Simulation of Tsiatis's example

2019-07-10

Highlight of Tsiatis's paper

1. Tsiatis defined two survival functions:

- Net survival function: $H_i(t) = P(Y_i > t)$, which can be treated as the true survival function. The probability that Y_i exceeds time t.
- Crude survival fucntion: $Q_i(t) = P(Y_i > t, all(j \neq i)Y_j > Y_i)$, which can be treated as the observed survival function, can be estimated from the data. The subject survivals time t and die of reason Y_i
- When $Y_i s$ are not mutually independent, the net function and crude function are not identical.

2. Theorems:

- theorem 1: the drivative $Q_i(t)'$ equals the partial derivative of net survival function, wrt t_i

$$Q_i(t)' = \frac{\partial H^{(k)}}{\partial t_i}|_{t_j=t}$$

- theorem 2: with the independent assumption, the net function can be estimated from the crude function:

$$H_i^* = exp(\big[\int_0^t Q_i(x)'/\sum_i Q_i(x)dx\big])$$

3. Examples:

- In Tsiatis's example, the net survival functions are ($k = 2$):

$$H(t_1, t_2) = P(T_1 > t_1, T_2 > t_2) = exp(-\lambda t_1 - \mu t_2 - \theta t_1 t_2)$$
$$H_1(t) = P(T_1 > t) = exp(-\lambda t), H_2(t) = P(T_2 > t) = exp(-\mu t)$$

- The parameters are: $\lambda = 0.1, \mu = 0.2$, $\theta = 0.1$ and $\theta = 0.02$.

4. Conclusion

- When not independent, the $Q_i(t) \neq H_i(t)$
- Tsiatis's example has shown significant differences between those two functions.
- Our goal: Trying to use Tsiatis's example to show that Slud's equation may not work well, however, the correction of Slud's equation may work better.

Simulation

Formula

Slud's estimation of survival function $S_p(t)$:

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_i - 1}\}$$

The corrected version of Slud's equation:

$$\hat{S}_{p,corrected}(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1}(1 - \frac{\rho(X_i)}{n_i})\}$$

To calculate $\hat{S}_p(t)$ and $\hat{S}_{p,corrected}(t)$, we need to get the ρ value:

$$\rho(t) = \big[\{f(t)/\phi(t)\} - 1\big]\big[\{S(t)/S_X(t)\}\big]^{-1}$$

Therefore, we need the values: $f(t), S(t), S_x(t), \psi(t)$.

Let $T_1 = T$ be the survival time. Let $T_2 = C$ be the censor time. From Tsiatis's example, we know that:

- $f(t) = \lambda exp(-\lambda t)$
- $S(t) = exp(-\lambda t)$
- $S_x(t) = P(T > t, C > t) = exp(-\lambda t - \mu t - \theta t^2)$

And the joint CDF:

$$\begin{aligned}
1 =& P(T > t, C > c) + P(T > t, C < c) + P(T < t, C > c) + P(T < t, C < c) \\
=& P(T > t, C > c) + P(T > t) - P(T > t, C > c) + P(C > c) - P(T > t, C > c) + P(T < t, C < c) \\
=& P(T > t) + P(C > c) - P(T > t, C > c) + P(T < t, C < c)
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(T < t, C < c) =& 1 + P(T > t, C > c) - P(T > t) - P(C > c) \\
=& 1 + exp(-\lambda t - \mu c - \theta t c) - exp(-\lambda t) - exp(-\mu c)
\end{aligned}$$

Then

$$\begin{aligned}
f(t,c) =& \frac{\partial P(T < t, C < c)}{\partial t \partial c} = \frac{\partial\big[1 + exp(-\lambda t - \mu c - \theta t c) - exp(-\lambda t) - exp(-\mu c)\big]}{\partial t \partial c} \\
=& \frac{\partial\big[-(\lambda + \theta c)exp(-\lambda t - \mu c - \theta c t) + \lambda exp(-\lambda t)\big]}{\partial c} \\
=& (\lambda\mu - \theta + \lambda\theta t + \mu\theta c + \theta^2 t c)exp(-\lambda t - \mu c - \theta c t)
\end{aligned}$$

Then the $\psi(t)$

$$\begin{aligned}
\psi(t) =& \int_t^\infty f(t,c)dc = \int_t^\infty (\lambda\mu - \theta + \lambda\theta t + \mu\theta c + \theta^2 t c)exp(-\lambda t - \mu c - \theta c t)dc \\
=& (\lambda + \theta t)exp(-\lambda t - \mu t - \theta t^2)
\end{aligned}$$

Therefore, the functions in the examples are:

Function	Description	Expression
$P(T < t, C < c)$	Joint CDF	$1 + exp(-\lambda t - \mu c - \theta t c) - exp(-\lambda t) - exp(-\mu c)$
$f(t,c)$	Joint PDF	$(\lambda\mu - \theta + \lambda\theta t + \mu\theta c + \theta^2 t c)exp(-\lambda t - \mu c - \theta c t)$
$f_t(t)$	Marginal PDF of T	$\lambda exp(-\lambda t)$
$S_t(t)$	Survival function of T	$exp(-\lambda t)$
$f_c(c)$	Marginal PDF of C	$\mu exp(-\mu c)$
$S_c(c)$	$P_c(C > c)$	$exp(-\mu c)$
$S_x(t)$	$P(T > t, C > t)$	$exp(-\lambda t - \mu t - \theta t^2)$
$\psi(t)$	$\int_t^\infty f(t,c)dc$	$(\lambda + \theta t)exp(-\lambda t - \mu t - \theta t^2)$

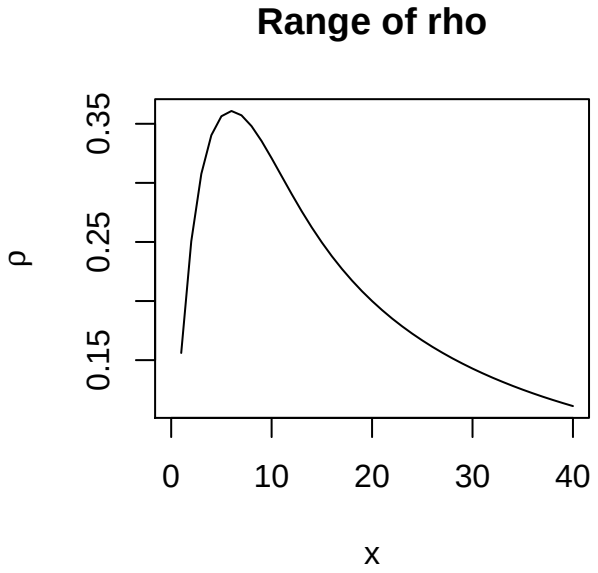
Parameter settings (consistent to Tsiatis's example):

- scenario 1: $\lambda = 0.1, \mu = 0.2, \theta = 0.02$
- scenario 2: $\lambda = 1, \mu = 1, \theta = 1$

Scenario 1:

ρ values

In this scenario, we could get a function of $\rho(t)$. The biggest value is around 0.37.

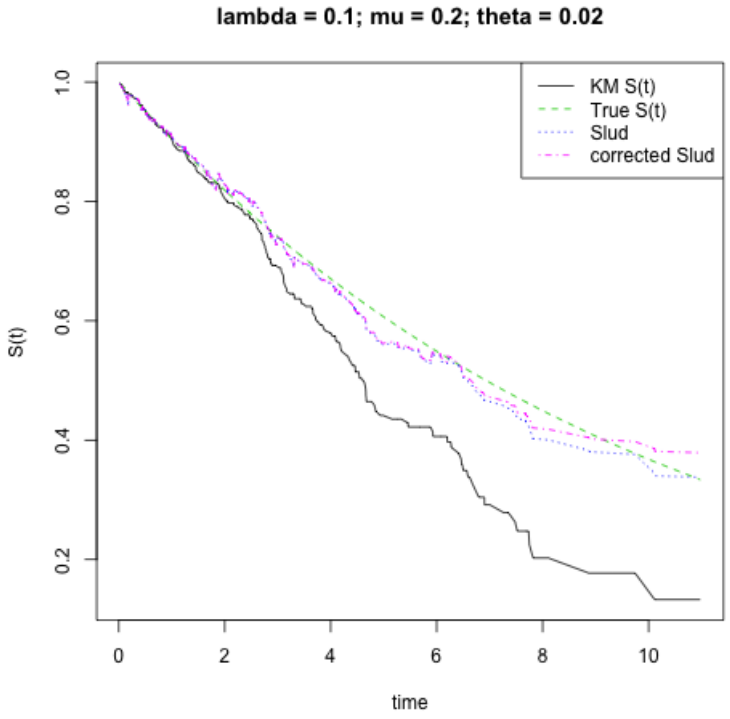


Generate dataset

The censoring rate is:

```
## [1] 0.406
```

The survival plot



Differences

The mean absolute difference between true value and the KM estimator (first 400 data):

```
## [1] 0.1465281
```

The mean absolute difference between true value and the Slud estimator (true $\rho(t)$ is used):

```
## [1] 0.009894739
```

The mean absolute difference between true value and the corrected Slud estimator (true $\rho(t)$ is used):

```
## [1] 0.00637633
```

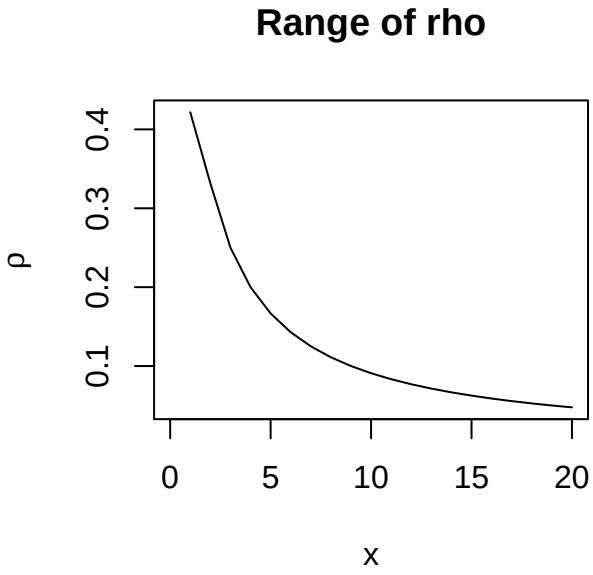
Conclusion

- In this example, the KM estimator does not work very well.
- The corrected Slud estimator works the best. However, it is only slightly better than the uncorrected estimator.

Scenario 2:

ρ values

In this scenario, we could get a function of $\rho(t)$. The biggest value is around 0.42.

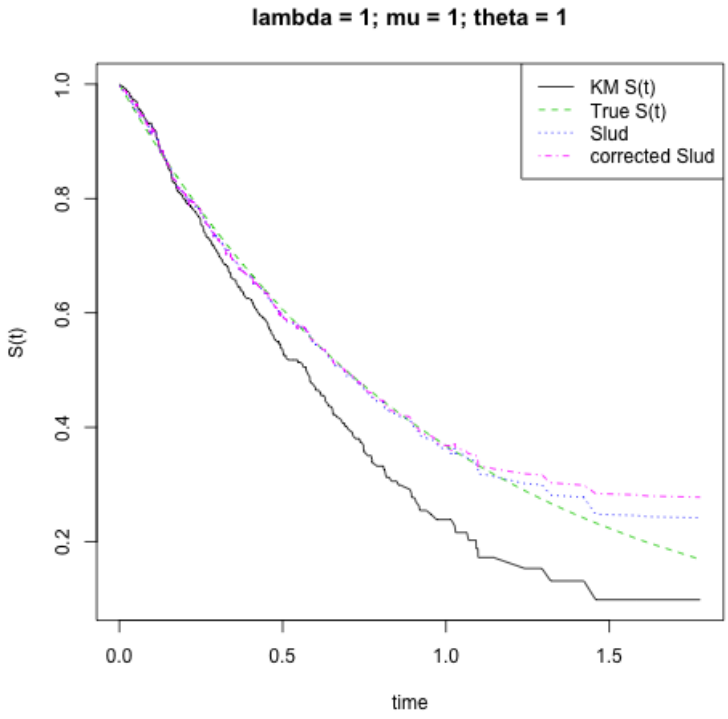


Generate dataset

The censoring rate is:

```
## [1] 0.498
```

The survival plot



Differences

The mean absolute difference between true value and the KM estimator (first 400 data):

```
## [1] 0.1036118
```

The mean absolute difference between true value and the Slud estimator (true $\rho(t)$ is used):

```
## [1] 0.003627692
```

The mean absolute difference between true value and the corrected Slud estimator (true $\rho(t)$ is used):

```
## [1] 0.00132013
```

Conclusion

- In this example, the KM estimator does not work very well.
- The corrected Slud estimator works the best. However, it is only slightly better than the uncorrected estimator.

Therefore, in Tsiatis's example, when the true ρ value is known, Slud's estimator can work well. Our corrected version of Slud estimator can slightly improve its performance. I think this maybe because that the true $\rho(t)$ values are relatively small (max less than 0.5). Therefore, there do not have big differences between those two equations:

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} (1 - \frac{\rho_i}{n_i + \rho_i - 1})\}$$

$$\hat{S}_{p,corrected}(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho(X_i)}{n_i})\}$$

Within those two equations, the different parts are:

$$part1 = \frac{1}{N} \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} (1 - \frac{\rho_i}{n_i + \rho_i - 1})$$

$$part2 = \frac{1}{N} c_{d(t)-1} + \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i})$$

$$\begin{aligned}
part2 - part1 =& \frac{1}{N} c_{d(t)-1} + \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i}) - \frac{1}{N} \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} (1 - \frac{\rho_i}{n_i + \rho_i - 1}) \\
=& \frac{1}{N} c_{d(t)-1} + \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i}) - \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i}) (\frac{1 - \frac{\rho_i}{n_i + \rho_i - 1}}{1 - \frac{\rho_i}{n_i}}) \\
& - \frac{1}{N} c_{d(t)-1} (1 - \frac{\rho_{d(t)-1}}{n_{d(t)-1} + \rho_{d(t)-1} - 1}) \\
=& \frac{1}{N} c_{d(t)-1} \frac{\rho_{d(t)-1}}{n_{d(t)-1} + \rho_{d(t)-1} - 1} + \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i}) \prod_{i=k+1}^{d(t)-1} (1 - \frac{1 - \frac{\rho_i}{n_i + \rho_i - 1}}{1 - \frac{\rho_i}{n_i}}) \text{ equation (*)}
\end{aligned}$$

If $\rho_i \leq 1$ (which is our case), $\frac{\rho_i}{n_i + \rho_i - 1} \geq \frac{\rho_i}{n_i}$

Then

$$\begin{aligned}
\text{equation (*)} \leq& \frac{1}{N} c_{d(t)-1} \frac{\rho_{d(t)-1}}{n_{d(t)-1} + \rho_{d(t)-1} - 1} + \frac{1}{N} \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho_i}{n_i}) A^{d(t)-1-k} \\
\leq& \frac{1}{N} c_{d(t)-1} \frac{\rho_{d(t)-1}}{n_{d(t)-1} + \rho_{d(t)-1} - 1} + (part2 - \frac{1}{N} c_{d(t)-1}) A \text{ equation (**)}
\end{aligned}$$

where A $= \max(1 - \frac{1-\frac{\rho_i}{n_i+\rho_i-1}}{1-\frac{\rho_i}{n_i}})$, $i \in [k+1, d(t)-1]$.

Then we need to find A, or $B = min(\frac{1-\frac{\rho_i}{n_i+\rho_i-1}}{1-\frac{\rho_i}{n_i}})$.

$$\frac{1-\frac{\rho_i}{n_i+\rho_i-1}}{1-\frac{\rho_i}{n_i}} = \frac{(n_i-1)n_i}{(n_i-1)n_i+\rho_i-\rho_i^2}$$

With fixed n_i, when $\rho_i = 0.5$, B get the min value; with fixed ρ_i, B is a monotune increasing function w.r.t n_i ($n_i \geq 1$).

Therefore, $A = \frac{(n_{d(t)-1}-1)n_{d(t)-1}}{(n_{d(t)-1}-1)n_{d(t)-1}+\rho_{0.5}-\rho_{0.5}^2}$, $\rho_{0.5}$ is the ρ value that closed to 0.5.

And equation (**) $\approx$ part2 A, since $c_{d(t)-1}/N$ is usually small.

Ususally, part2 is small (less than 0.01) and A is small and less than 1. Therefore, the difference between Slud's equation and corrected Slud's equation is relatively small (usually less than 0.01).

We could check the max difference value (the part2 A) from our simulation.

Function to calculate A:

```
A_part = function(t){
  N = dim(data)[1]
  dt = sum(data$time <= t & data$status == 1)
  nis = ni(dt - 1)
  res = 1 - (nis - 1) * nis / ((nis - 1) * nis + 0.5 - 0.5^2)
  return(res)
}
```

The max A value based on the dataset is 0.0059. Therefore, the Slud's equation and the corrected equation may not have big difference.

p.s. in previous simulation, the differences seems larger, since I did not contain the $\frac{1}{N}c_{d(t)-1}$

$$\hat{S}_{p,corrected}(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1}(1 - \frac{\rho(X_i)}{n_i})\}$$

Another question

I don't know whether I calculated the $f_{tc}(t,c)$ correct or not. Since when there is a setting $\lambda\mu - \theta < 0$, $f_{tc}(t,c)$ can smaller than 0.

$$f(t,c) = (\lambda\mu - \theta + \lambda\theta t + \mu\theta c + \theta^2 tc)exp(-\lambda t - \mu c - \theta ct)$$