HW 02 - Colin Busby BSTA 662

Sunday, February 12, 2023 12:18 PM

662_HW2-S2023

BSTA 662 Survival Analysis **HOMEWORK TWO**

Topic: Censoring, Truncation, and Non-parametric Estimation/Tests

Reading: Klein 2nd ed., Ch 4.1 to 4.3, 5.4, 7.3; SAS textbook, Page 30 to 58

Problem 1
Dataset:

I have drunk the remains of a morning cup of coffee, with cream, in the late afternoon or evening many a time. In case of food poisoning, I have decided to keep a diary of how long it takes me to finish a 12-oz cup of coffee. I have also decided to toss out the coffee if I do not finish it by 8 pm. The following shows a one-week excerpt from my coffee diary:

Day	Time 1	Time 2	Elapsed Time (in minutes)	Remarks
1	12:05 pm	6:00 pm	355	Finished coffee
2	10:00 am	8:00 pm	600	Tossed out the remaining coffee
3	11:55 am	4:55 pm	300	Finished coffee
4	09:45 am	4:00 pm	375	Left unfinished coffee in a classroom
5	10:30 am	5:30 pm	420	Finished coffee

Here Time 1 is the time I start drinking my coffee; Time 2 is the last recorded time; and Elapsed time is the difference between Time 1 and Time 2. We are estimating the survival function for the time it takes me to finish a 12-oz cup of coffee.

1) Fill in the table the observed survival time and its status (1 for uncensored and 0 for censored) for each day:

Day	Observed survival time	Status	Censoring type if censored
1	355 min	1	
2	600 min	0	Type I Right
3	300 min	1	
4	375+ min	0	Type I right or random
5	420 min	1	

2) Compute the Kaplan-Meier estimate of the survival function and the Nelson-Aalen estimate of the cumulative hazard function.
   a) Let $t_{(j)}, j = 1, 2, \ldots k$ be the distinct uncensored survival times; $\hat{S}_j$ be the Kaplan-Meier estimate of the survival rate at time $t_{(j)}$; and $\hat{H}_j$ be the Nelson-Aalen estimate of the cumulative hazard rate at time $t_{(j)}$.

$t_{(j)}$	300 min	355 min	420
$\hat{S}_j$	0.80	0.60	0.40
$\hat{H}_j$	0.20	0.45	0.783

   b) Sketch the Kaplan-Meier estimated survival curve; please clearly label the plot.

300 355+
375+ 420
600+

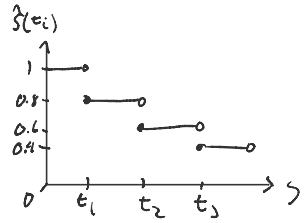
N-A

j	t_j	n_j	d_j	$\hat{S}_{t_j}$
1	300	5	1	$(1-\frac{1}{5}) = \frac{4}{5}$
2	355	4	1	$\hat{S}_1 * (1-\frac{1}{4}) = \frac{3}{5}$
3	420	3	1	$\hat{S}_1 * \hat{S}_2 * (1-\frac{1}{3}) = \frac{2}{5}$

N-A

$$\sum_{j=1}^{i} \frac{d_j}{n_j} = \hat{H}_j$$

6)



$$\hat{H}(t_i) = \left(\sum_{j=1}^{i} \frac{d_j}{n_j}\right) + \sum_{j=1}^{i} \left[\frac{(d_j/n_j)^2}{2} + \ldots\right]$$

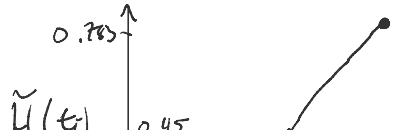
$\tilde{N}(t_i)$

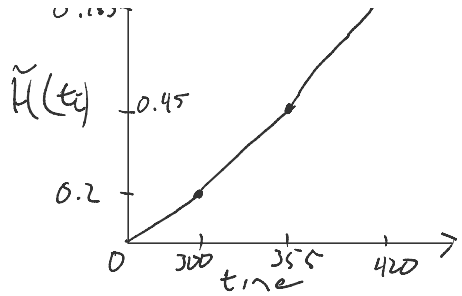
c) $\hat{H}(t_i) = -\sum_{j=1}^{i} \log\left(1-\frac{d_j}{n_j}\right)$

j	n_j	d_j	$\hat{H}(t_i)$
1	5	1	0.0969
2	4	1	0.2218
3	3	1	0.397

j	n_j	d_j	$\tilde{N}(t_i)$	t_j
1	5	1	0.20	300
2	4	1	0.45	355
3	3	1	0.783	420

Since we have a small sample, $\tilde{H}(t_i)$ is better (Nelson-Aalen estimate)





The Kaplan-Meier cumulative hazard curve is useful for large datasets & involves taking the log of the K-M estimator, while the Nelson-Aalen estimate is better for small datasets & is just the sum of events/failures over total observations $\left(\sum \frac{d_i}{n_i}\right)$

3) $\hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i-d_i)}$

$S_t \pm 1.96 * SE(S_t)$

j	t_j	n_j	d_i	$\hat{S}(t_j)$
1	300	6	1	$1-\frac{1}{6} = 5/6$
2	355	5	1	$\frac{5}{6}\times\left(1-\frac{1}{5}\right) = \frac{2}{3}$
3	360	4	1	$\frac{2}{3}\left(1-\frac{1}{4}\right) = \frac{1}{2}$
4	420	3	1	$\frac{1}{2}\left(1-\frac{1}{3}\right) = \frac{1}{3}$

$$\hat{S}(360)^2\left[\sum_{i=1}^{3} \frac{d_i}{n_i(n_i-d_i)}\right] = Var\left[\hat{S}(360)\right]$$

$$= \left(\frac{1}{2}\right)^2\left[\frac{1}{6(6-1)} + \frac{1}{5(5-1)} + \frac{1}{4(4-4)}\right]$$

$$Var\left[\hat{S}(360)\right] = 0.041667$$

$$\hat{S}(u_0) \pm z_{1-\alpha/2}\sqrt{Var[\hat{S}(u_0)}$$

$$\hat{S}(360) \pm 1.96\sqrt{Var\left[\hat{S}(360)\right]}$$

$$0.5 \pm 1.96\sqrt{1/24}$$

$$0.5 \pm 0.400083325$$

$$(0.0999, 0.9) \quad 95\% \text{ CI}$$

We are 95% confident that the survival rate of coffee is between 0.0999 & 0.9 (and SAS agrees)

4) I would not use life-table estimation for this data, given the small size of the data set.
Additionally, the data doesn't contain intervals.

5) $E[x] = m = \frac{1}{\lambda}$ $\quad \lambda = \frac{1}{m}$ $\quad h(x) = \lambda$ $\quad S(x) = \exp[-\lambda x]$ $\quad \exp[-\frac{1}{m}x]$ $\quad f(x) = \lambda \exp[-\lambda x]$

$$L(m) = m^r \exp[-m \cdot S_t], \quad r = \Sigma \delta_i, \quad m = \frac{1}{\lambda}$$

$$= [f(355) f(300) f(420)] * S(375) S(600)$$

$$= (m)^3 \exp[1075m] * \exp[975m]$$

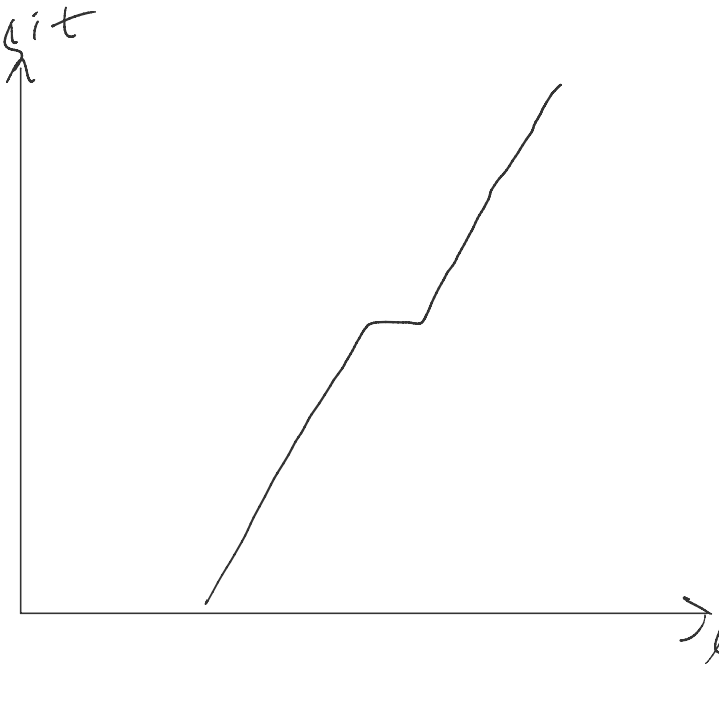
$$L(m) = (m)^3 \exp[2050m]$$

$$\ell(m) = \ln L(m) = 3 \log m - 2050m$$

$$\frac{\partial \ell}{\partial m} = \frac{3}{m} - 2050 = 0$$

$$\frac{3}{m} = 2050$$

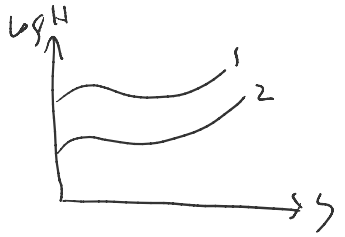
$m = 683.33$ min

6) 

the linear log relationship indicates that the survival curve follows an exponential distribution

7) The assumption of proportional hazards, in this context, assumes that the risk of you finishing your coffee versus your office mate is constant at time t. i.e. At time t, the risk of you finishing your coffee is double that of your office mate.

Ex: $\frac{h_1(t)}{h_2(t)} = 2$



8) We would use the log-rank test since it is the most powerful test, if the

j

Since it is the most powerful test if the proportional Hazards Assumption holds, which for this problem we assume it must.

survival plot for checking log-logistic model

The LIFETEST Procedure

Survival Function and Cumulative Hazard Rate

time		Product-Limit Survival	Failure	Survival Standard Error	Nelson-Aalen Cumulative Hazard	Cum Haz Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	.	0	5
300.000		0.8000	0.2000	0.1789	0.2000	0.2000	1	4
355.000		0.6000	0.4000	0.2191	0.4500	0.3202	2	3
375.000	*	.	.	.	.	.	2	2
420.000		0.3000	0.7000	0.2387	0.9500	0.5937	3	1
600.000	*	0.3000	.	.	.	.	3	0

Note: The marked survival times are censored observations.

Summary Statistics for Time Variable time

Quartile Estimates

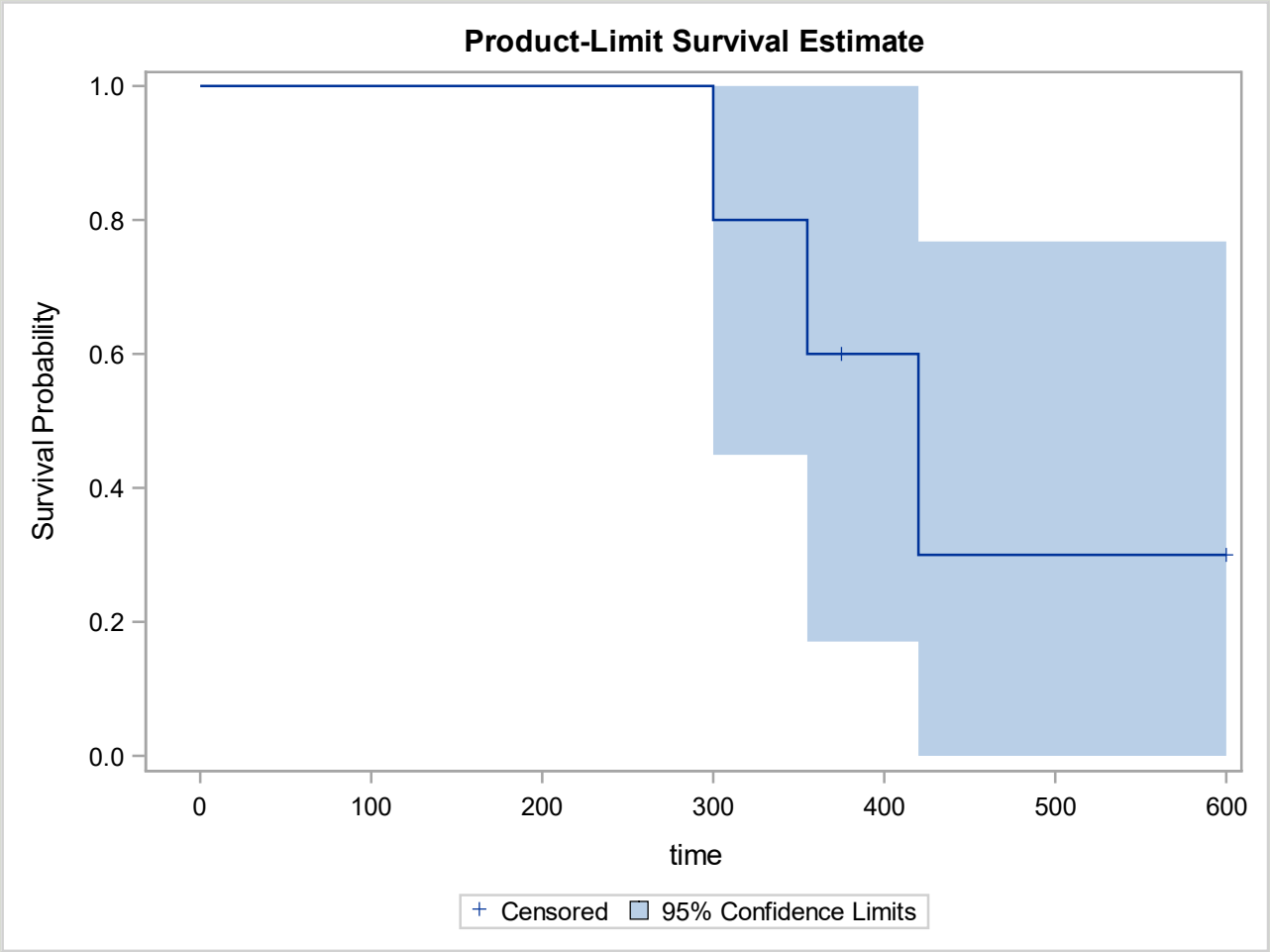
Percent	Point Estimate	95% Confidence Interval Transform	[Lower	Upper)
75	.	LINEAR	355.000	.
50	420.000	LINEAR	300.000	.
25	355.000	LINEAR	300.000	.

Mean	Standard Error
383.000	26.586

Note: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

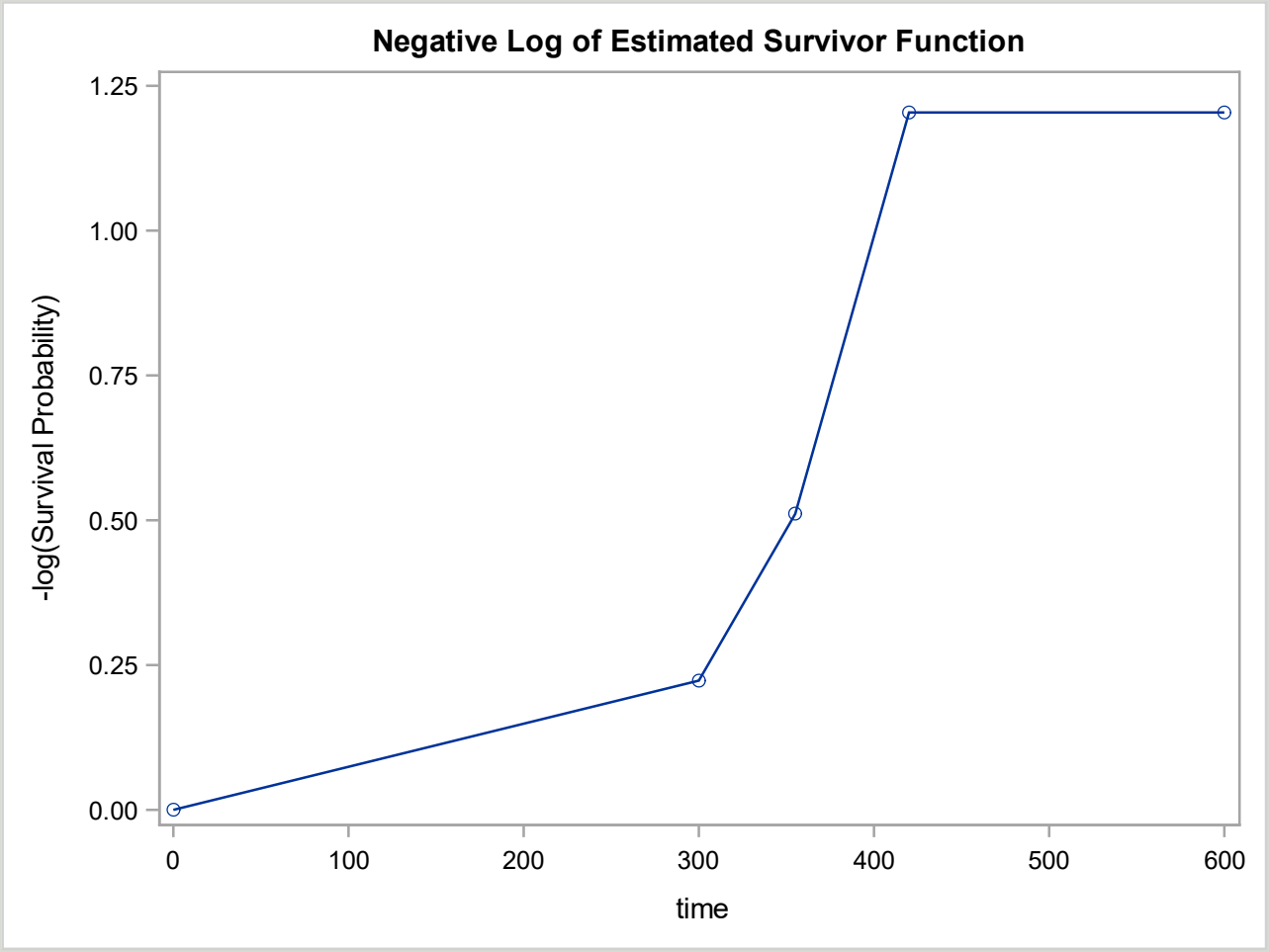
survival plot for checking log-logistic model

The LIFETEST Procedure



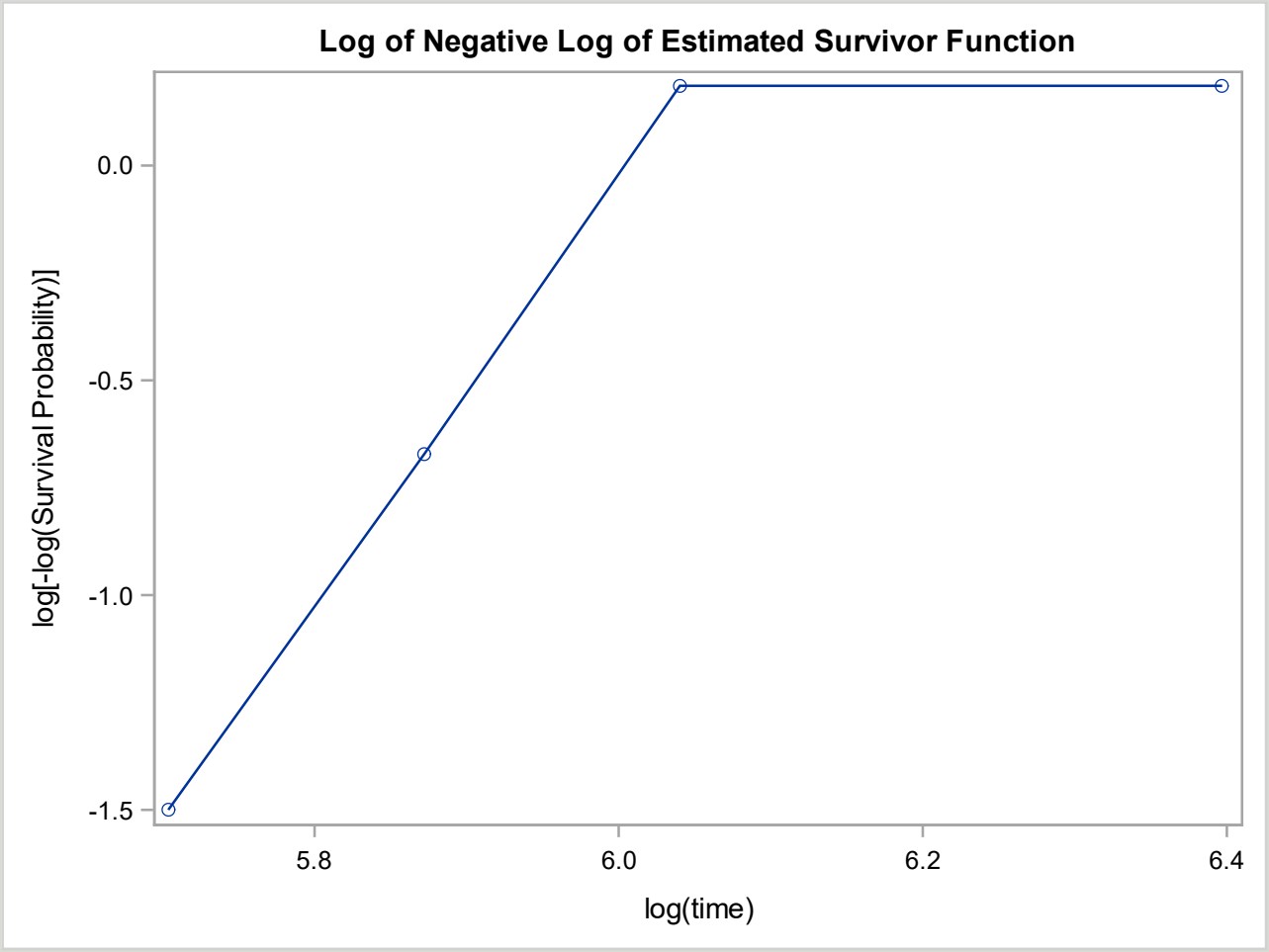
survival plot for checking log-logistic model

The LIFETEST Procedure



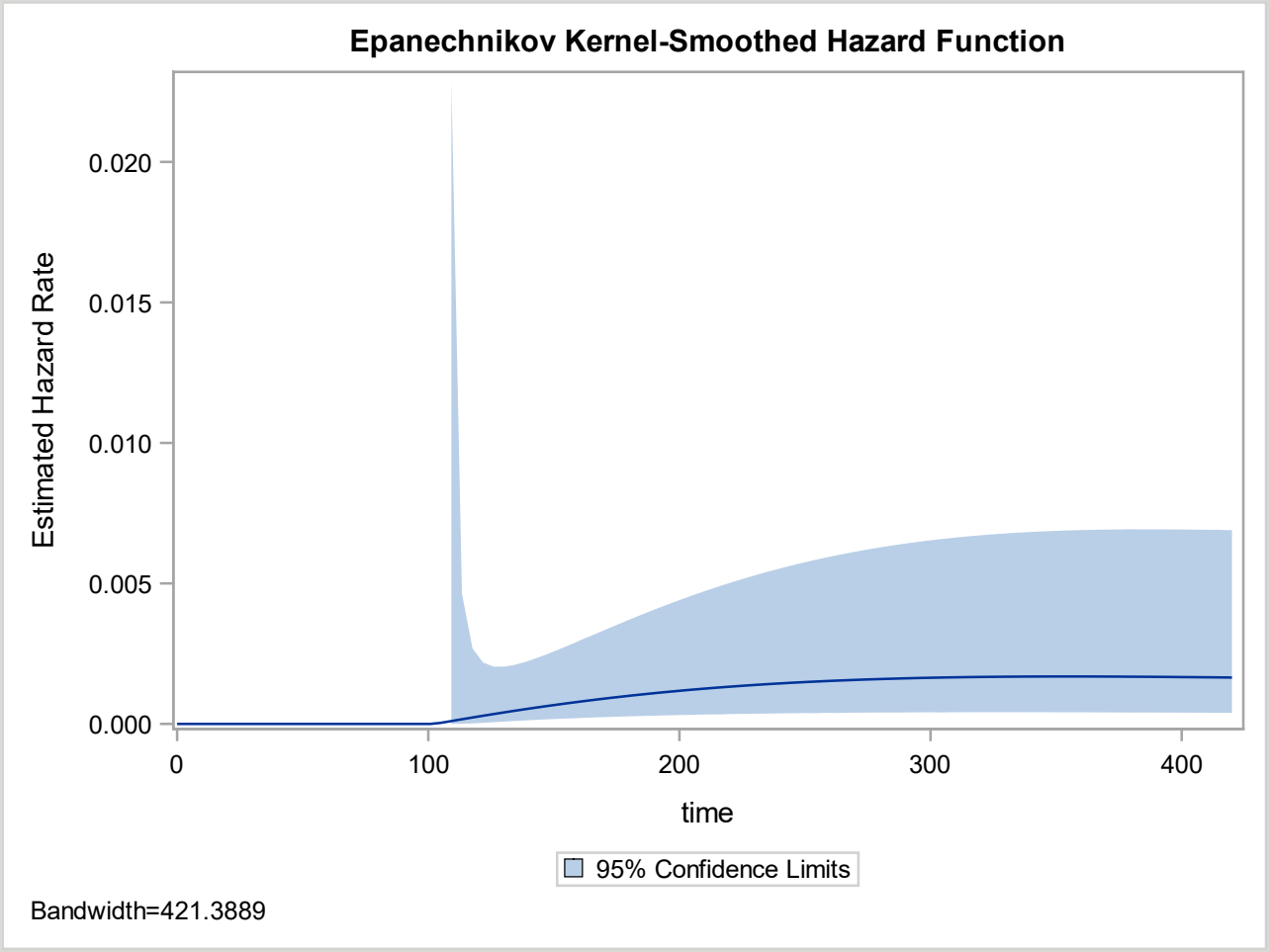
survival plot for checking log-logistic model

The LIFETEST Procedure



survival plot for checking log-logistic model

The LIFETEST Procedure



Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
5	3	2	40.00

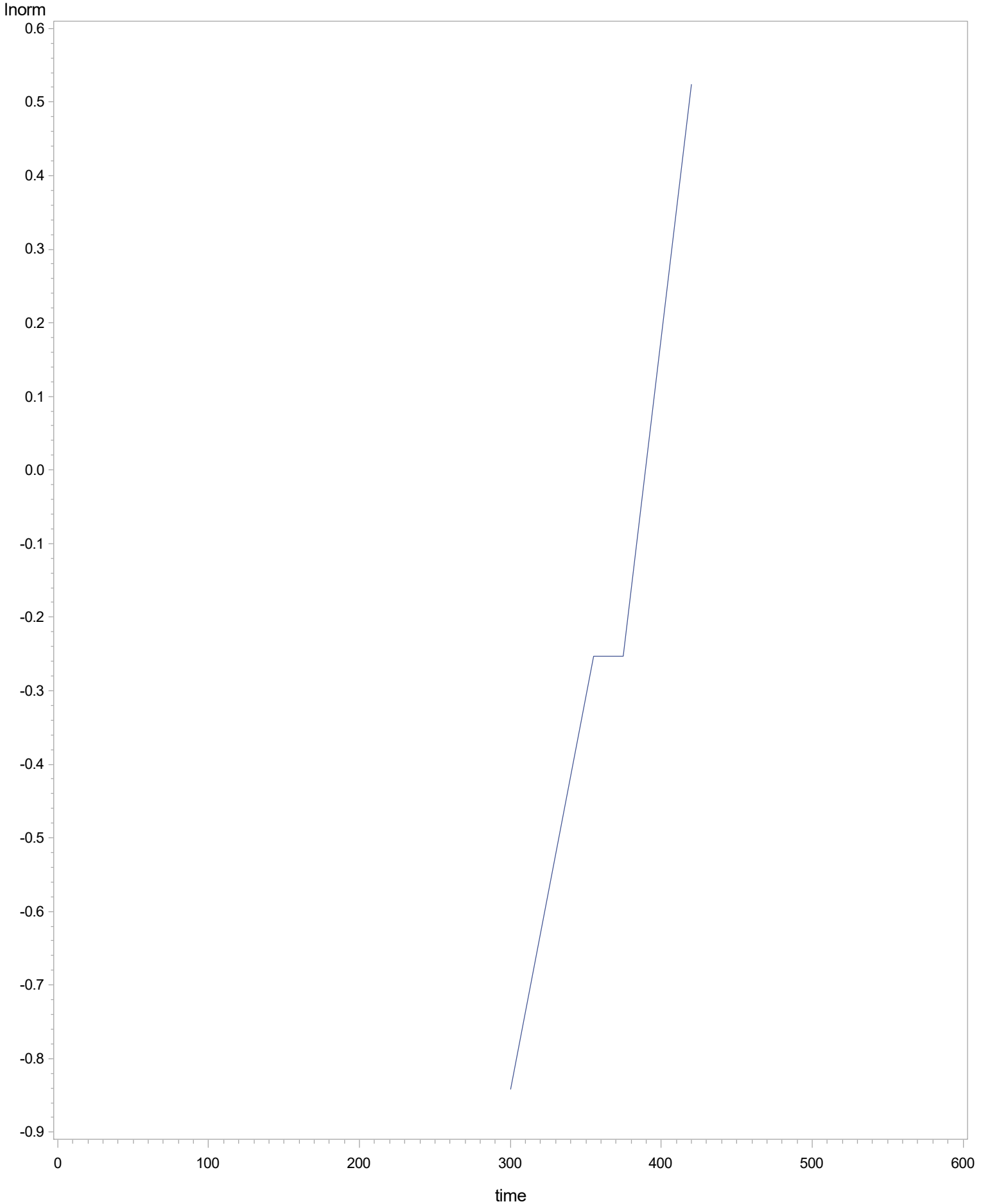
survival plot for checking log-logistic model

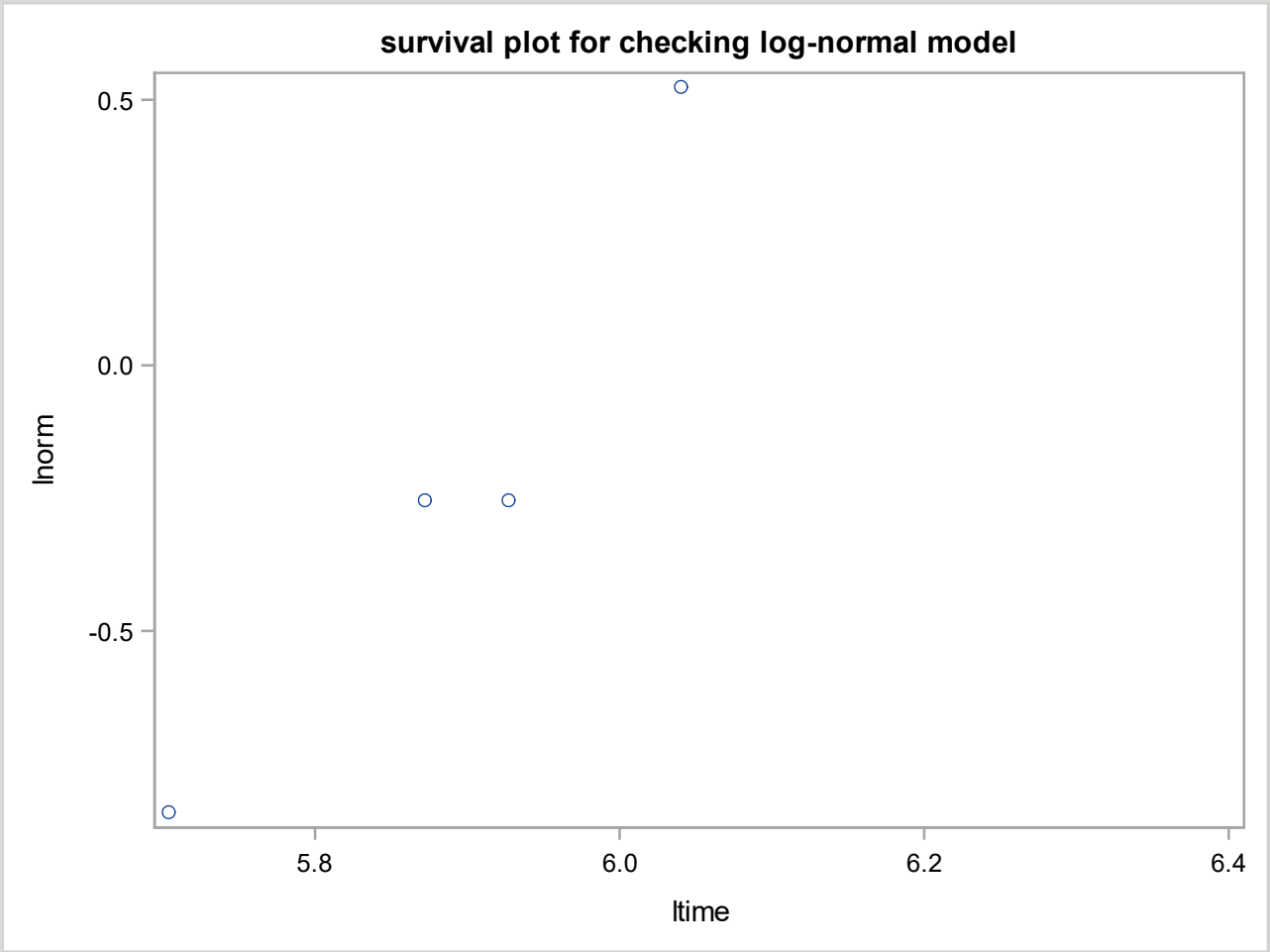
Obs	time	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
1	0	.	1.0	1.00000	1.00000
2	300	0	0.8	0.44939	1.00000
3	355	0	0.6	0.17059	1.00000
4	375	1	0.6	.	.
5	420	0	0.3	0.00000	0.76793
6	600	1	.	.	.

survival plot for checking log-logistic model

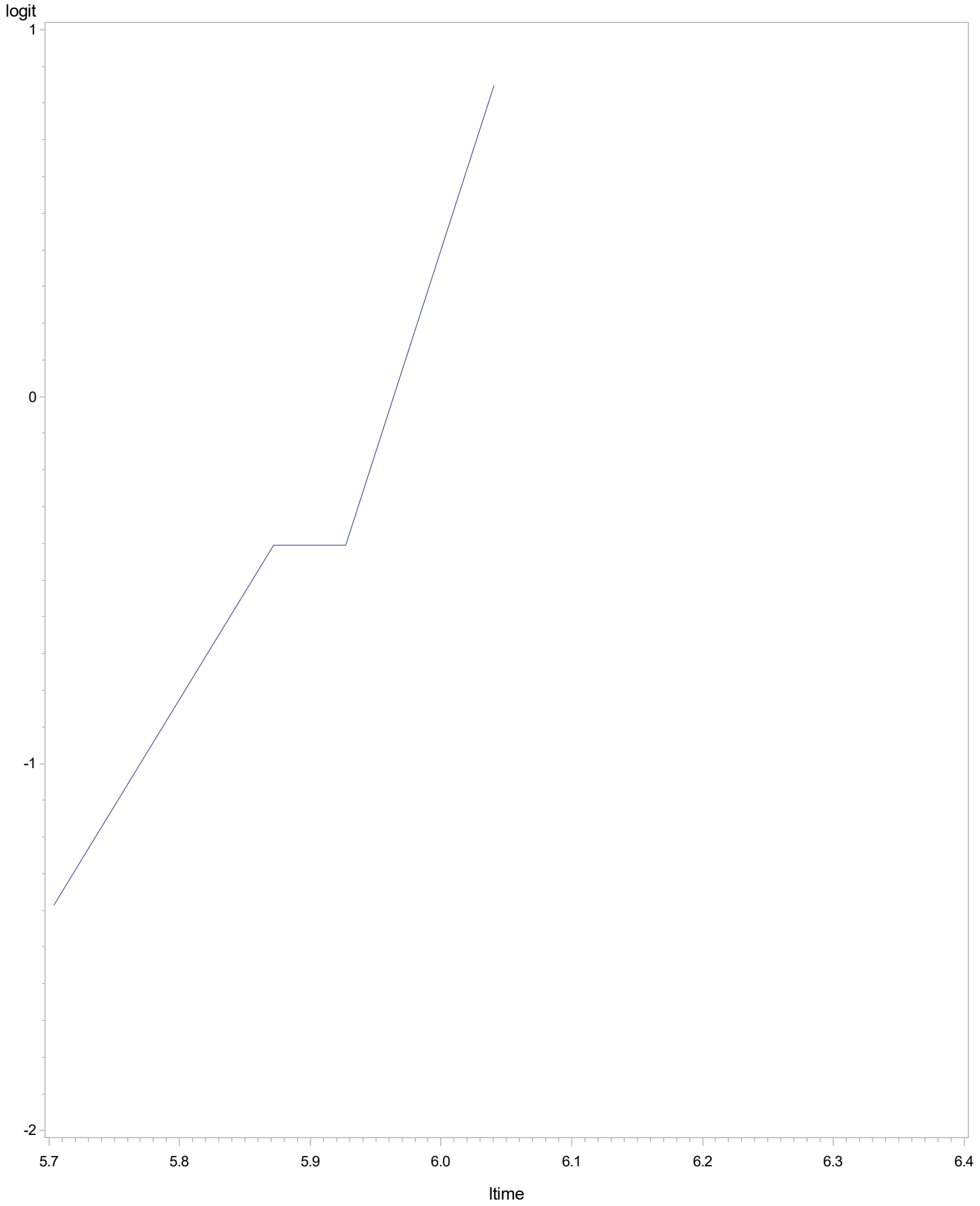
Obs	time	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL	s	logH	lnorm	logit	ltime
1	0	.	1.0	1.00000	1.00000	1.0	.	.	.	.
2	300	0	0.8	0.44939	1.00000	0.8	-1.49994	-0.84162	-1.38629	5.70378
3	355	0	0.6	0.17059	1.00000	0.6	-0.67173	-0.25335	-0.40547	5.87212
4	375	1	0.6	.	.	0.6	-0.67173	-0.25335	-0.40547	5.92693
5	420	0	0.3	0.00000	0.76793	0.3	0.18563	0.52440	0.84730	6.04025
6	600	1	.	.	.	.	.	.	.	6.39693

survival plot for checking log-normal model




survival plot for checking log-normal model
0.5
0.0
-0.5
lnorm
5.8
6.0
6.2
6.4
ltime

survival plot for checking log-logistic model



survival plot for checking log-logistic model

The LIFETEST Procedure

Survival Function and Cumulative Hazard Rate

time		Product-Limit Survival	Failure	Survival Standard Error	Nelson-Aalen Cumulative Hazard	Cum Haz Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	.	0	6
300.000		0.8333	0.1667	0.1521	0.1667	0.1667	1	5
355.000		0.6667	0.3333	0.1925	0.3667	0.2603	2	4
360.000		0.5000	0.5000	0.2041	0.6167	0.3609	3	3
375.000	*	.	.	.	.	.	3	2
420.000		0.2500	0.7500	0.2041	1.1167	0.6167	4	1
600.000	*	0.2500	.	.	.	.	4	0

Note: The marked survival times are censored observations.

Summary Statistics for Time Variable time

Quartile Estimates

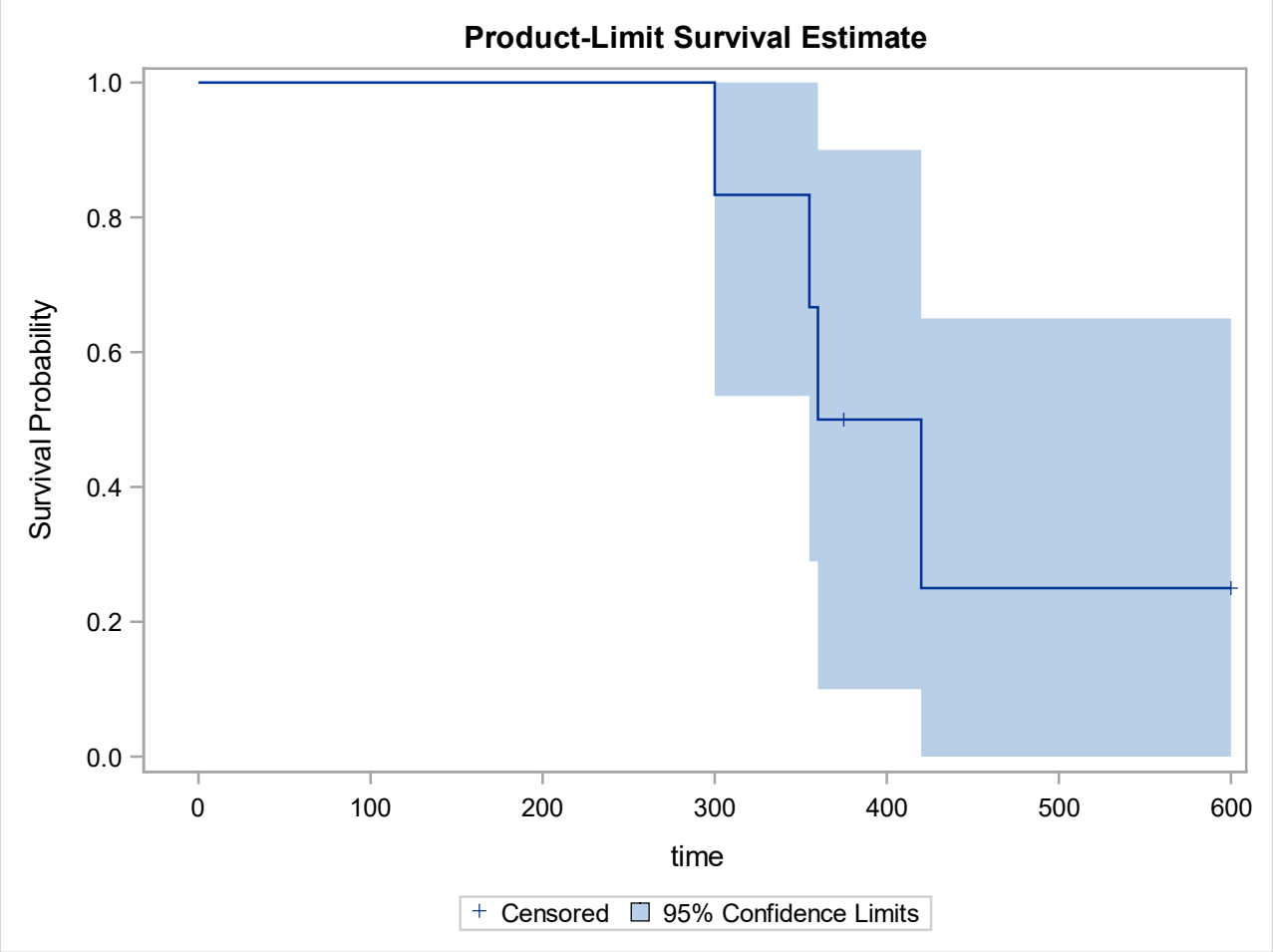
Percent	Point Estimate	95% Confidence Interval Transform	[Lower	Upper)
75	.	LINEAR	360.000	.
50	390.000	LINEAR	355.000	.
25	355.000	LINEAR	300.000	420.000

Mean	Standard Error
379.167	21.275

Note: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

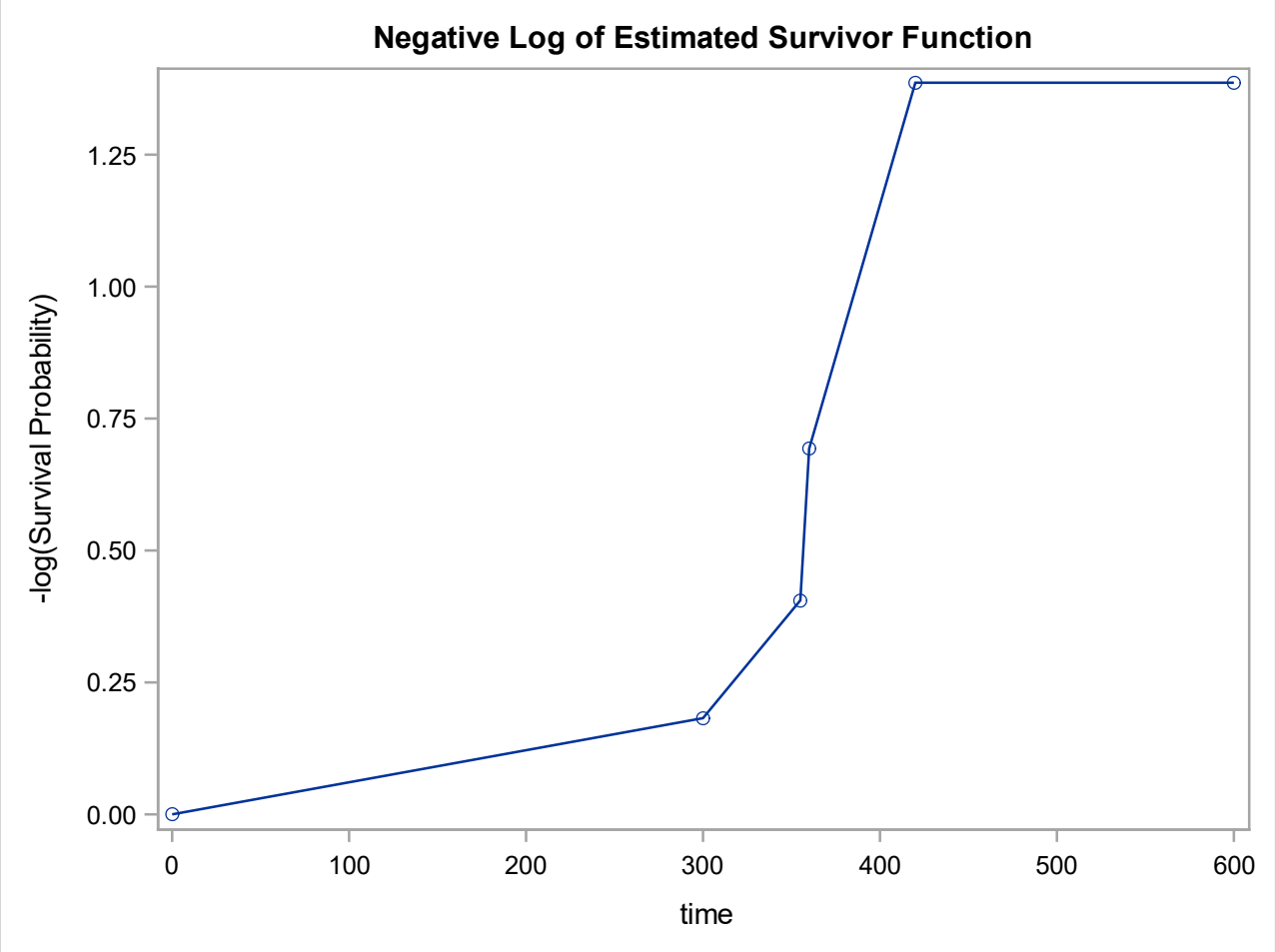
survival plot for checking log-logistic model

The LIFETEST Procedure



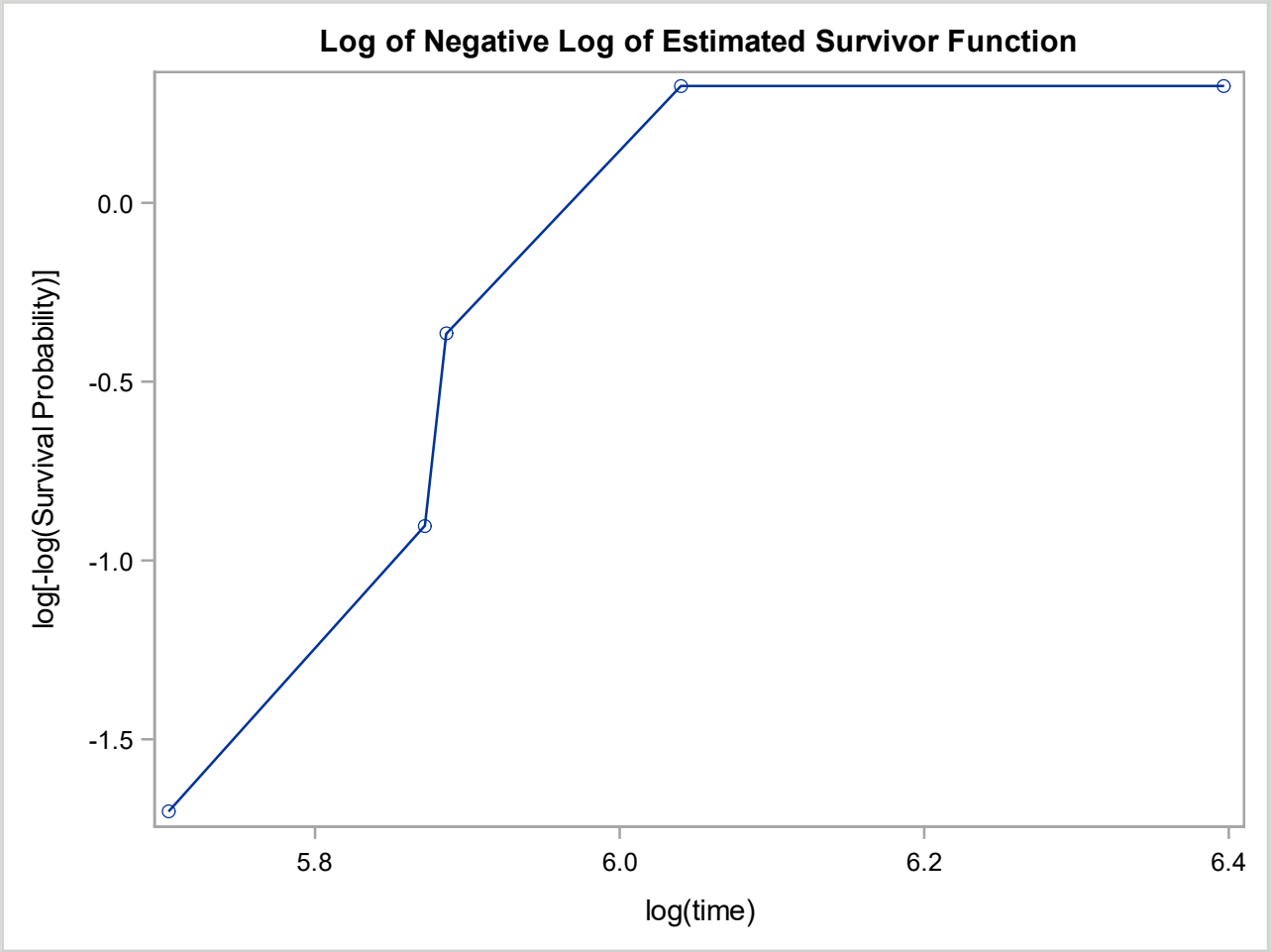
survival plot for checking log-logistic model

The LIFETEST Procedure



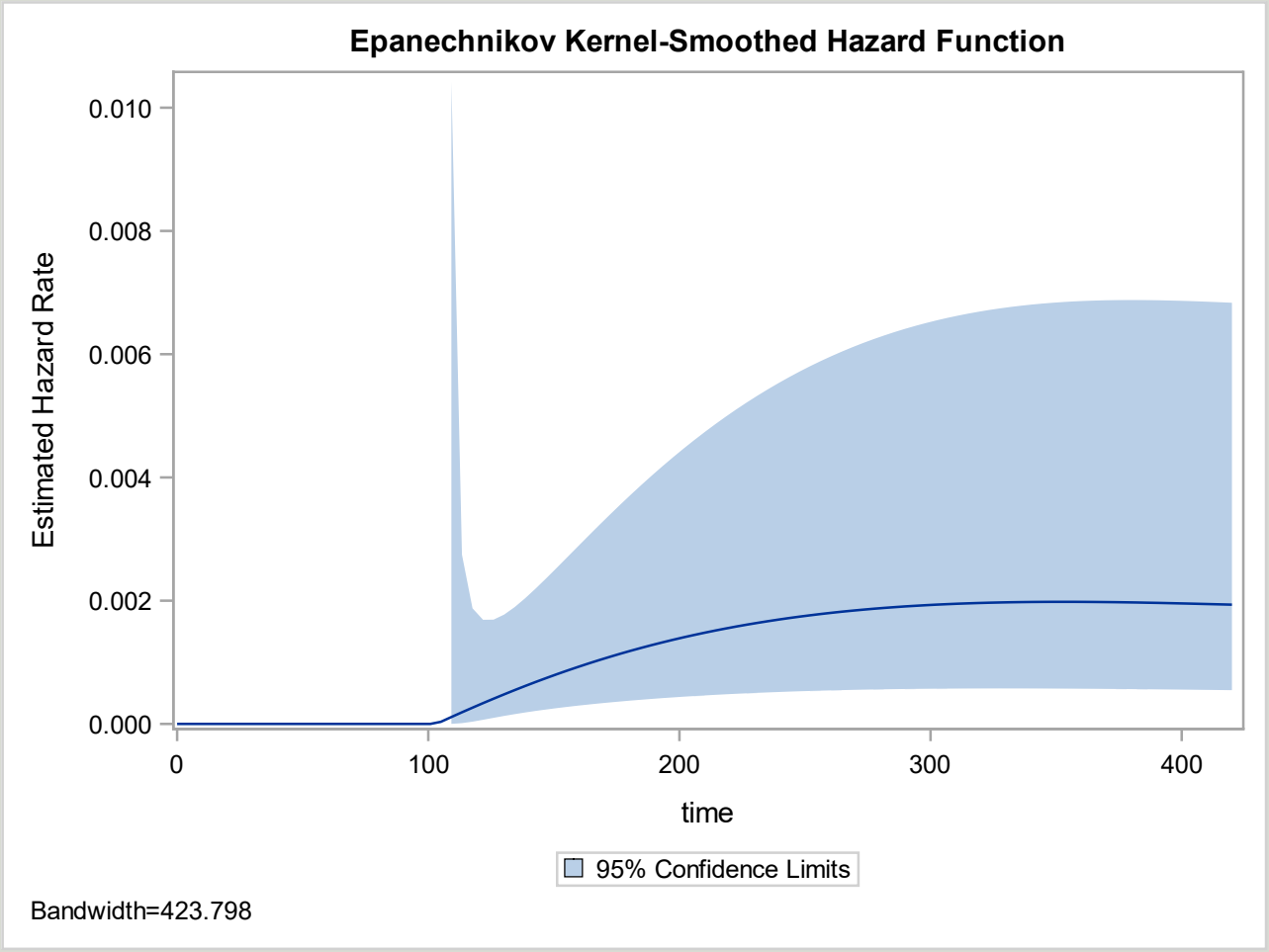
survival plot for checking log-logistic model

The LIFETEST Procedure



survival plot for checking log-logistic model

The LIFETEST Procedure



Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
6	4	2	33.33

survival plot for checking log-logistic model

Obs	time	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
1	0	.	1.00000	1.00000	1.00000
2	300	0	0.83333	0.53513	1.00000
3	355	0	0.66667	0.28947	1.00000
4	360	0	0.50000	0.09992	0.90008
5	375	1	0.50000	.	.
6	420	0	0.25000	0.00000	0.65008
7	600	1	.	.	.

survival plot for checking log-logistic model

The LIFETEST Procedure

Life Table Survival Estimates

Interval [Lower,	Interval Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error
0	2	77	5	924.5	0.0833	0.00909	1.0000	0	0	11.1513	0.5864
2	3	71	3	843.5	0.0842	0.00956	0.9167	0.0833	0.00909	10.6362	0.5627
3	5	119	6	768.0	0.1549	0.0131	0.8395	0.1605	0.0121	11.0119	0.5401
5	7	75	9	641.5	0.1169	0.0127	0.7095	0.2905	0.0150	11.3311	0.4994
7	11	109	7	558.5	0.1952	0.0168	0.6265	0.3735	0.0160	11.4810	0.8643
11	17	148	5	443.5	0.3337	0.0224	0.5042	0.4958	0.0166	11.4674	0.7806
17	25	107	0	293.0	0.3652	0.0281	0.3360	0.6640	0.0158	14.4054	1.3879
25	37	74	0	186.0	0.3978	0.0359	0.2133	0.7867	0.0138	15.5765	1.2836
37	53	85	0	112.0	0.7589	0.0404	0.1284	0.8716	0.0113	10.5412	0.9960
53	.	27	0	27.0	1.0000	0	0.0310	0.9690	0.00586	.	.

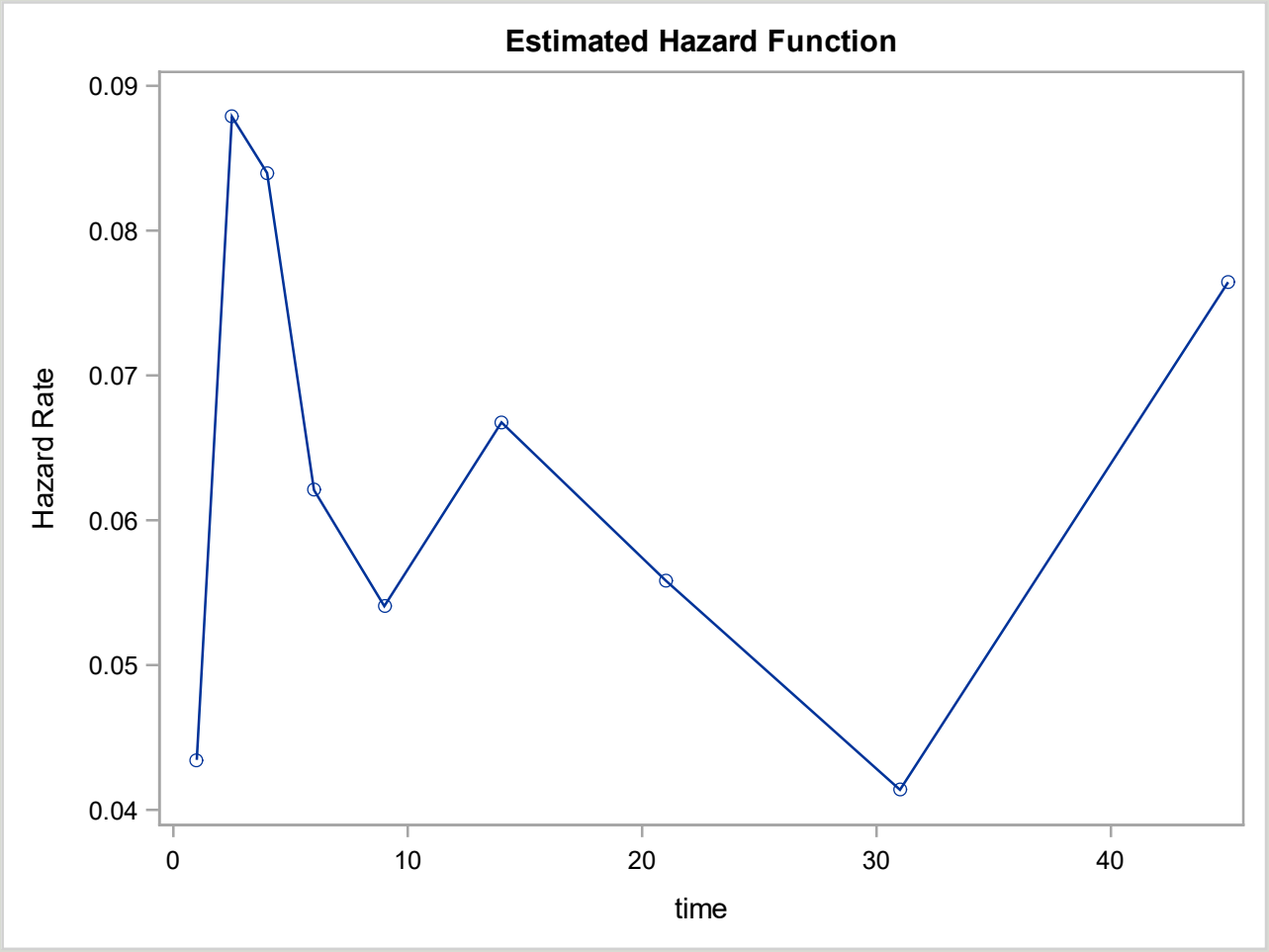
Interval [Lower,	Interval Upper)	Evaluated at the Midpoint of the Interval PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	2	0.0416	0.00454	0.043454	0.004947
2	3	0.0772	0.00880	0.087871	0.010418
3	5	0.0650	0.00556	0.08398	0.007671
5	7	0.0415	0.00458	0.062086	0.007155
7	11	0.0306	0.00274	0.054067	0.005148
11	17	0.0280	0.00210	0.066757	0.005376
17	25	0.0153	0.00138	0.055846	0.005262
25	37	0.00707	0.000785	0.041387	0.00466
37	53	0.00609	0.000626	0.076439	0.00656
53	.	.	.	.	.

survival plot for checking log-logistic model

The LIFETEST Procedure

survival plot for checking log-logistic model

The LIFETEST Procedure



Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
927	892	35	3.78

Note: 4 observations with invalid time, censoring, or frequency values were deleted.

```
ods pdf file='C:\Users\cianb\Documents\BSTA662_hw2.pdf' style=statistical;


data prob1;
input time status;
datalines;
355 1
600 0
300 1
375 0
420 1
360 1
;
run;


proc lifetest data=prob1 method=km nelson conftype=linear plots=(survival(cl),hazard(cl))
plots=(s,ls, lls) graphics outsurv=b;
time time*status(0);
*/strata drug/ tests=all;
run;




proc print data=b;
run;


data b2;
set b;
s=survival;
```

```
logH=log(-log(s));

lnorm=probit(1-s);

logit=log((1-s)/s);

ltime=log(time);

run;

proc print data=b2;

run;

proc gplot data=b2;

title "survival plot for checking log-normal model";

symbol i=join;

plot lnorm*time;

run;

proc sgplot data=b2;

scatter x=ltime y=lnorm;

run;

proc gplot data=b2;

title "survival plot for checking log-logistic model";

symbol i=join;

plot logit*ltime;

run;



data prob2;

input time status number;

datalines;

1 1 77

1 0 5

2.5 1 71
```

```
2.5 0 3

4 1 119

4 0 6

6 1 75

6 0 9

9 1 109

9 0 7

14 1 148

14 0 5

21 1 107

21 0 0

31 1 74

31 0 0

45 1 85

45 0 0

53.5 1 27

53.5 0 0

;

run;


proc lifetest method=life data=prob2 intervals=2 3 5 7 11 17 25 37 53

plots=(s,h) outsurv=a;

time time*status(0);

freq number;

run;


ods rtf file="hw2.rtf";
```

```
proc document name=temp(write);
   import textfile="C:\Users\cianb\Documents\Biostats\BSTA 662_HW2_Colin_Busby.sas" to ^;
   replay;
run;
ods rtf close;


ods pdf close;
```