Examples

2019-10-05

Introduction

We denote $Y_i, i = 1, ..., N$ are the independent, identically, distributed (iid) lifetimes, whose corresponding cumulative distriubtion 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 ovserved time is $Z_i = Y_i \wedge C_i$, whose CDF is H and PDF is h. The $\delta_i = I_{[T_i \leq C_i]}$ is the status indicator, which shows whether subject i is censored ($\delta_i = 0$) or not ($\delta_i = 0$). The corresponding hazard function of lifetime is λ_F and cumulative hazard function is Λ_F.

Instead of the strong assumption of independent between Y_i and C_i, we proposed that

$$P(C > t | T = t) = P(C > t | T > t) \tag{1}$$

under which, the Gerhard Dikta's model is still hold.

Given (Eq 1), we could derive that

$$P(\delta = 1 | X = t) = \frac{P(C > t, T = t)}{P(X = t)} = \frac{P(T = t)}{P(X = t)} \frac{P(C > t, T > t)}{P(T > t)} = \frac{f(t) S_x(t)}{h(t) S(t)} = \frac{\lambda_F(t)}{\lambda_H(t)}$$

where $\lambda_H(t)$ is the hazard function corresponding to Z, which is known as crude hazard rate as well.

We may define $m(t) = P(\delta = 1 | X = t) = E(\delta | X = t)$. Then

$$m(t) = \frac{\lambda_F(t)}{\lambda_H(t)} \tag{2}$$

which is the same parameter defined in Dikta's papers. Therefore, the independence between Y and C is not the necessory condition for equation (2).

We give several examples to support our conclusion.

Example 1

For a joint pdf function $f_{T1,T2}(t_1, t_2)$, if it equals to

$$f_{T1,T2}(x, y) = 16(x - \frac{1}{2})(y - \frac{1}{2})(x - y)(x - y + 1) + 1$$

Then we have survival function $S_{T_1,T_2} = P(T_1 > t_1, T_2 > t_2)$ as:

$$
\begin{aligned}
S_{T_1,T_2} =& P(T_1 > t_1, T_2 > t_2) = \int_{t_2}^{1}\int_{t_1}^{1} f_{T_1,T_2}(x,y)dxdy \\
=& \int_{t_2}^{1}\int_{t_1}^{1} f_{T_1,T_2}(x,y)dxdy \\
=& \int_{t_2}^{1}\int_{t_1}^{1} \Big[16(x-\frac{1}{2})(y-\frac{1}{2})(x-y)(x-y+1)+1\Big]dxdy \\
=& \int_{t_2}^{1} \Big\{4(y-\frac{1}{2})\Big[x^4-2x^3+(-2y^2+2y+1)x^2+(2y^2-2y)x\Big]+x\Big\}|_{t_1}^{1}dy \\
=& \int_{t_2}^{1} \Big\{(2-4y)t_1^4+(8y-4)t_1^3+(8y^3-12y^2+2)t_1^2+(-8y^3+12y^2-4y-1)t_1+1\Big\}dy \\
=& (t_1-1)y(2t_1y^3-4t_1y^2+(-2t_1^3+2t_1^2+2t_1)y+2t_1^3-2t_1^2-1)|_{t_2}^{1} \\
=& (1-t_1)(1-t_2)(1-2t_1t_2(t_2-t_1)(t_1+t_2-1))
\end{aligned}
$$

The marginal function for the survival time and censoring time are all uniform distributions:

$$
\begin{aligned}
f_{t_1}(x) =& \int_0^1 f_{t_1,t_2}(x,y)dy \\
=& \Big\{y-4(x-\frac{1}{2})(y^4-2y^3+(-2x^2+2x+1)y^2+(2x^2-2x)y)\Big\}|_0^1 \\
=& 1
\end{aligned}
$$

$$
\begin{aligned}
f_{t_2}(y) =& \int_0^1 f_{t_1,t_2}(x,y)dx \\
=& \Big\{4(y-\frac{1}{2})\Big[x^4-2x^3+(-2y^2+2y+1)x^2+(2y^2-2y)x\Big]+x\Big\}|_0^1 \\
=& 1
\end{aligned}
$$

That is,

$$
f_{T_1}(t_1) = I_{[0,1]}(t_1),\ f_{T_2}(t_2) = I_{[0,1]}(t_2)
$$
$$
P(T_1 > t_1) = 1 - t_1,\ P(T_2 > t_2) = 1 - t_2
$$

Therefore, the hazard rate function λ_F for the survival time is:

- $S_F(t) = 1 - t$, $\Lambda_F(t) = -log(1-t)$, $\lambda_F(t) = \frac{1}{1-t}$

The hazard rate function λ_H for the observed time is:

- $S_H(t) = P(Z > t) = (1-t)^2$, $\Lambda_H(t) = -2log(1-t), \lambda_H(t) = \frac{2}{1-t}$

Then

$$
m(t) = \frac{\lambda_F(t)}{\lambda_H(t)} = 0.5
$$

Let's make a simulation to show it works.

Data generation

T_2 is generated from the UNI(0,1).

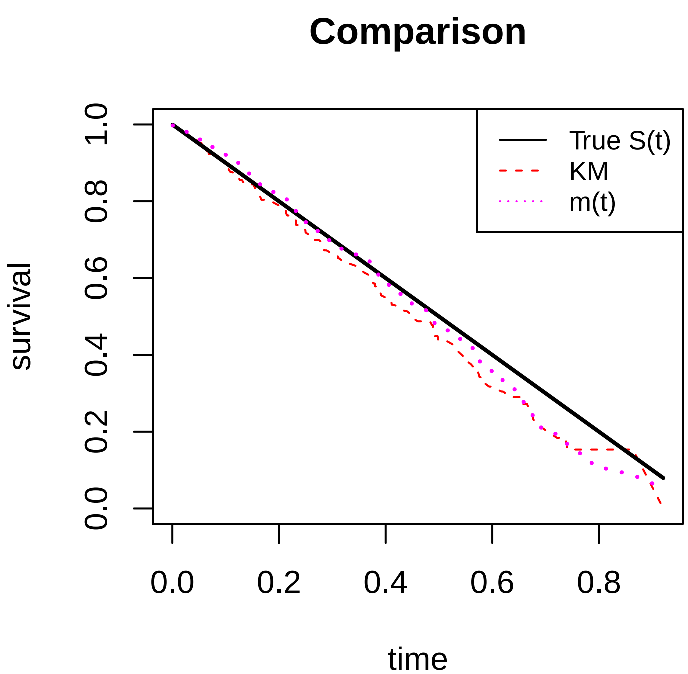
Given T_2, T_1 is generated from $f_{T_1|T_2}(x|y) = \frac{f_{T_1,T_2}(x,y)}{f_{T_2}(y)} = f_{T_1,T_2}(x,y)$, since $f_{T_2}(y) = 1$.

Then $F_{T_1|T_2}(x|y) = x((4y-2)x^3 + (4-8y)x^2 + (-8y^3 + 12y^2 - 2)x + 8y^3 - 12y^2 + 4y + 1)$. Then sample x by inverse probability sampling.

Results:

Censoring percentage: 52.5%

The KM estimator:



Bias:

Kaplan Meier:

```
mean(abs(fit_km$surv - Sx(fit_km$time)))
```

```
## [1] 0.03419431
```

Semi parametric model: $m(t) = \frac{\lambda_F(t)}{\lambda_H(t)}$

```
mean(abs(sest - Sx(fit_km$time)))
```

```
## [1] 0.02045551
```

If we do not know the $m(t)$ function, but know that it is a constant, i.e. $m(t;\theta) = \theta$,we many estimate the parameter by using the MLE:

$$L_n(\theta) = \prod_{i=1}^{n} m(\theta)^{\delta_i}(1-m(\theta))^{\delta_i}$$

The estimated value is $m(t) = 0.525$. The bias is

```
## [1] 0.0263961
```

Example 2

The other examples in paper are not usable.

A family of exponential example

$$f(x,y) = ae^{-x}e^{-y} + be^{-x-y} + c(2I_{y>x} - 1)e^{min(x,y)-max(x,y)}$$

That is

$$f(x,y) = \begin{cases} ae^{-x}e^{-y} + be^{-x-y} + cexp(x-y) & y \geq x \\ ae^{-x}e^{-y} + be^{-x-y} - cexp(y-x) & y < x \end{cases}$$

The marginal distribution is

$$\begin{aligned} f_x(x) =& \int_0^\infty f(x,y)dy \\ =& \int_0^x \Big(ae^{-x}e^{-y} + be^{-x-y} - cexp(y-x)\Big)dy + \\ & \int_x^\infty ae^{-x}e^{-y} + be^{-x-y} + cexp(x-y)dy \\ =& (a+b+c)e^{-x} - (a+b)e^{-2x} - c + c + (a+b)e^{-2x} \\ =& (a+b+c)e^{-x} \end{aligned}$$

$$\begin{aligned} f_y(y) =& \int_0^\infty f(x,y)dx \\ =& \int_y^\infty \Big(ae^{-x}e^{-y} + be^{-x-y} - cexp(y-x)\Big)dx + \\ & \int_0^y ae^{-x}e^{-y} + be^{-x-y} + cexp(x-y)dx \\ =& (-c + (a+b)e^{-2y}) + (c + (a+b-c)e^{-y} - (a+b)e^{-2y}) \\ =& (a+b-c)e^{-x} \end{aligned}$$

$f_{x,y}(t,t) = ae^{-2t}$ and $f_x(t)f_y(t) = (a+b-c)(a+b+c)e^{-2t}$

$\rightarrow a+b = (a+b-c)(a+b+c)$, that is

$$A = A^2 - C^2$$

We may let $A = 2, C = \sqrt{2}$ or other values