Dr. Ying's Example, with parameter

2019-11-10

CDF with parameter

In Zhiliang Ying's paper, the Joint CDF is:

$$S(T \geq x, C \geq y) = \begin{cases} e^{-x} e^{-(e^y-1)\left((x-y)^2+1\right)} & x \geq y \\ e^{-x} e^{-(e^y-1)} & x < y \end{cases}$$

Let's add a parameter θ in the model:

$$S(T \geq x, C \geq y) = \begin{cases} e^{-\theta x} e^{-(e^{\theta y}-1)\left((\theta x-\theta y)^2+1\right)} & x \geq y \\ e^{-\theta x} e^{-(e^{\theta y}-1)} & x < y \end{cases}$$

PDF with parameter

Since

$$\begin{aligned} P(T \geq x, C \geq y) =& P(T \geq x) - P(T \geq x, C < y) \\ =& P(T \geq x) - (P(C < y) - P(C < y, T < x)) \\ =& P(T \geq x) + P(C \geq y) + P(C < y, T < x) - 1 \end{aligned}$$

Then

$$P(C < y, T < x) = 1 + P(T \geq x, C \geq y) - P(T \geq x) - P(C \geq y)$$

When $x \geq y$, the pdf is

$$\frac{\partial}{\partial x} P(C < y, T < x) = \theta e^{-\theta x} + \Big(-(e^{\theta y}-1)(2\theta^2 x - 2\theta^2 y) - \theta \Big) e^{-\theta x} e^{-(e^{\theta y}-1)\left((\theta x-\theta y)^2+1\right)}$$

$$\begin{aligned} & f_{T,C}(x,y) = \frac{\partial}{\partial x \partial y} P(C < y, T < x) \\ & = ((2\theta^2 x - 2\theta^2 y)(1-e^{\theta y}) - \theta)((2\theta^2 y - 2\theta^2 x)(1-e^{\theta y}) - \theta(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)e^{\theta y}) \times \\ & e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y}) - \theta x} \\ & + (\theta(2\theta^2 y - 2\theta^2 x)e^{\theta y} - 2\theta^2(1-e^{\theta y}))e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y}) - \theta x} \end{aligned}$$

When $x < y$, the pdf is

$$\frac{\partial}{\partial x} P(C < y, T < x) = \theta e^{-\theta x} - \theta e^{-\theta x} e^{-(e^{\theta y}-1)}$$

$$\frac{\partial}{\partial x \partial y} P(C < y, T < x) = \theta^2 e^{-e^{\theta y} + \theta y - \theta x + 1}$$

Therefore, the total pdf is

$$f_{T,C}(x,y) = \begin{cases} ((2\theta^2 x - 2\theta^2 y)(1-e^{\theta y}) - \theta)((2\theta^2 y - 2\theta^2 x)(1-e^{\theta y}) - \theta(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)e^{\theta y})\times \\ e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y}) - \theta x} \\ +(\theta(2\theta^2 y - 2\theta^2 x)e^{\theta y} - 2\theta^2(1-e^{\theta y}))e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y}) - \theta x} & x \geq y \\ \\ \theta^2 e^{-e^{\theta y} + \theta y - \theta x + 1} & x < y \end{cases}$$

$m()$ function, $\rho()$ function

Then

$$S_T(t) = P(T > t) = P(T > t, C > 0) = e^{-\theta t} e^{-(e^{\theta 0}-1)\left((t-0)^2+1\right)} = e^{-\theta t}$$
$$f_T(t) = \frac{\partial}{\partial t}(1 - S_T(t)) = \frac{\partial}{\partial t}(1 - e^{-\theta t}) = \theta e^{-\theta t}$$
$$S_x(t) = P(T > t, C > t) = e^{-\theta t} e^{-(e^{\theta t}-1)} = e^{-e^{\theta t} - \theta t + 1}$$
$$f_x(t) = \frac{\partial}{\partial t}(1 - S_x(t)) = 1 - e^{-e^{\theta t} - \theta t + 1} = \theta(1 + e^{\theta t})e^{-e^{\theta t} - \theta t + 1}$$
$$\psi(t) = \int_t^\infty f(t,c)dc = \int_t^\infty \theta^2 e^{-e^{\theta c} + \theta c - \theta t + 1} dc = \theta e^{-e^{\theta t} - \theta t + 1}$$

Therefore, the $m()$ function is:

$$m(t) = \frac{\lambda_F(t)}{\lambda_H(t)} = \frac{f_T(t)}{S_T(t)} / \frac{f_X(t)}{S_X(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} / \frac{\theta(1 + e^{\theta t})e^{-e^{\theta t} - \theta t + 1}}{e^{-e^{\theta t} - \theta t + 1}} = \frac{1}{1 + e^{\theta t}}$$

And for the $\rho()$ function,

$$\begin{aligned} \rho =& \frac{f(t)/\psi(t) - 1}{S(t)/S_x(t) - 1} \\ =& \frac{\theta e^{-\theta t}/(\theta e^{-e^{\theta t} - \theta t + 1}) - 1}{e^{-\theta t}/e^{-e^{\theta t} - \theta t + 1} - 1} \\ =& 1 \end{aligned}$$

Simulation

Data generation

Censoring percentage

$$\begin{aligned} P(T < C) =& \int_0^\infty \int_0^y \theta^2 e^{-e^{\theta y} + \theta y - \theta x + 1} dx dy \\ =& \int_0^\infty \theta(e^{\theta y} - 1)e^{-e^{\theta y} + 1} \\ \approx& 0.4 \end{aligned}$$

Conditional distribution

Since $f_t(x) = \theta e^{-\theta x}$,

- when $X < Y$:

$$f_{c|t}(y) = \frac{f_{t,c}(x,y)}{f_t(x)} = \theta^2 e^{-e^{\theta y}+\theta y-\theta x+1}/(\theta e^{-\theta x}) = \theta e^{-e^{\theta y}+\theta y+1}$$

$$F_{c|t}(x) = \int_0^x f_{c|t}(y)dy = \int_0^x \theta e^{-e^{\theta y}+\theta y+1}dy = 1 - e^{1-e^{\theta x}}$$

$$F_{c|t}^{-1}(x) = \frac{1}{\theta}\ln(1-\ln(1-x))$$

Which also means that, when $x < y$, the simulation of T and C can be independent.

- when $X \geq Y$:

$$\begin{aligned} f_{c|t}(y) =& \frac{f_{t,c}(x,y)}{f_t(x)} = \frac{f_{t,c}(x,y)}{\theta e^{-\theta x}} \\ =& ((2\theta x - 2\theta y)(1-e^{\theta y})-1)((2\theta^2 y - 2\theta^2 x)(1-e^{\theta y}) - \theta(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)e^{\theta y})\times \\ & e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y})} \\ & + ((2\theta^2 y - 2\theta^2 x)e^{\theta y} - 2\theta(1-e^{\theta y}))e^{(\theta^2 y^2 - 2\theta^2 xy + \theta^2 x^2 + 1)(1-e^{\theta y})} \end{aligned}$$

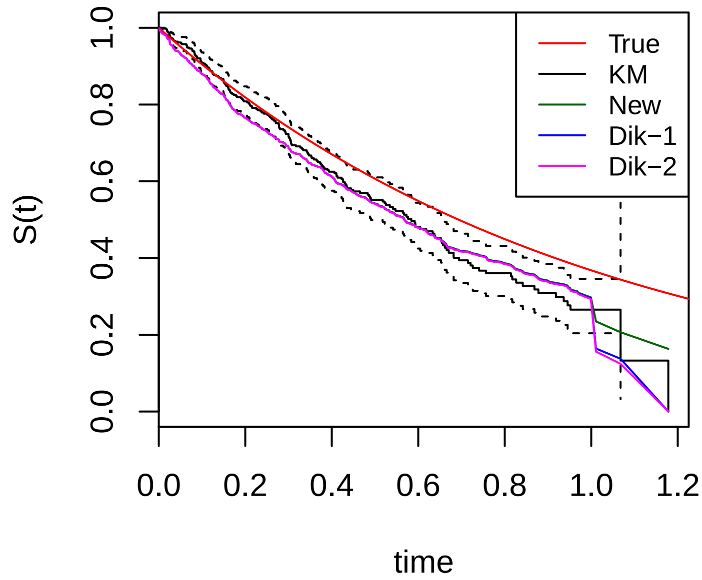
$$\begin{aligned} F_{c|t}(c) =& \int_0^c f_{c|t}(y)dy \\ =& 1 - \left((2e^{\theta c+1} - 2e)e^{\theta^2 c^2}\theta x + (-2\theta c e^{\theta c+1} + 2\theta e c + e)e^{\theta^2 c^2} e^{-e^{\theta c}\theta^2 x^2 + \theta^2 x^2 + 2\theta c e^{\theta c}\theta x - 2\theta^2 c x - \theta^2 c^2 e^{\theta c} - e^{\theta c}}\right) \end{aligned}$$

Result

When $\theta = 1$

```
theta = 1
```

The plot



mean absolute difference between true and KM

```
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03287375
```

mean absolute difference between true and new estiamte based on $m()$

```
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.04866821
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.05185118
```

mean absolute difference between true and Dikta formular 2 based on $m()$

```
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

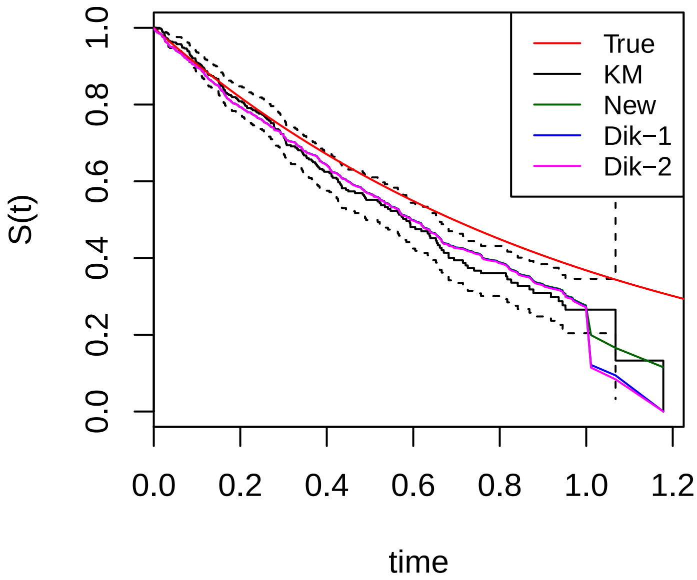
```
## [1] 0.05240246
```

If we do not know the value of m()

We may use logistic regression to estimate it

```
lg = glm(status ~ time,  data = data, family = "binomial")
lgres = predict(lg, type = 'response')
# use etimated m
m = function(t){
  ii = which(data$time == t)
  return(mean(lgres[ii]))
}
```

The plot



mean absolute difference between true and KM

```r
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03287375
```

mean absolute difference between true and new estiamte based on $m()$

```r
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.03146126
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```r
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.03477909
```

mean absolute difference between true and Dikta formular 2 based on $m()$

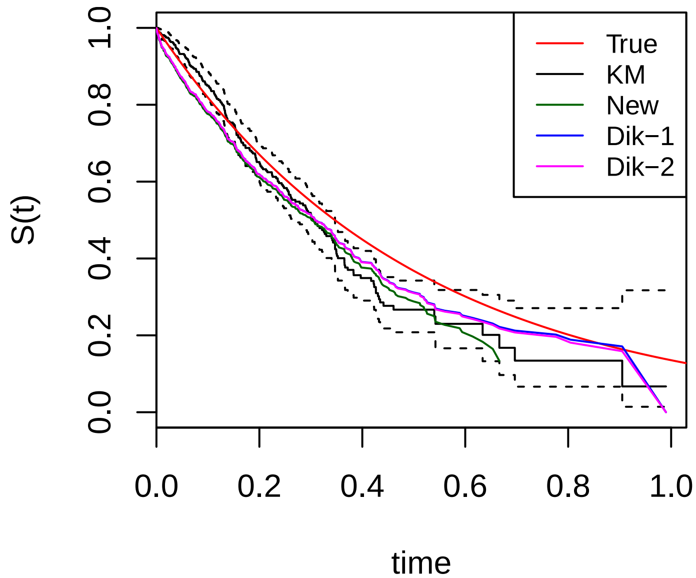
```r
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.03529491
```

When $\theta = 2$

```r
theta = 2
```

The plot



mean absolute difference between true and KM

```
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03421633
```

mean absolute difference between true and new estiamte based on $m()$

```
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.04668945
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.03861953
```

mean absolute difference between true and Dikta formular 2 based on $m()$

```
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

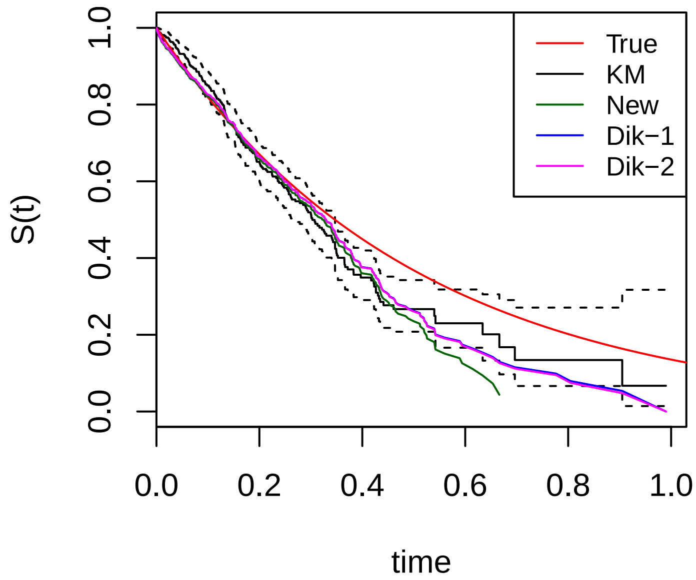
```
## [1] 0.03898382
```

If we do not know the value of m()

We may use logistic regression to estimate it

```
lg = glm(status ~ time,  data = data, family = "binomial")
lgres = predict(lg, type = 'response')
# use etimated m
m = function(t){
  ii = which(data$time == t)
  return(mean(lgres[ii]))
}
```

The plot



mean absolute difference between true and KM

```
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03421633
```

mean absolute difference between true and new estiamte based on $m()$

```
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.02344271
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.01927498
```

mean absolute difference between true and Dikta formular 2 based on $m()$

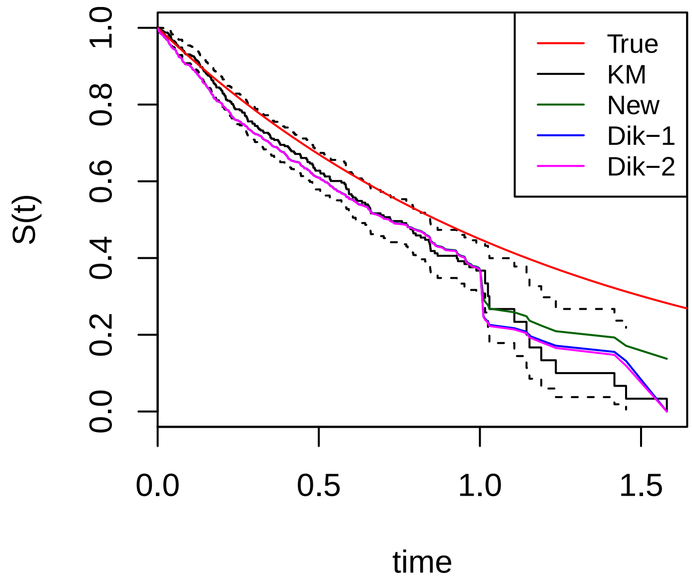
```
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.01955071
```

When $\theta = 0.8$

```
theta = 0.8
```

The plot



mean absolute difference between true and KM

```
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03596879
```

mean absolute difference between true and new estiamte based on $m()$

```
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.05462399
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.05854623
```

mean absolute difference between true and Dikta formular 2 based on $m()$

```
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

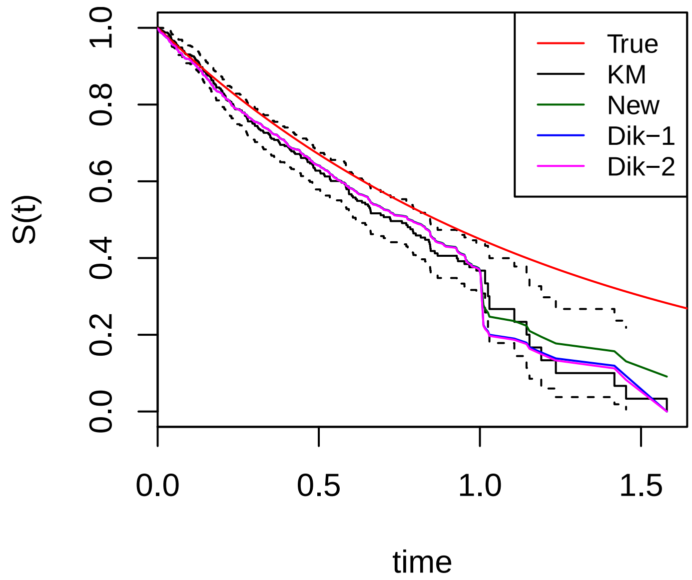
```
## [1] 0.05899001
```

If we do not know the value of m()

We may use logistic regression to estimate it

```
lg = glm(status ~ time,  data = data, family = "binomial")
lgres = predict(lg, type = 'response')
# use etimated m
m = function(t){
  ii = which(data$time == t)
  return(mean(lgres[ii]))
}
```

The plot



mean absolute difference between true and KM

```
mean(abs(fit_km$surv - exp(-theta *fit_km$time)), na.rm = TRUE)
```

```
## [1] 0.03596879
```

mean absolute difference between true and new estiamte based on $m()$

```
mean(abs(res1 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.03654611
```

mean absolute difference between true and Dikta formular 1 based on $m()$

```
mean(abs(res2 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.04076176
```

mean absolute difference between true and Dikta formular 2 based on $m()$

```
mean(abs(res3 - exp(-theta *data$time)), na.rm = TRUE)
```

```
## [1] 0.04120029
```