relaxed Assumption

We denote $T_i, i = 1, ..., N$ are the independent, identically, distributed (iid) lifetimes, whose corresponding cumulative distribution function (CDF) is F, probability distribution function (PDF) is f; the censoring time is defined as $C_i, i = 1, ..., N$. C_is are also iid, with CDF denoted as G and PDF denoted as g. We set the censors happen on the right and the observed time is $Z_i = T_i \wedge C_i$, whose CDF is H and PDF is h. The $\delta_i = I_{[T_i \leq C_i]}$ is the status indicator, which shows whether the event of the ith subject is censored ($\delta_i = 0$) or observed ($\delta_i = 1$). The corresponding hazard function of lifetime is λ_F and cumulative hazard function is Λ_F. Besides, we set λ_H as the hazard function for the observed time, which is known as crude hazard rate as well, and its cumulative hazard function is Λ_H. The most commonly applied Kaplan Meier prodcut limit estimator is defined as

$$S^{KM}(t) = \prod_{Z_i \leq t} \Big(1 - \frac{\delta_i}{n - R_{i,n} + 1}\Big)$$

Dikta (1998) proposed another product limited estimator defined as

$$S^{D1} = \prod_{Z_i \leq t} \Big(1 - \frac{m_n(Z_{k:n})}{n - R_{i,n} + 1}\Big)$$

where the $m(t)$ is defined as the conditional expectation of δ given observed time Z

$$m(t) = P(\delta = 1|Z = t) = E(\delta|Z = t)$$

He argued that the semiparameter S^{D1} is unbiased and has less variance than Kaplan Meier estimator. He also produced a new semi-parametric estimator based on it as

$$S^{D2} = \prod_{Z_i \leq t} \Big(1 - \frac{m_n(Z_{k:n})}{n - R_{i,n} + m_n(Z_{k:n})}\Big)$$

and its self consistency has been proved by Dikta (2011). However, all those estimators need independence between T and C, which is hard to satisfy in practice. Instead of the strong condition, Slud demonstrated an alternative assumption on the independence between survival time T and the censoring time C. He defined a function $\rho(t)$ as

$$\rho(t) = \lim_{\delta \to 0} \frac{P(t < T < t + \delta|T > t, C \leq t)}{P(t < T < t + \delta|T > t, C > t)} \tag{1}$$

If $\rho(t) > 1$ for all t, we have positive dependence between death and censoring while if $\rho(t) < 1$ uniformly, we have negative dependence. Besides, if the censoring time and the death time are independent, the $\rho(t) = 1$ for all t. However, when $\rho(t) = 1$, it does not equivalent to the independence but the diagonal independence, i.e. this assumption is weaker then the independence since it only restricts on the timepoint where $T = t = C$. However, he did not give detalied illstration about its application. We then propose a more relaxed assumption,

$$\lim_{dt \to 0} \Big\{ P(T > t + dt, C > t) - P(T > t + dt)P(C > t) \Big\} = 0 \tag{2}$$

As well as

$$P(C > t, T \geq t) = P(C > t)P(T \geq t) \tag{3}$$

Or we may write it as

$$\exists \epsilon > 0, s.t. \text{ for } \forall |dt| < \epsilon, P(T \geq t + dt, C > t) - P(T \geq t + dt)P(C > t) = 0, \text{ for } \forall |dt| < \epsilon \tag{4}$$

The new relaxed condtioin is weaker than Slud's assumption of independence, i.e., $\rho(t) = 1$, since when the relaxed assumption is satifsied, the $\rho(t)$ is not necessory to be 1. Proof of the relationship between these two assumption can be found in the Appendix.

When the relaxed condition holds, we have

$$\begin{aligned} m(t) = P(\delta = 1|Z = t) =& \frac{P(C > t, T = t)}{P(Z = t)} = \frac{P(C > t|T = t)P(T = t)}{P(Z = t)} \\ =& \frac{P(C > t|T > t)P(T = t)}{P(Z = t)} = \frac{P(T = t)}{P(Z = t)} \frac{P(C > t, T > t)}{P(T > t)} \\ =& \frac{f(t)S_x(t)}{h(t)S(t)} = \frac{\lambda_F(t)}{\lambda_H(t)} \end{aligned}$$

Therefore,

$$m(t) = \frac{\lambda_F(t)}{\lambda_H(t)} \tag{5}$$

Therefore, we derive the same $m(t)$ function as proposed by Dikta (1998) under the weaker independence assumption. We show that Dikta's methods still give good estimation under our relaxed assumption in the next section.

Numerical Simulation

In this section, we conduct a simulation study to illustrate our assumptions mentioned in the above section: (i) The both the semi-parametric estimators proposed by Dikta and the new constructed estimator have good performances under the relaxed assumption regarding of survival time and censoring time; (ii) The semi-parametric methods can have better estimation than Kaplan Meier estimator and Cox PH model when there are some covariates related to censoring.

The simulated datasets are generated as the following joint distribution:

$$S_{T,C}(x,y)=\begin{cases} e^{-\theta_1 x}e^{-(e^{\theta_2 y}-1)(x-y+1)} & \text{when } x\geq y\\ e^{-\theta_1 x}e^{-(e^{\theta_2 y}-1)} & \text{when } x<y\end{cases}$$

where T and C are random variables representing survival time and censoring time, respectively. Besides, the θ_1 and θ_2 are parameters that are associated with other covariates.

$$\theta_1=\beta_1^T\mathbf{X}_1,\theta_2=\beta_2^T\mathbf{X}_2$$

$\mathbf{X}_1,\mathbf{X}_2$ are two covariates vectors.

The survival functions and censoring functions can be intutively derived from the joint distribution,

$$S_T(x)=P(T>x)=P(T>x,C>0)=e^{-\theta_1 x}$$

$$S_C(x)=P(C>x)=P(T>0,C>x)=e^{-(e^{\theta_2 x}-1)}$$

Notice that the survival time and the censoring time are not independent but satisfy the relaxed independence assumption. The observed time Z is defined as $Z=T\wedge C$, with distribution

$$S_Z(x)=P(T>x,C>x)=e^{-e^{\theta_2 x}-\theta_1 x+1}$$

The censoring indicator $\delta=I(T<C)$. And the $m(t,x)$ function, which shows the $E(\delta|H=t,X=x)$ can be calculated as:

$$m(t)=\frac{\lambda_T(x)}{\lambda_Z(x)}=\frac{\theta_1}{\theta_1+\theta_2 e^{\theta_2 x}}=\frac{1}{1+\frac{\theta_2}{\theta_1}e^{\theta_2 x}}$$

Consider the covariates X_1,X_2, which can be measures for the population, such as age, education level, blood pressure, etc. Let X_1 present the blood pressure levle, which is set as a one dimension discrete covariate with four levels $1,2,3,4$; Let

$$\theta_1=\beta_1 X_1,\ \theta_2=\beta_1 X_1\times I(\text{sex}=\text{Female})$$

That is, two discrete covariates are included in the simulation, the blood pressure X_1 with four levels and the gender with two levels. We may also notice that the censoring dependent on gender. For males, $\theta_2=\beta_1 X_1\times I(\text{ sex}=\text{F})=0$. The corresponding CDF is

$$S_{T,C}(x,y)=\begin{cases} e^{-\theta_1 x} & \text{when } x\geq y\\ e^{-\theta_1 x} & \text{when } x<y\end{cases}\ \text{,since}\ \ e^{-(e^{\theta_2 y}-1)}=1$$

That is, for males, the joint distribution only dependents on T. The censoring will not happen. On the other hand, for the females, $\theta_1=\beta_1 X_1$, $\theta_2=\beta_1 X_1\times I(\text{ sex}=\text{F})=\beta_1 X_1$. That is $\theta_1=\theta_2$. Therefore, within females,

$$m(t)=E(\delta=1|Z=t,X=x)=\frac{1}{1+e^{\beta_1 xt}}=\frac{1}{1+e^{\beta_1(xt)}}$$

Also, notice that, the $m(t)$ follows a logistic distribution in terms of xt.

In this model, the censoring depends on gender. Only females may drop the trial and bring censor. Besides, the censoring is also related to the covariate X_1. We set $\beta_1 = 1.5$. The propotion of females is 0.5, with sample size $n = 2000$. The four levels of X_1 are all have the same probabilty to be included in the dataset, with $p = 0.25$. 1000 repetitions have been conducted for the simulation. Four methods, the Cox PH model, two semi-parametric estimators from Dikta and the new propsed model are applied to estimate the data. Illustrations and codes about the model and the simulation can be found in the Appendix.

Results

When facing a data set with censoring and covariates, the Cox PH model is commonly used by researchers. However, the this simulation setting shows that the Cox PH can be biased, since by fitting a Cox PH model with both X_1 and gender, the p value for gender is not significant and probably be excluded from the model. However, by fitting a logsitic regression with outcome as censoring status δ and covariates X_1 and gender, both X_1 and gender are significant and are related to censoring and thus may affect the survival time estimation. The p-values are reported in the following times.

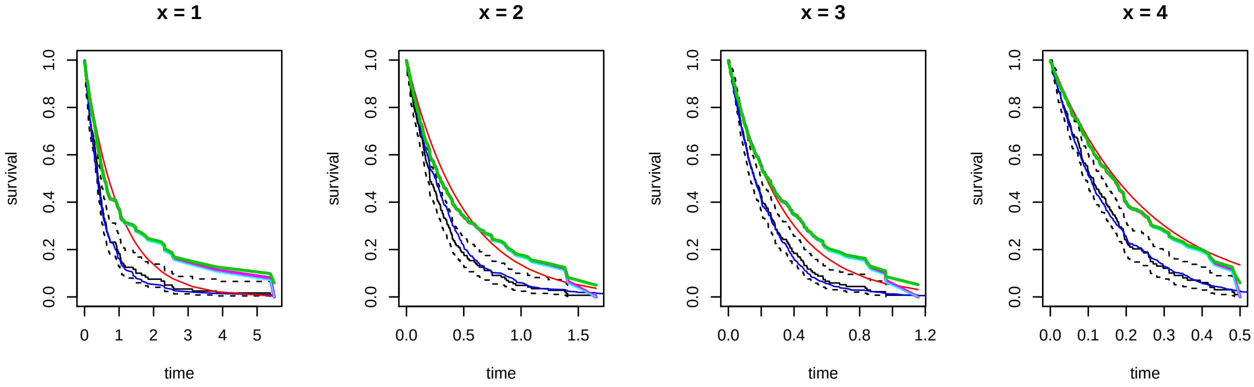
Table 1: Cox PH model

	X1	Sex
p-value	<0.001	0.521

Table 2: Logistic regression

	Time	X1	Sex
p-value	<0.001	<0.001	<0.001

Since the sex is not significant in the Cox PH model, its only fitted with covariate X_1. How the survival function T is estimated can be intutively observed for the survival plots:



The time-dependent AUC at time $t = 0.15$ for the methods is shown in the Table 3. The Cox ph method have smaller AUC comparing to the other methods.

If the m function is correctly specified, the parameter β can be estimated well and the AUC calculated with the true $m()$ or $\hat{m}()$ with $\hat{\beta}$ have similar values. All of them are higher than the AUC estimated by cox PH model, which ignored the effect of gender.

Given the covariate X_1, we may also estimate the quantiles, i.e. $90\%, 75\%, 50\%, 25\%, 10\%$. The bias, standard deviation and mean square error at each quantile above with different methods are listed in the following tables.

Table 3: Time-dependent AUC

		True m()			Est m()		
	coxph	Dikta1	Dikta2	Exp m	Dikta1	Dikta2	Exp m
Mean	0.6858	0.7582	0.7582	0.7516	0.7582	0.7582	0.7516
SD	0.0167	0.0153	0.0153	0.0171	0.0153	0.0153	0.0171

Table 4: Mean absolute differences

		True m()			Est m()		
	coxph	Dikta1	Dikta2	Exp m	Dikta1	Dikta2	Exp m
x = 1	0.0907	0.0261	0.0260	0.0264	0.0264	0.0264	0.0267
x = 2	0.0607	0.0236	0.0236	0.0239	0.0242	0.0242	0.0244
x = 3	0.0717	0.0232	0.0232	0.0237	0.0239	0.0239	0.0244
x = 4	0.1058	0.0210	0.0210	0.0216	0.0215	0.0215	0.0220

Table 5: Mean absolute difference between estimated and true S()

		True m()			Est m()		
Quantile	coxph	Dikta1	Dikta2	Exp m	Dikta1	Dikta2	Exp m
x = 1							
0.90	-0.0696	-0.0191	-0.0192	-0.0189	-0.0193	-0.0195	-0.0192
0.75	-0.1507	-0.0309	-0.0312	-0.0305	-0.0322	-0.0326	-0.0318
0.50	-0.2090	-0.0059	-0.0069	-0.0045	-0.0094	-0.0104	-0.0080
0.25	-0.1548	0.0406	0.0384	0.0442	0.0377	0.0356	0.0413
0.10	-0.0732	0.0476	0.0431	0.0579	0.0461	0.0417	0.0563
x = 2							
0.90	-0.0318	-0.0143	-0.0145	-0.0145	-0.0146	-0.0147	-0.0147
0.75	-0.0718	-0.0253	-0.0256	-0.0251	-0.0267	-0.0271	-0.0266
0.50	-0.1112	-0.0045	-0.0055	-0.0035	-0.0085	-0.0094	-0.0074

0.25	-0.0989	0.0388	0.0367	0.0420	0.0356	0.0335	0.0388
0.10	-0.0525	0.0454	0.0407	0.0531	0.0438	0.0391	0.0514
x = 3							
0.90	-0.0339	-0.0146	-0.0147	-0.0149	-0.0149	-0.0150	-0.0152
0.75	-0.0767	-0.0236	-0.0239	-0.0242	-0.0251	-0.0254	-0.0257
0.50	-0.1187	-0.0062	-0.0071	-0.0058	-0.0105	-0.0114	-0.0101
0.25	-0.1057	0.0407	0.0386	0.0429	0.0371	0.0350	0.0393
0.10	-0.0594	0.0420	0.0374	0.0489	0.0401	0.0356	0.0470
x = 4							
0.90	-0.0533	-0.0104	-0.0105	-0.0110	-0.0106	-0.0107	-0.0112
0.75	-0.1158	-0.0206	-0.0210	-0.0213	-0.0221	-0.0224	-0.0228
0.50	-0.1666	-0.0025	-0.0034	-0.0025	-0.0067	-0.0076	-0.0067
0.25	-0.1404	0.0370	0.0349	0.0386	0.0334	0.0313	0.0350
0.10	-0.0745	0.0346	0.0300	0.0428	0.0329	0.0283	0.0410

Table 6: Standard deviation of estimated S()

		True m()			Est m()		
Quantile	coxph	Dikta1	Dikta2	Exp m	Dikta1	Dikta2	Exp m
x = 1							
0.90	0.0126	0.0148	0.0148	0.0149	0.0148	0.0148	0.0149
0.75	0.0206	0.0204	0.0204	0.0204	0.0202	0.0203	0.0202
0.50	0.0247	0.0275	0.0276	0.0276	0.0272	0.0273	0.0273
0.25	0.0161	0.0334	0.0337	0.0333	0.0327	0.0330	0.0325
0.10	0.0102	0.0390	0.0395	0.0329	0.0384	0.0389	0.0323
x = 2							
0.90	0.0094	0.0140	0.0140	0.0141	0.0140	0.0140	0.0142
0.75	0.0146	0.0198	0.0198	0.0198	0.0198	0.0199	0.0199
0.50	0.0185	0.0275	0.0276	0.0270	0.0283	0.0284	0.0278
0.25	0.0149	0.0293	0.0295	0.0289	0.0297	0.0299	0.0293
0.10	0.0091	0.0342	0.0355	0.0325	0.0341	0.0354	0.0324
x = 3							
0.90	0.0112	0.0130	0.0130	0.0129	0.0130	0.0130	0.0129
0.75	0.0140	0.0193	0.0194	0.0195	0.0198	0.0198	0.0200
0.50	0.0183	0.0262	0.0263	0.0262	0.0270	0.0270	0.0271
0.25	0.0149	0.0347	0.0350	0.0343	0.0351	0.0354	0.0349
0.10	0.0070	0.0395	0.0402	0.0353	0.0388	0.0395	0.0347
x = 4							
0.90	0.0151	0.0149	0.0149	0.0154	0.0149	0.0149	0.0154
0.75	0.0192	0.0193	0.0193	0.0199	0.0194	0.0194	0.0200
0.50	0.0223	0.0273	0.0273	0.0275	0.0275	0.0276	0.0278
0.25	0.0167	0.0313	0.0317	0.0322	0.0312	0.0315	0.0320

0.10	0.0068	0.0427	0.0431	0.0359	0.0421	0.0425	0.0353

Table 7: MSE of estimated S()

		True m()			Est m()		
Quantile	coxph	Dikta1	Dikta2	Exp m	Dikta1	Dikta2	Exp m
x = 1							
0.90	0.0031	0.0003	0.0003	0.0004	0.0003	0.0003	0.0004
0.75	0.0138	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
0.50	0.0283	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008
0.25	0.0200	0.0023	0.0022	0.0025	0.0021	0.0020	0.0022
0.10	0.0056	0.0030	0.0027	0.0031	0.0028	0.0026	0.0029
x = 2							
0.90	0.0031	0.0003	0.0003	0.0004	0.0003	0.0003	0.0004
0.75	0.0138	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
0.50	0.0283	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008
0.25	0.0200	0.0023	0.0022	0.0025	0.0021	0.0020	0.0022
0.10	0.0056	0.0030	0.0027	0.0031	0.0028	0.0026	0.0029
x = 3							
0.90	0.0031	0.0003	0.0003	0.0004	0.0003	0.0003	0.0004
0.75	0.0138	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
0.50	0.0283	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008
0.25	0.0200	0.0023	0.0022	0.0025	0.0021	0.0020	0.0022
0.10	0.0056	0.0030	0.0027	0.0031	0.0028	0.0026	0.0029
x = 4							
0.90	0.0031	0.0003	0.0003	0.0004	0.0003	0.0003	0.0004
0.75	0.0138	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
0.50	0.0283	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008
0.25	0.0200	0.0023	0.0022	0.0025	0.0021	0.0020	0.0022
0.10	0.0056	0.0030	0.0027	0.0031	0.0028	0.0026	0.0029