Semi-parameter model estimation, with quantiles

2019-11-25

Changes

- use inverse function in R package instead of using my own function
- calculated the estimated $\hat{S}(t)$ at time quantiles (e.g. $t_{0.1}, t_{0.25}, t_{0.5}, t_{0.75}, t_{0.9}$), instead of using order statistics of time.

Results highlight

- Four methods: KM: Kaplan Meier; Exp m(): estimate λ_F with $\exp(m(t)\lambda_H)$; Dikta 1: use Dikta's first formula; Dikta 2: use Dikta's updated formula.
- The 4 methods have similar mean absolute difference between $\hat{S}(t)$ and true $S(t)$
- The standard deviation of KM is larger than the other three methods.
- The other three methods have similar standard deviation over 500 iterations.
- When using the logistic regression estimated $m(t)$, the standard deviation get larger, but still smaller than the Kaplan Meier's sd

Simulation

All the simulations below have generated data with size 500. And 500 times iterations.

Example 3: exponential + extreme distribution

The survival time is denoted as T and censor time as C, the observed time is marked as Z

And

$$S_T(t) = P(T > t) = P(T > t, C > 0) = e^{-\theta t} e^{-(e^{\theta 0}-1)\left((t-0)^2+1\right)} = e^{-\theta t}$$

$$f_T(t) = \frac{\partial}{\partial t}(1 - S_T(t)) = \frac{\partial}{\partial t}(1 - e^{-\theta t}) = \theta e^{-\theta t}$$

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

$$f_Z(t) = \frac{\partial}{\partial t}(1 - S_x(t)) = 1 - e^{-e^{\theta t}-\theta t+1} = \theta(1 + e^{\theta t})e^{-e^{\theta t}-\theta t+1}$$

$$\psi(t) = \int_t^\infty f(t,c)dc = \int_t^\infty \theta^2 e^{-e^{\theta c}+\theta c-\theta t+1} dc = \theta e^{-e^{\theta t}-\theta t+1}$$

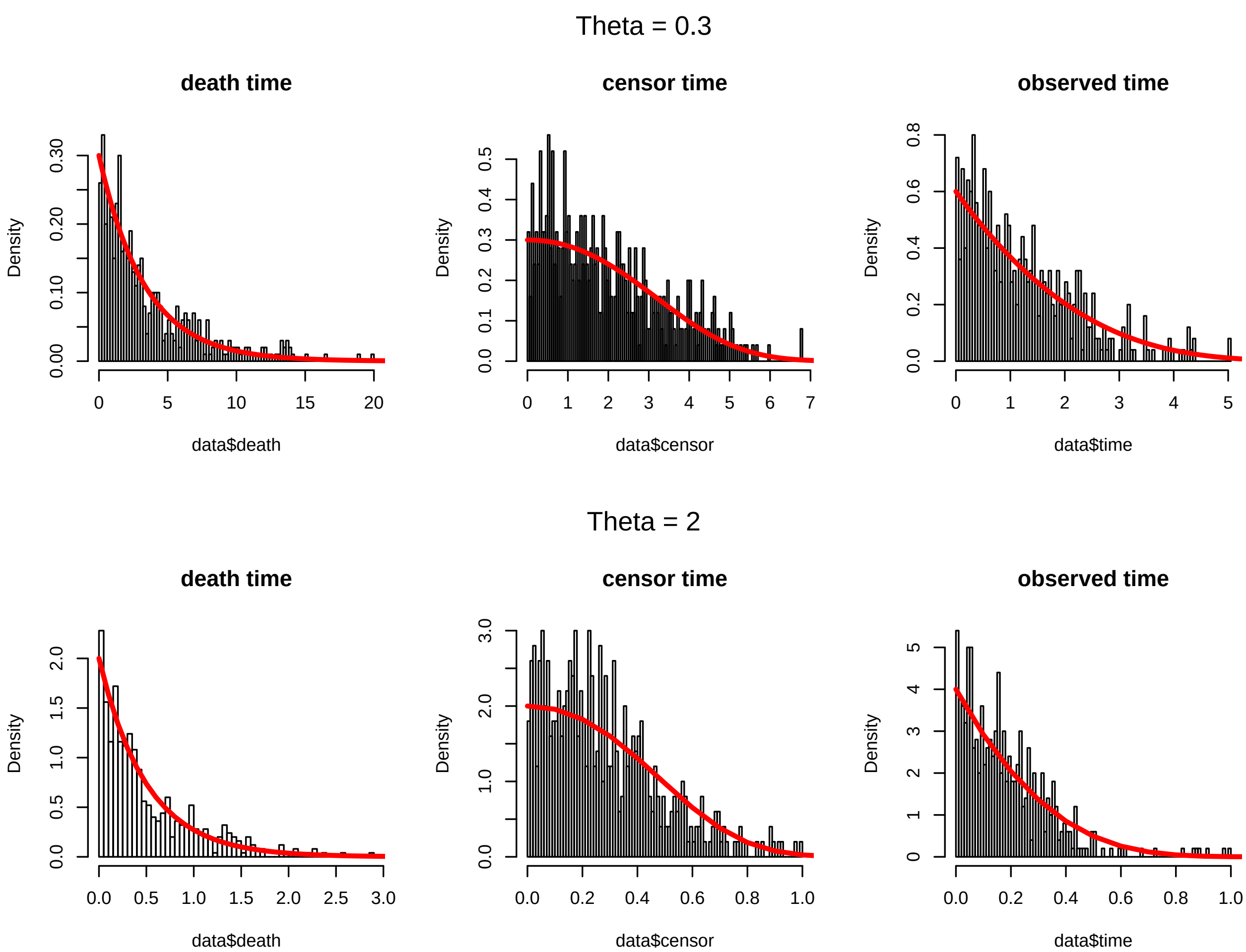
Therefore, the $m()$ function is:

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

And for the $\rho()$ function,

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

After changing the inverse function, the data is generated more appropriately. For example:



Results

Table 1: Mean absolute difference between estimated and true S()

	With true m()				With estimated m()			
Quantile	KM	Exp m()	Dikta 1	Dikta 2	KM	Exp m()	Dikta 1	Dikta 2
theta = 0.3								
t0.1	0.01071	0.00956	0.00962	0.00965	0.01071	0.00976	0.00982	0.00985
t0.25	0.01460	0.02100	0.02128	0.02143	0.01460	0.02346	0.02376	0.02392
t0.5	0.04125	0.06398	0.06511	0.06580	0.04125	0.07632	0.07746	0.07815
t0.75	0.10133	0.04082	0.05102	0.05665	0.10133	0.06283	0.07307	0.07850
t0.9	0.05235	0.05139	0.09990	0.09986	0.05235	0.02789	0.10000	0.09996
theta = 0.8								
t0.1	0.01001	0.00846	0.00850	0.00852	0.01001	0.00853	0.00857	0.00859
t0.25	0.01416	0.01863	0.01888	0.01901	0.01416	0.02015	0.02041	0.02055
t0.5	0.03326	0.05150	0.05255	0.05321	0.03326	0.06035	0.06142	0.06207
t0.75	0.09109	0.04377	0.05576	0.06180	0.09109	0.06079	0.07236	0.07823
t0.9	0.05912	0.05143	0.09992	0.09989	0.05912	0.03319	0.09989	0.09985
theta = 1								
t0.1	0.00947	0.00826	0.00831	0.00834	0.00947	0.00839	0.00844	0.00847
t0.25	0.01380	0.01903	0.01928	0.01942	0.01380	0.02075	0.02102	0.02116
t0.5	0.03366	0.05310	0.05417	0.05484	0.03366	0.06186	0.06294	0.06359
t0.75	0.09109	0.04340	0.05253	0.05865	0.09109	0.05979	0.06886	0.07481
t0.9	0.05637	0.05244	0.10000	0.10000	0.05637	0.03446	0.10000	0.10000
theta = 1.5								
t0.1	0.00986	0.00811	0.00813	0.00814	0.00986	0.00813	0.00816	0.00817
t0.25	0.01415	0.01674	0.01695	0.01707	0.01415	0.01795	0.01818	0.01831
t0.5	0.03170	0.04993	0.05097	0.05163	0.03170	0.05853	0.05958	0.06023
t0.75	0.06719	0.01941	0.02279	0.02549	0.06719	0.03227	0.03648	0.03962
t0.9	0.05985	0.05800	0.09995	0.09991	0.05985	0.03937	0.09981	0.09978
theta = 2								
t0.1	0.01000	0.00772	0.00774	0.00775	0.01000	0.00776	0.00779	0.00780
t0.25	0.01425	0.01652	0.01674	0.01686	0.01425	0.01779	0.01803	0.01816
t0.5	0.03329	0.05195	0.05302	0.05367	0.03329	0.06001	0.06110	0.06176
t0.75	0.05248	0.01777	0.01798	0.01860	0.05248	0.02427	0.02601	0.02751
t0.9	0.07273	0.08153	0.10000	0.09997	0.07273	0.06150	0.09987	0.09983
theta = 5								
t0.1	0.01083	0.01598	0.01608	0.01613	0.01083	0.01687	0.01697	0.01703
t0.25	0.01591	0.03690	0.03722	0.03741	0.01591	0.04328	0.04361	0.04380
t0.5	0.07241	0.10069	0.10195	0.10278	0.07241	0.12810	0.12941	0.13020
t0.75	0.18745	0.06923	0.08175	0.08904	0.18745	0.11240	0.12385	0.13007
t0.9	0.06338	0.06325	0.06141	0.05180	0.06338	0.02267	0.02741	0.02676

Table 2: Standard deviations of the estimated S()

	With true m()				With estimated m()			
Quantile	KM	Exp m()	Dikta 1	Dikta 2	KM	Exp m()	Dikta 1	Dikta 2
theta = 0.3								

t0.1	0.01318	0.01040	0.01042	0.01042	0.01318	0.01040	0.01042	0.01042
t0.25	0.01768	0.01430	0.01432	0.01433	0.01768	0.01431	0.01433	0.01435
t0.5	0.02559	0.02005	0.02015	0.02021	0.02559	0.02166	0.02175	0.02180
t0.75	0.04435	0.02678	0.04130	0.04231	0.04435	0.02961	0.04100	0.04172
t0.9	0.06024	0.00853	0.00844	0.00776	0.06024	0.02219	0.01039	0.00969
theta = 0.8								
t0.1	0.01256	0.00968	0.00969	0.00970	0.01256	0.00969	0.00970	0.00970
t0.25	0.01760	0.01540	0.01543	0.01544	0.01760	0.01541	0.01544	0.01545
t0.5	0.02290	0.01909	0.01917	0.01923	0.02290	0.01978	0.01986	0.01990
t0.75	0.05050	0.02613	0.04558	0.04598	0.05050	0.03048	0.04555	0.04580
t0.9	0.06710	0.00800	0.00885	0.00819	0.06710	0.02131	0.00818	0.00754
theta = 1								
t0.1	0.01178	0.00934	0.00935	0.00935	0.01178	0.00935	0.00936	0.00936
t0.25	0.01710	0.01493	0.01495	0.01496	0.01710	0.01507	0.01509	0.01510
t0.5	0.02282	0.01927	0.01936	0.01942	0.02282	0.02049	0.02057	0.02062
t0.75	0.05134	0.02637	0.03997	0.04102	0.05134	0.03117	0.04155	0.04230
t0.9	0.06793	0.00856	0.00000	0.00000	0.06793	0.02315	0.00000	0.00000
theta = 1.5								
t0.1	0.01245	0.00989	0.00990	0.00990	0.01245	0.00988	0.00989	0.00990
t0.25	0.01750	0.01474	0.01475	0.01476	0.01750	0.01460	0.01461	0.01463
t0.5	0.02505	0.01981	0.01990	0.01995	0.02505	0.02039	0.02047	0.02052
t0.75	0.04198	0.02202	0.02550	0.02639	0.04198	0.02685	0.02949	0.03010
t0.9	0.07040	0.00714	0.00922	0.00852	0.07040	0.02248	0.00653	0.00591
theta = 2								
t0.1	0.01227	0.00935	0.00936	0.00937	0.01227	0.00935	0.00936	0.00937
t0.25	0.01760	0.01468	0.01470	0.01472	0.01760	0.01467	0.01469	0.01471
t0.5	0.02349	0.01913	0.01922	0.01927	0.02349	0.01987	0.01995	0.02000
t0.75	0.03840	0.02193	0.02295	0.02379	0.03840	0.02811	0.02892	0.02952
t0.9	0.04559	0.01157	0.01027	0.00965	0.04559	0.02698	0.00763	0.00701
theta = 5								
t0.1	0.01357	0.01039	0.01040	0.01040	0.01357	0.01047	0.01048	0.01048
t0.25	0.01796	0.01552	0.01554	0.01555	0.01796	0.01592	0.01593	0.01595
t0.5	0.02835	0.02001	0.02014	0.02022	0.02835	0.02056	0.02066	0.02071
t0.75	0.02984	0.02156	0.03704	0.03758	0.02984	0.02336	0.03257	0.03285
t0.9	0.02260	0.01651	0.05320	0.05011	0.02260	0.02168	0.04103	0.03849

The estimate of $m()$

Table 3: mean absolute difference between hat m() and true m()

0.3	0.8	1	1.5	2	5
0.0142715	0.0118019	0.012252	0.0116686	0.0113792	0.0266302

The row name shows the θ value

Table 4: standard deviation of estimated m()

0.3	0.8	1	1.5	2	5
0.009696	0.0083783	0.0084335	0.0083854	0.0083391	0.0110993

The row name shows the θ value

Table 5: estimated theta from logitic regression

0.3	0.8	1	1.5	2	5
0.2538652	0.7098642	0.8885693	1.337076	1.794715	3.311591

The row name shows the true θ value

Example 4: exponential + weibull distribution

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

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.

Results

Table 6: Mean absolute difference between estimated and true S()

	With true m()				With estimated m()			
Quantile	KM	Exp m()	Dikta 1	Dikta 2	KM	Exp m()	Dikta 1	Dikta 2
theta = 1								
t0.1	0.01433	0.01276	0.01281	0.01282	0.01433	0.01234	0.01239	0.01240
t0.25	0.02735	0.02253	0.02279	0.02284	0.02735	0.02368	0.02394	0.02399
t0.5	0.05885	0.04872	0.04944	0.04977	0.05885	0.06795	0.06868	0.06901
t0.75	0.18087	0.09472	0.10128	0.10543	0.18087	0.12850	0.13453	0.13802
t0.9	0.09934	0.00619	0.09649	0.09619	0.09934	0.03384	0.09689	0.09719
theta = 2								
t0.1	0.01085	0.01061	0.01059	0.01059	0.01085	0.01059	0.01059	0.01059
t0.25	0.01295	0.01280	0.01274	0.01273	0.01295	0.01238	0.01241	0.01242
t0.5	0.01333	0.01345	0.01340	0.01339	0.01333	0.01156	0.01169	0.01178
t0.75	0.02056	0.01573	0.01545	0.01547	0.02056	0.01769	0.01728	0.01706
t0.9	0.05968	0.01864	0.09742	0.09697	0.05968	0.03254	0.09762	0.09724

Table 7: Standard deviations of the estimated S()

	With true m()				With estimated m()			
Quantile	KM	Exp m()	Dikta 1	Dikta 2	KM	Exp m()	Dikta 1	Dikta 2
theta = 1								
t0.1	0.01475	0.01409	0.01411	0.01411	0.01475	0.01417	0.01418	0.01419
t0.25	0.01825	0.01694	0.01697	0.01697	0.01825	0.01741	0.01743	0.01744
t0.5	0.02041	0.01986	0.01993	0.01995	0.02041	0.01822	0.01829	0.01831
t0.75	0.02947	0.02235	0.02852	0.02945	0.02947	0.01954	0.02412	0.02473
t0.9	0.00341	0.00733	0.02207	0.02012	0.00341	0.01146	0.01541	0.01390
theta = 2								
t0.1	0.01341	0.01263	0.01265	0.01265	0.01341	0.01306	0.01307	0.01307
t0.25	0.01620	0.01537	0.01538	0.01539	0.01620	0.01569	0.01570	0.01571
t0.5	0.01650	0.01671	0.01669	0.01671	0.01650	0.01345	0.01343	0.01344
t0.75	0.02582	0.01904	0.01906	0.01929	0.02582	0.01984	0.01982	0.02002
t0.9	0.06723	0.00757	0.02660	0.02443	0.06723	0.01955	0.02779	0.02567

The estimate of $m()$

Table 8: mean absolute difference between hat m() and true m()

1	2
0.0398907	0.0207512

The colname shows the true θ value.

Table 9: standard deviation of estimated m()

1	2
0.0116082	0.0108452

The colname shows the true θ value.

Table 10: estimated k from logitic regression (true k = 2)

1	2
1.865846	2.140893

The colname shows the true θ value.

Table 11: estimated theta from logitic regression

1	2
0.5951885	1.663312

The colname shows the true θ value.

Table 12: estimated theta from logitic regression with true k

1	2
0.7664011	2.414013

The colname shows the true θ value.