Example 4- Exponential and Weibull

Consider the joint distribution who has:

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

Then

- $S_T(x) = P(T \geq x, C \geq 0) = S(x,0) = e^{-\theta x}$, $f_T(x) = \frac{1 - S_T(x)}{x} = \theta e^{-\theta x}$
- $S_C(x) = P(T \geq 0, C \geq x) = S(0,x) = e^{-\theta 0} e^{-(\theta x)^k} = e^{-(\theta x)^k}$, $f_C(x) = \frac{1 - S_C(x)}{x} = k\theta(\theta y)^{k-1} e^{-(\theta x)^k}$

The death time is from an exponential distribution with paramter θ, the censor time is from a Weibull distribution with shape parameter k and scale parameter $1/\theta$.

Beisdes,

- $S_Z(x) = P(T > x, C > x) = e^{-\theta x - (\theta x)^k}$, $f_Z(x) = (\theta + k\theta(\theta x)^{k-1}) e^{-\theta x - (\theta x)^k}$

Therefore the $m()$ function is

$$m(x) = \frac{f_T(x)/S_T(x)}{f_Z(X)/S_Z(x)} = \frac{\theta e^{-\theta x} / e^{-\theta x}}{(\theta + k\theta(\theta x)^{k-1}) e^{-\theta x - (\theta x)^k} / e^{-\theta x - (\theta x)^k}} = \frac{1}{1 + k(\theta x)^{k-1}}$$

We could also transform $m()$ function as:

$$m(x) = \frac{1}{1 + \exp(\log(k(\theta x)^{k-1}))} = \frac{1}{1 + \exp(\log(k) + (k-1)\log(\theta) + (k-1)\log(x))}$$

We can then estimate the k and θ by fitting logistic regression.

And when $x < y$, the pdf is

$$f_{T,C}(x,y) = \frac{\partial^2}{\partial x \partial y} S(x,y) = \frac{\partial^2}{\partial x \partial y} e^{-\theta x} e^{-(\theta y)^k} = k\theta^{k+1} y^{k-1} \exp(-\theta x - (\theta y)^k)$$

Then

$$\begin{aligned} \psi(x) &= \int_x^\infty f_{T,C}(x,y) dy \\ &= \int_x^\infty k\theta^{k+1} y^{k-1} \exp(-\theta x - (\theta y)^k) dy \\ &= \theta \exp(-\theta x - (\theta x)^k) \end{aligned}$$

The $\rho(x)$ is:

$$\rho(x) = \frac{f_T(x)/\psi(x) - 1}{S(x)/S_Z(x) - 1} = \frac{\theta \exp(-\theta x)/(\theta \exp(-\theta x - (\theta x)^k)) - 1}{\exp(-\theta x)/exp(-\theta x - (\theta x)^k) - 1} = 1$$

Data generation

When $x < y$, (death time < censor time), the death time is generated from an exponentaion with parameter θ, the censor time is independently generated from a Weibull distribution. The pairs that $x < y$ are selected.

When $x > y$, firstly generate death time an exponentaion with parameter θ. Given a known death time, generate censor time from the conditional distribution:

$$
\begin{aligned}
F_{C|T}(y|x) = &\int_0^y f_{C|T}(v|x)dv \\
= &\int_0^y \frac{\partial}{\partial y}\frac{\partial F_{T,C}(x,y)}{\partial x f_T(x)}|_{y=v}dv \\
= &\frac{\partial F_{T,C}(x,y)}{\partial x f_T(x)}|_0^y \\
= &(2\theta(\theta y)^k(y-x)-1)\exp(-(\theta y)^k\Big((\theta x-\theta y)^2+1\Big))|_0^y \\
= &1-(2\theta(\theta y)^k(x-y)+1)\exp(-(\theta y)^k\Big((\theta x-\theta y)^2+1\Big))
\end{aligned}
$$

The censoring percentage

$$
\begin{aligned}
P(T < C) = &\int_0^\infty \int_0^y f_{T.C}(x,y)dxdy = \int_0^\infty \int_0^y k\theta^{k+1}y^{k-1}e^{-\theta x-(\theta y)^k}dxdy \\
= &\int_0^\infty k\theta^k y^{k-1}(1-e^{-\theta y})e^{-(\theta y)^k}dy
\end{aligned}
$$

To make things easy, we can set $k = 2$, then

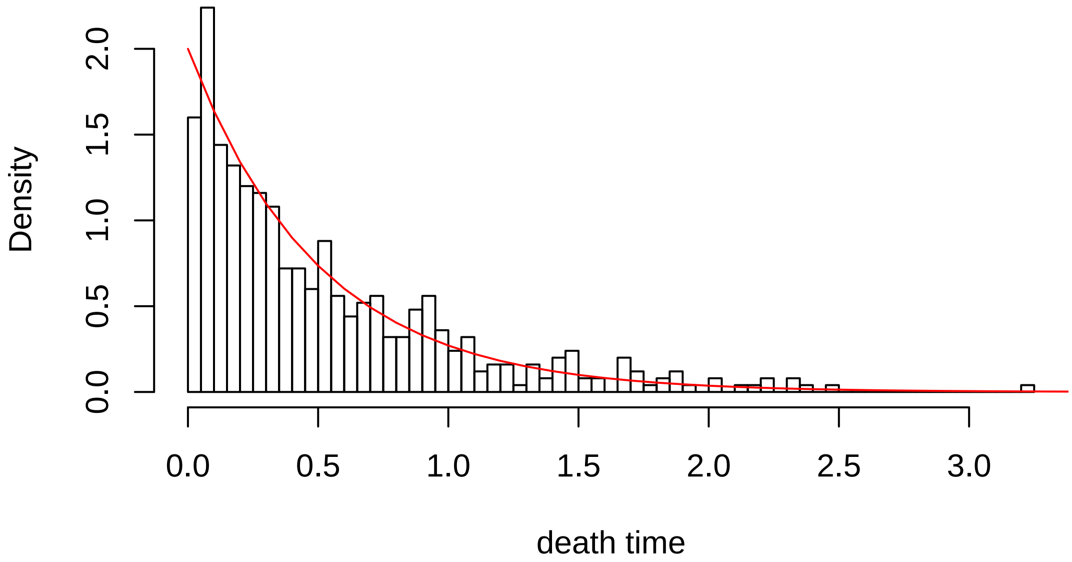
$$
\begin{aligned}
P(T < C) = &\int_0^\infty 2\theta^2 y(1-e^{-\theta y})e^{-(\theta y)^2}dy \\
= &\int_0^\infty 2\theta^2 ye^{-(\theta y)^2} - 2\theta^2 ye^{-\theta y-(\theta y)^2}dy \\
= &-e^{-(\theta y)^2}|_0^\infty - \int_0^\infty 2\theta^2 ye^{-\theta y-(\theta y)^2}dy \\
\approx &0.546
\end{aligned}
$$

Data generation, check margnial

For example, when theta = 1,

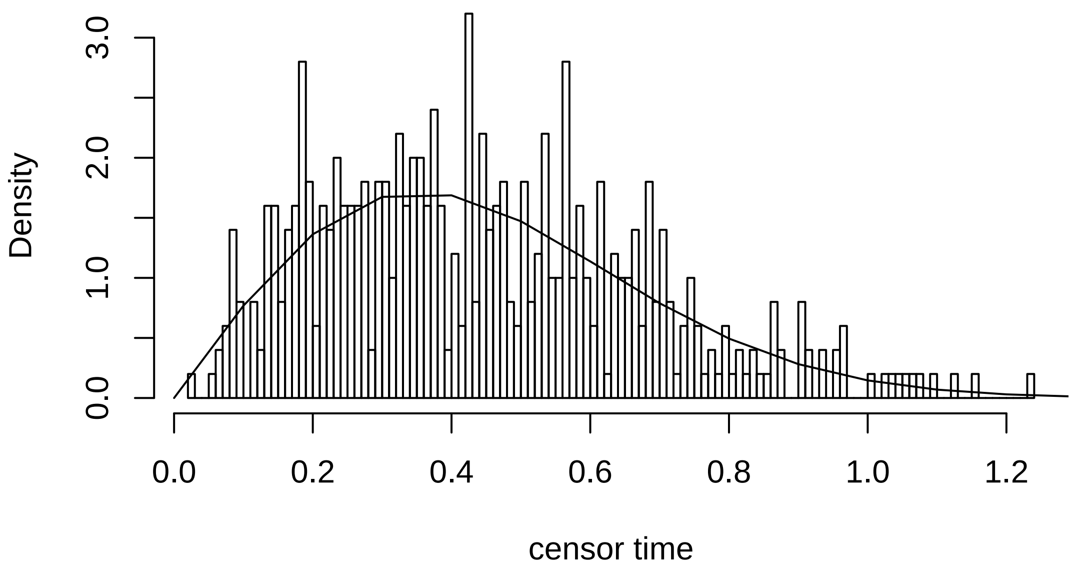
- Histogram of T estimation

The red line is the true density of T



- Histogram of C estimation

The red line is the true density of C



Results

Estimation of θ by logistic regression

	1	2
Estimated intercept	0.4197	1.5544

	1	2
Estimated slope	0.8658	1.1409

Estimation of $m()$ by logistic regression

Plug in the estimated $\hat{\theta}$ to get the estimated $\hat{m}(t)$. Calculate the mean absolute difference between true $m(t)$ and $\hat{m}(t)$.

	1	2
|hat m(t) - m(t)|	0.0399	0.0208

Estimation of $S(t)$ with true $m(t)$

KM	new m()	Dikta1	Dikta2
0.0457	0.0475	0.0482	0.0485
0.0136	0.0115	0.0116	0.0117

Estimation of $S(t)$ with logistic regression estimated $m(t)$

KM	new m()	Dikta1	Dikta2
0.0457	0.0483	0.0491	0.0494
0.0136	0.0122	0.0124	0.0124

Variances of the quantiles of $S(t)$

Let's look at the variance of the estimation: S(t) at the '10th', '20th', '50th', '125th', '250th', '325th', '400th' out of 500 subjects.

The following table shows the fraction of standard deviation of new methods over Kaplan Meier estiamter: $\frac{v}{v_{km}}$.

1. Estimated by $\lambda_F(t) = m(t)\lambda_H(t)$ ($\frac{v}{v_{km}}$)

```
## Warning in rep(linesep, length.out = nrow(x) - 2): 'x' is NULL so the
## result will be NULL
```

With true m()		Logistic estimated m()	
1	2	1	2

10th	0.5926	0.2560	0.6115	0.2571
20th	0.7046	0.4397	0.7376	0.4415
50th	0.6925	0.7202	0.7259	0.7261
125th	0.7565	0.7169	0.8096	0.7257
250th	0.6932	0.6813	0.8140	0.6989
325th	0.6124	0.4550	0.7632	0.4887
400th	0.7232	0.3356	0.8465	0.3881

2. Estimated by Dikta's 1st formula ($\frac{v}{v_{km}}$)

```
## Warning in rep(linesep, length.out = nrow(x) - 2): 'x' is NULL so the
## result will be NULL
```

	With true m()		Logistic estimated m()	
	1	2	1	2
10th	0.5935	0.2563	0.6124	0.2574
20th	0.7056	0.4402	0.7386	0.4420
50th	0.6934	0.7209	0.7268	0.7268
125th	0.7571	0.7185	0.8103	0.7266
250th	0.6957	0.6818	0.8152	0.6998
325th	0.6143	0.4525	0.7644	0.4869
400th	0.7265	0.3264	0.8499	0.3783

3. Estimated by Dikta's 2nd formula ($\frac{v}{v_{km}}$)

```
## Warning in rep(linesep, length.out = nrow(x) - 2): 'x' is NULL so the
## result will be NULL
```

	With true m()		Logistic estimated m()	
	1	2	1	2
10th	0.5932	0.2561	0.6120	0.2572
20th	0.7051	0.4398	0.7381	0.4416
50th	0.6929	0.7203	0.7264	0.7262
125th	0.7567	0.7179	0.8099	0.7260
250th	0.6957	0.6813	0.8152	0.6993
325th	0.6148	0.4522	0.7648	0.4866
400th	0.7285	0.3271	0.8517	0.3789

Plots of the variances