DATA Analysis Report

Xinning Chu

April 17, 2018

The George Washington University

Contents

Part A

1. Data Description

There are 150 observations of one response variable(time), one status variable and an explanatory variable in this dataset. And obviously the variable group has 3 levels of value.

2. Survivor Fuction

We should introduce a definiton of survivor function before estimating. The probility that the survival time T is greater than or equal to t can be writen as S(t) = Pr(T⩾t), where S(t) is called “survivor function”.

2.1 Product limit method

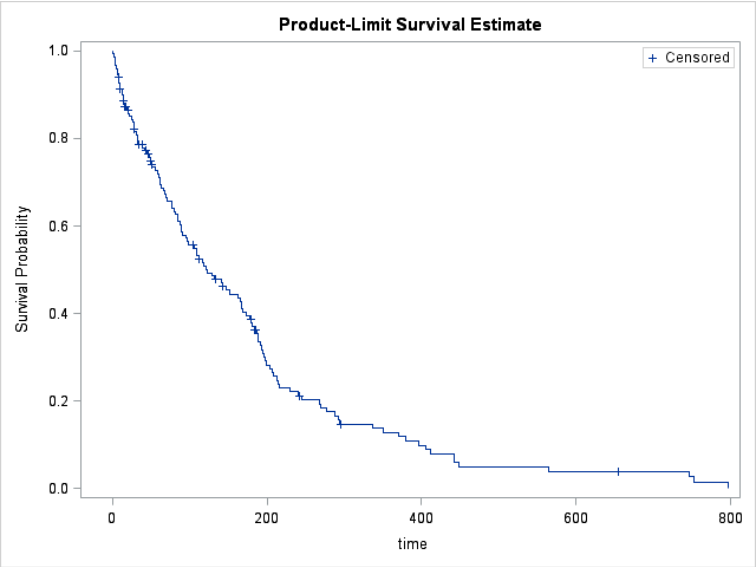
Initially, the product limit is used to estimate the parameters of the survivor function. Product limit is a nonparametric method, which can estimate the survivor function from censored data.

Table 1: Summary statistics for time variable time

Percent	Point estimate	Transform	Lower CI	Upper CI
75	212.840	LOGLOG	191.580	276.500
50	122.260	LOGLOG	89.520	166.990
25	48.410	LOGLOG	28.320	67.610

According to the table above, the point estimate of this method is 212.84 with 75 percent, 122.26 with 50 percent and 48.41 with 25 percent. Other estimates are also shown in the table.

Figure 1: Product-limit survival estimate



The fitting plot of the product limit method is shown in the figure above. Apparently, the curve is not so smooth. Thus, possibly some other methods are better to fit the function. Further research will be done in the next part.

2.2 Life Table Method

In this part, life table method is used to fit the survival function. Trough this way, numbers of censored and uncensored observations are useful to estimate the survival function.

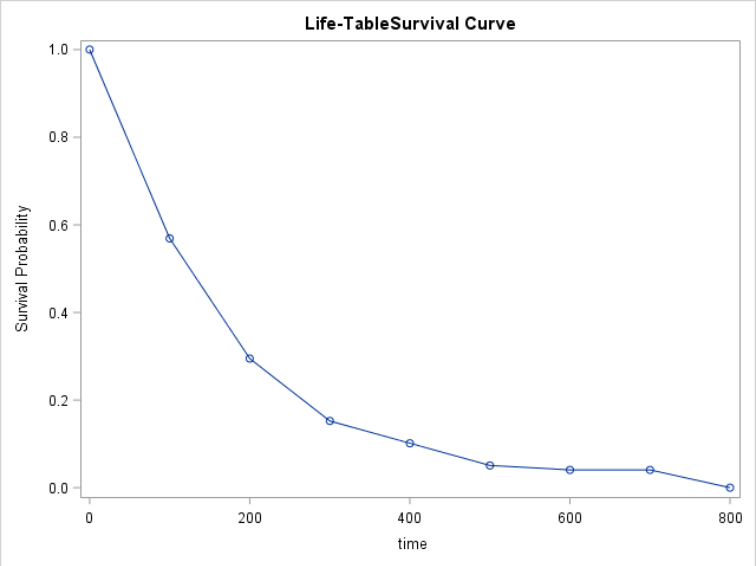
Table 2: Summary Statistics of the life table method

lower	upper	number failed	number censored	survival	failure
0	100	61	17	1.0000	0
100	200	33	7	0.5689	0.4311
200	300	15	2	0.2948	0.7052
300	400	5	0	0.1522	0.8478
400	500	5	0	0.1014	0.8986
500	600	1	0	0.0507	0.9493
600	700	0	1	0.0406	0.9594
700	800	3	0	0.0406	0.9594
800	.	0	0	0	1.0000

Some summary statistics of this method is shown in the Table 2, such as

the number of survival or failure, the number of censored or failed observations and the number of lower and upper.

Figure 2: Time table survival function



The plot of survival under time table method is shown in the Figure 2. Though compared with the last one, this curve is smoothier, it's not good enough.

Thus, we can draw a conclusion that there is not a significant difference between the survivor functions estimated by these two methods.

2.3 Confidence interval plot

Figure 3: The confidence interval of survival function

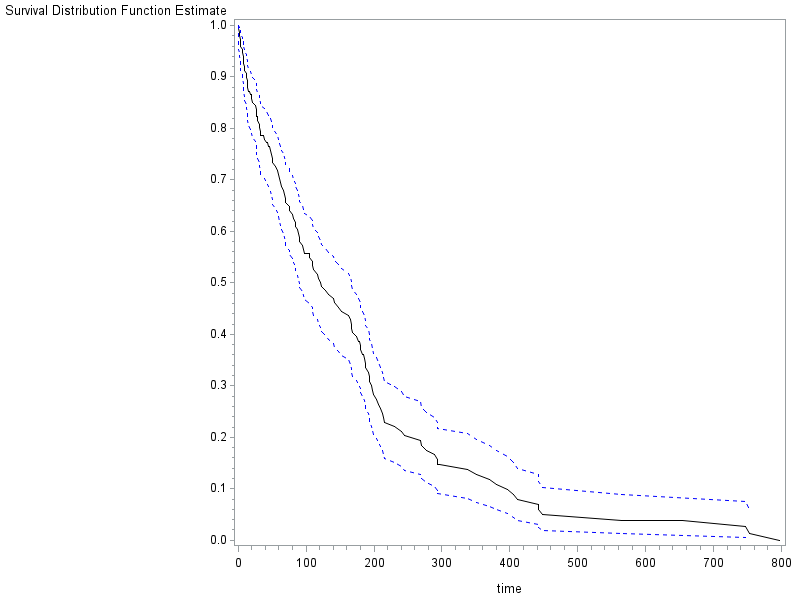
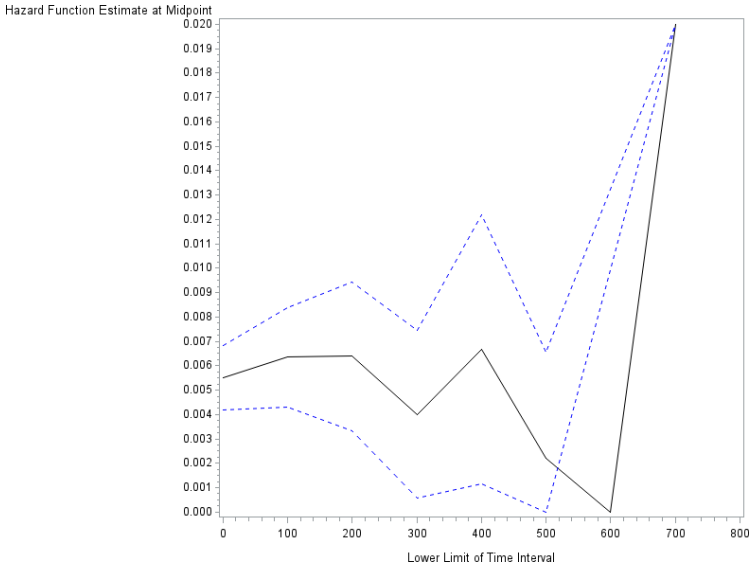


Figure 4: The confidence interval of survival function



The figure 3 shows the confidence interval plot of survivor function by product limit method, while the figure 4 shows the confidence interval plot of Hazard function. The trend of survival function and Hazard are

clear.

2.4 survivor function under different group

Then the survival function under different group is considered and it's used to verify the relation between each group. In the following tables, we can see the result of survival under different groups.

Table 3: Group 1

Percent	Point estimate	Transform	Lower CI	Upper CI
75	166.990	LOGLOG	111.530	214.340
50	80.970	LOGLOG	42.010	116.030
25	27.530	LOGLOG	13.600	58.300

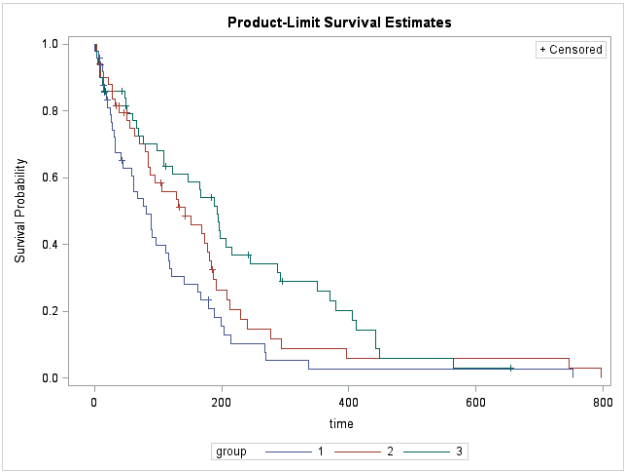
Table 4: Group 2

Percent	Point estimate	Transform	Lower CI	Upper CI
75	208.920	LOGLOG	176.930	276.500
50	141.910	LOGLOG	83.700	180.710
25	54.740	LOGLOG	21.730	87.830

Table 5: Group 3

Percent	Point estimate	Transform	Lower CI	Upper CI
75	369.960	LOGLOG	207.140	441.740
50	193.530	LOGLOG	109.230	244.490
25	66.680	LOGLOG	14.740	122.260

Figure 5: The confidence interval of survival function



Some statistical test is performed to evaluate the different between estimated survivor function for different group, and results are shown in the Table 6. All the p-values of are smaller than 0.05, so we can conclude the survivor function of different groups are significantly different.

Table 6

Test	Chi Square	DF	P value
Log-Rank	10.4587	2	0.0054
Wilcoxon	8.7038	2	0.0129
Tarone	10.6033	2	0.0050
Peto	9.5432	2	0.0085
Modified Peto	9.5032	2	0.0086
Fleming(1)	9.5519	2	0.0084

Part B

1. Data Description

There are 100 observations of one response variable(time), one status variable and an explanatory variable in this dataset. And obviously the variable group has 3 levels of value.

2. Cox's regression analysis

In this part, the cox's regression analysis with x and z is performed. The table 7 show the result of maximum likelihood estimates of variable x and y. Both the p-values of x and z are greater than 0.05, so we conclude there are not significant variables associated with the survival time.

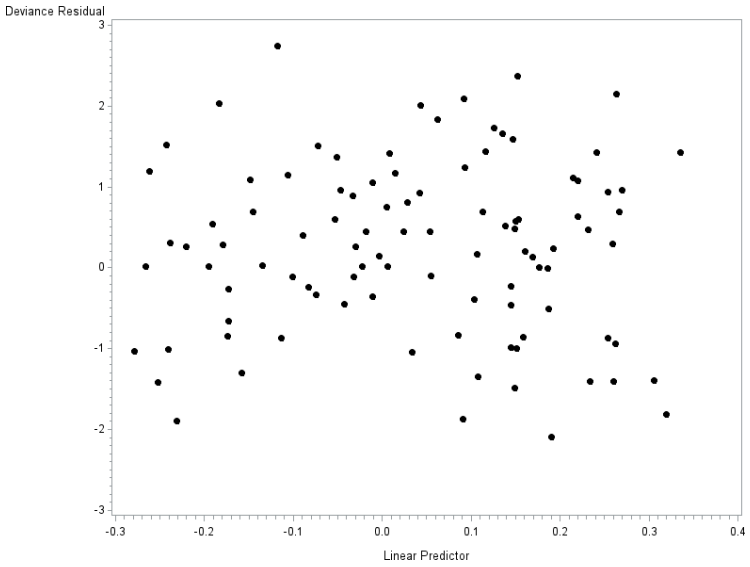
Table 7 Analysis of maximum likelihood estimates

parameter	DF	parameter estimate	SE	CHI square	p value	Hazard ratio
x	1	-0.00339	0.00437	0.6030	0.4375	0.997
z	1	0.04109	0.03787	1.1776	0.2778	1.042

3. Model Diagnostics

The residual plot is performed to verify the assumption of the modeland check the dispersion of the plot. In the figure 6, we can see the result of residual plot, and find that the plot distributed randomly between y=0. So the model is good.

Figure 6: The residuals plot



Part C

1. Data description

The dataset contains 428 observations. X is a fixed variable and the z is a time varying variable.

2. Cox's regression with time dependent covariate

In this step, the cox's regression is performed. When both x and z included, the performance of the model from the table below. The p-value of z is smaller than 0.05, so the variable z is significant. In addition, the p value of x variable is greater than 0.05, so it is an

insignificant variable. The confidence interval of Hazard ration of x variable is (0.855,1.181) and the CI of z variable is (1.056,1.180).

Table 11: Analysis of maximum likelihood estimates

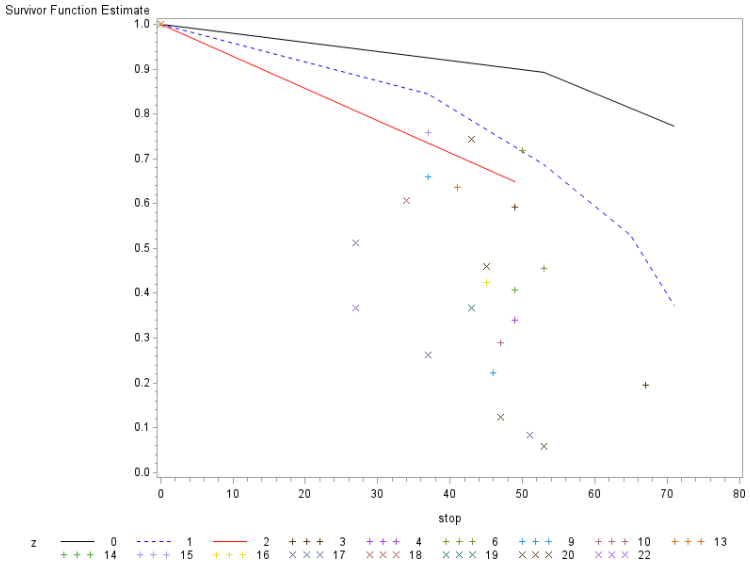
par	DF	par estimate	SE	CHI square	p value	Hazard ratio	Lower CI	Upper CI
x	1	0.06013	0.05429	1.2266	0.2681	1.062	0.955	1.181
z	1	0.11014	0.02837	15.0723	0.0001	1.116	1.056	1.180

3. Cox's regression with stratified analysis(based on x)

Table 12: Analysis of maximum likelihood estimates

par	DF	par estimate	SE	CHI square	p value	Hazard ratio	Lower CI	Upper CI
x	1	0.07183	0.08997	0.6376	0.4246	1.074	0.901	1.282

Figure 7: The residuals plot



The variable of x become significant when the variable z is the time dependent covariate. Also, we know the confidence interval of Hazard ration is (0.901,1.282)