Homework 1: BST 665: Survival Analysis

Devin Koehl

Spring Semester 2019

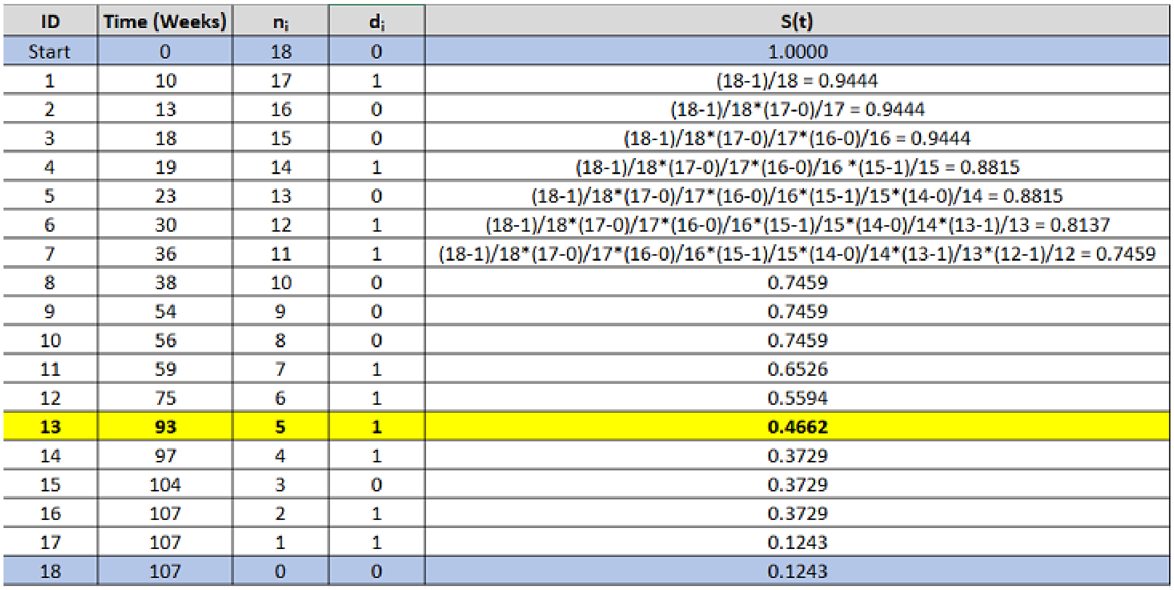
1. A gynecologist has collected data on women using a particular intrauterine device (IUD). She is interested in estimating how long women will use the IUD before discontinuing, or removing, the device due to adverse events. The following data gives the number of weeks from beginning using the IUD until discontinuation. Times for women who were lost to follow-up or discontinued use for reasons other than an adverse event (i.e., censored times) are labelled with a + sign.

10, 13+, 18+, 19, 23+, 30, 36, 38+, 54+, 56+, 59, 75, 93, 97, 104+, 107, 107, 107+

Use these data to complete Parts A-C. Each of these problems should be done by hand(you can use SAS to check your answers if needed).

A. Compute and sketch the Kaplan-Meier estimator for these data.

Screen shot of hand drawn Kaplan-Meier and table:



ID	Time (Weeks)	n_i	d_i	S(t)
Start	0	18	0	1.0000
1	10	17	1	(18-1)/18 = 0.9444
2	13	16	0	(18-1)/18*(17-0)/17 = 0.9444
3	18	15	0	(18-1)/18*(17-0)/17*(16-0)/16 = 0.9444
4	19	14	1	(18-1)/18*(17-0)/17*(16-0)/16 *(15-1)/15 = 0.8815
5	23	13	0	(18-1)/18*(17-0)/17*(16-0)/16*(15-1)/15*(14-0)/14 = 0.8815
6	30	12	1	(18-1)/18*(17-0)/17*(16-0)/16*(15-1)/15*(14-0)/14*(13-1)/13 = 0.8137
7	36	11	1	(18-1)/18*(17-0)/17*(16-0)/16*(15-1)/15*(14-0)/14*(13-1)/13*(12-1)/12 = 0.7459
8	38	10	0	0.7459
9	54	9	0	0.7459
10	56	8	0	0.7459
11	59	7	1	0.6526
12	75	6	1	0.5594
13	**93**	**5**	**1**	**0.4662**
14	97	4	1	0.3729
15	104	3	0	0.3729
16	107	2	1	0.3729
17	107	1	1	0.1243
18	107	0	0	0.1243

At each point we are using the following equation:

$\hat{S}(t_i) = \prod_{k=1}^{i} \frac{n_k - d_k}{n_k}$, I demonstrated this on the first 8 rows to show the algorithm. We continue down using the same approach each time.

If an observation is censored, it is not included in the calculation because the fraction will become 1.

B. Estimate the median time to discontinuation.

$min\{t : \hat{S}(t) \leq 0.50\}$

The median survival time is highlighted in yellow in the above table. This is the first instance where we drop $\leq$ 0.50. This occurs at 93 weeks where $\hat{S}(t)$ = 0.4662.

C. The gynecologist would like to know what proportion of women will use the IUD for at least a year (i.e., 52 weeks). Provide an answer to this question.

Of the 18 subjects, 10 of them had IUD for at least a year. $\frac{10}{18} = .55$.

2. In survival analysis, using a fully parametric regression model means proposing a distribution for the survival time, T. While the Weibull and exponential distributions are the most common choices, any continuous distribution that restricts T to be positive can be used. Suppose we want to use the following as the cumulative distribution function for T:

F(t) $= \frac{t^3}{1+t^3}, t > 0$

A. What is the survival function for T?

The CDF is related to the survivor function by the following equation:

S(t) = 1 - F(t)

Which yields S(t) $= 1 - \frac{t^3}{1+t^3}$

We can check this is true by also observing the following relationship:

f(t) $= \frac{dF(t)}{dt}$

Which yields

f(t) $= \frac{3t^2}{(1+t^3)^2}$ by using the quotient rule.

f(t) $= \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$

Going backwards and taking the -$\frac{d}{dt}$ we get the same function for f(t), $\frac{3t^2}{(1+t^3)^2}$

B. What is the hazard function?

The hazard function is the ratio of the probability density function to the survival function, S(x).

h(x) = $\frac{f(x)}{S(x)} = \frac{f(x)}{1-F(x)}$

We could also use the -log of the survivor function.

h(x) =$\frac{1-\frac{t^3}{1+t^3}}{\frac{3t^2}{(t^3+1)^2}} = \frac{3t^2}{1+t^3}$

To check this we could also do

$\frac{d}{dt} - log(1 - \frac{t^3}{1+t^3})$ which yields the same answer $\frac{3t^2}{1+t^3}$

C. What is the median survival time?

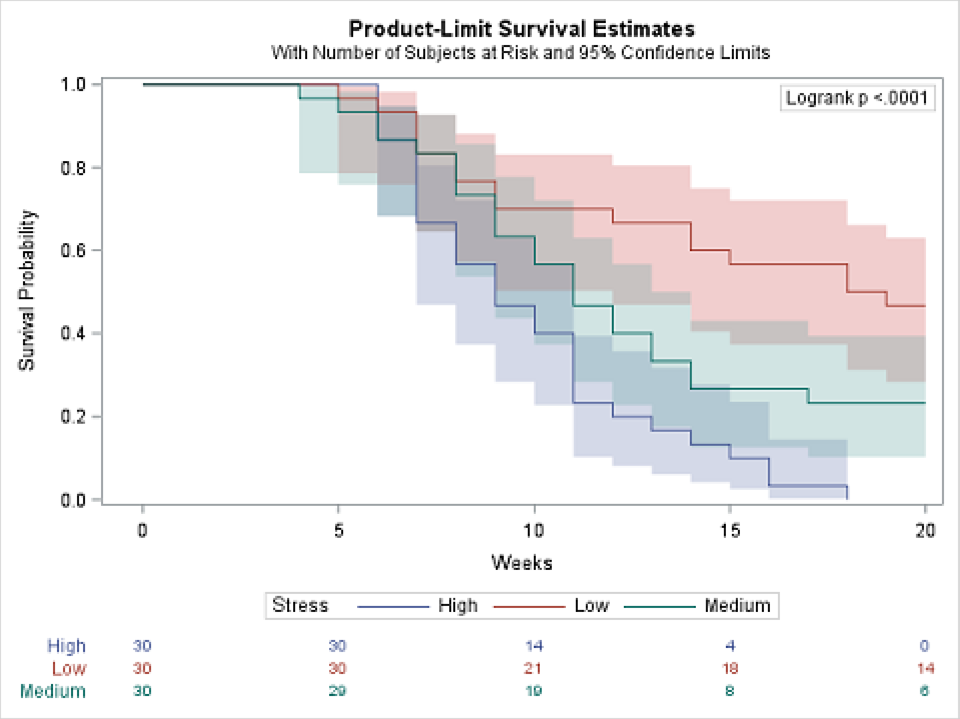
We can evauluate the survivor function at with the condition that it must be $\leq$ 0.50.

S(t) = 1-$\frac{t^3}{1+t^3} \leq 0.50$.

If we solve for values which satitsfy this equation, using S(1.0), we get exactly S(t) = 0.50. This means the median survival time would be where t=1.0

3. A researcher is interested in the relationship between environmental stress and the development of tumors in rats. Ninety rats were divided into three groups and exposed to lowstress, medium-stress, and high-stress environments. The investigator injected tumor cells into the rats and observed them for 20 weeks. The outcome of interest was the tumor-free time (i.e., the number of weeks between injection and development of a tumor).

A. On the same graph, plot the Kaplan-Meier estimators for the three environments. Which group had the poorest survival? Which group had the best?



From the Kaplan-Meier, the group with the poorest survival is the high stress group. The group with the best survival is the low stress group. The medium stress group lied in between both of these.

B. What was the median survival time for each of the three groups?

The median survival time for the low stress group was 18.5 weeks which occured at S(t) = 0.5000. The median survival time for the medium stress group was 11 weeks which occured at S(t) = 0.4667. The median survival time for the high stress group was 9.0 weeks which occured at S(t) = 0.4667.

C. Is there evidence to conclude that environmental stress is associated with tumor development? Support your answer using the log-rank test. (Make sure you state the null and alternative hypotheses, level of significance used, test statistic, p-value and conclusion in terms of the problem).

Yes, there is evidence to conclude that environmental stress is associated with tumor development. The log-rank test between the groups was significant.

Null hypothesis:

$H_0 : S_1 = S_2 = S_3$ There is no difference in the survival curves

Alternative Hypothesis:

$H_1 : S_1 \neq S_2 \neq S_3$ There is a difference in the survival curves

A p-value less than 0.05 is considered significant. The log-rank test is ¡0.0001 between the two groups, with a test statistic of 20.33. Therefore, we reject the null hypothesis and conclude there is a significant difference in the survival curves.

D. Write a short summary (1-2 paragraphs) of the results of this study. Be sure to include your interpretation of the results and to take the stated goal of the study into consideration.).

A researcher was interested in the reationship between environmental stress and the developement of tumor in rats. Ninety rats were divided into three groups, each consisting of thirty rats. The researcher then injeted tumor cells into the rats. The outcome of the study was the tumor-free time and how environmental stress affected the rats tumor-free time. Kaplan-Meier Survival methods were used to analyze this problem and a log-rank test was reported. The results show that the rats subjected to low ennvironmental stress had longer median survival time (18.5 days), while the rats subjected to high environmental stress were much shorter (9.0 weeks). The log-rank test between the groups was significant (p¡0.0001) and shows the trajectory of each group was different due to stress level. The low rats faired best, followed by medium stress. Lasltly, the high stress rats had the poorest survival. Overall, it was shown that environemental stress is associated with to tumor onset.