

12. Survival Analysis

In survival analysis we are interested in the time interval between entry into the study and an event. The outcome of interest is time to an event. Survival analysis was originally developed for studying time from commencement of treatment until death. This was commonly used for evaluating treatment efficacy in fatal conditions like cancer. Hence the name. But survival analysis is applicable to many other situations in addition to mortality. For example, commencement of Hormone Replacement Therapy and thrombotic episode; time to exercise to maximum tolerance; exclusive breast feeding and time to another pregnancy; time for leg fracture to heal; and so on. By convention one mentions survival data and survival analysis in all such cases regardless of the nature of the event. In industry the same approach is used to test the reliability of appliances. Here the outcome of interest is time to breakdown. Hence some computer packages (e.g. Minitab) refer to the procedure as reliability analysis.

The problem with studying the time between one event and another is twofold:

1. Time interval may vary from one subject to another. At the end of the follow up period the chances are that the event would probably not have happened for all the subjects. Therefore we cannot expect time intervals to be distributed normally.
2. In studies requiring prolonged observation, some subjects are invariably lost to follow up. So the only information we can have about them is that they were still alive at the time of the last follow up. These are termed **censored observations**. Also those subjects in whom the event has not occurred by the end of the follow up period are considered censored. All studies last for a finite time. At the end we do not know when the remaining subjects will experience the event. We do not know when they will experience the event

In all survival studies we are making certain assumptions. Patients are recruited over a period and followed up to a fixed date beyond the end of recruitment. Some would stay in the study for longer time than those who were recruited more recently. It is assumed that the survival prospects stay the same throughout the study.

We also assume that patients lost to follow up have the same prognoses as those who continue in the study. A critical assumption is that the probability of an individual subject to be censored is unrelated to the probability of suffering the endpoint event.

From the study design point of view survival should always be evaluated in a cohort of patients followed forward in time from a particular start point such as randomization, even if the cohort is historical.

The data is best looked at using the Kaplan- Meier Survival Curve. To construct such a curve the following procedure is followed:

Starting with the shortest survival time the interval of time is tabulated in ascending order. At each event (which could be survival or death or censoring) the number alive immediately before is calculated. Before the first event all the subjects are alive, and so the proportion surviving is = 1. If we denote the start time of the study as t_0 then at t_0 we have $S(t_0) = 1$. We can now begin to calculate the survival time for each value of time interval from 1 to n using the following formula

$$S(t_i) = \frac{r_i - d_i}{r_i} \times S(t_{i-1})$$

where d_i is the number of events (deaths) at time t_i and r_i is the number alive just before t_i .

This calculation is done for the events and not for the censored observations. The survival curve is unchanged at the time of the censored observation, but at the next event after the censored observation the number of people 'at risk' is reduced by the number censored between the two events.

If h_0 is the baseline hazard at time t it is written as $h_0(t)$. The model links the hazard for an individual subject at time t (i.e. $h_i(t)$) to the baseline hazard $h_0(t)$ by the following equation

$$\text{Log}_e h_i(t) = \log_e h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

where X_1 and X_2 are variables associated with the subject.

$$\text{Then } h_i(t) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}.$$

The baseline hazard $h_0(t)$ is the reference point and can be thought of as the intercept β_0 in multiple regression. It represents the hazard for an individual for whom $X_1, X_2 \dots X_p$ are all zero. But there is an important difference from the intercept in multiple regression in that it changes with time whereas in multiple regression the intercept remains constant. Unlike the model in logistic regression which yields an odds ratio this model yields relative risk. If there was one binary explanatory variable X_1 then e^β is the relative risk of death for $X = 1$ compared to $X = 0$. Thus Cox's regression is very like logistic regression with a binary outcome variable representing an event, but with the added information on length of time to event.

We assume that anything which affects the hazard does so by the same ratio at all times. Thus something that doubles the risk of an endpoint (e.g. death) on day one will also double the risk of an endpoint on day two, three, and so on. Thus, if $h_0(t)$ is the hazard function for a subject with all the predictor values equal to zero and $h(t)$ is the hazard function for a subject with some other values for the predictor variables then $h(t)/h_0(t)$ depends only on the predictor variables and not on time t . We call $h(t)/h_0(t)$ the **hazard ratio**. It is the relative risk of an end point occurring at time t .

Let us now see how survival analysis works in practice.

The survival time of 49 patients with Duke's colorectal cancer was recorded after randomly assigning patients to either γ linoleic acid or control treatment. The data are tabulated below where for Event 1= Event (i.e. .death); 2= Censored; and for Factor 1 = γ linoleic acid; 2 = Control

The data set is provided below:

ID	Time(ms)	<u>Event</u>		<u>Factor</u>	
		1=dth	1=linol	2=cnsr.	2=cntrl
1	1	2	1		
2	5	2	1		
3	6	1	1		
4	6	1	1		
5	9	2	1		
6	10	1	1		
7	10	1	1		
8	10	2	1		
9	12	1	1		
10	12	1	1		
11	12	1	1		
12	12	1	1		
13	12	2	1		
14	13	2	1		
15	15	2	1		
16	16	2	1		
17	20	2	1		
18	24	1	1		
19	24	2	1		
20	27	2	1		
21	32	1	1		
22	34	2	1		
23	36	2	1		
24	36	2	1		
25	44	2	1		
26	3	2	2		
27	6	1	2		
28	6	1	2		
29	6	1	2		
30	6	1	2		
31	8	1	2		
32	8	1	2		
33	12	1	2		
34	12	1	2		
35	12	2	2		
36	15	2	2		
37	16	2	2		
38	18	2	2		
39	18	2	2		
40	20	1	2		
41	22	2	2		
42	24	1	2		
43	28	2	2		
44	28	2	2		
45	28	2	2		
46	30	1	2		
47	30	2	2		
48	33	2	2		
49	42	1	2		

In Minitab the data columns under “Time(ms.)” and “Event” need to be unstacked prior to analysis using the corresponding values in the column “Factor”, as shown. We would then have two columns for time arranged by linoleic acid and control treatment. We would also have two columns by event arranged by linoleic acid and control treatment as shown below:

Linol. Time	Cntrl Time	Linol. Event	Cntrl. Event
1	3	2	2
5	6	2	1
6	6	1	1
6	6	1	1
9	6	2	1
10	8	1	1
10	8	1	1
10	12	2	1
12	12	1	1
12	12	1	2
12	15	1	2
12	16	1	2
12	18	2	2
13	18	2	2
15	20	2	1
16	22	2	2
20	24	2	1
24	28	1	2
24	28	2	2
27	28	2	2
32	30	1	1
34	30	2	2
36	33	2	2
36	42	2	1
44		2	

[In Minitab Stat → Reliability/Survival → Non-parametric Distribution Analysis – right censored (Right censored because censoring happens as one follows the progress of the subject. Under variables place Linol Time and Cntrl Time. Click on censor button. In the resulting box place Linol.Event Cntrl Event. In the box for censoring value type 1)].

Minitab describes the above type of study as *multiply censored*, meaning thereby that failures (i.e. deaths) were intermixed with censoring. Multiply censored data are more common in clinical practice.

Minitab requires two columns for each group of subjects – one column indicating the time when event occurred, and a corresponding column giving the code with regard to failure and censoring as set out above. The codes for censoring could be numbers or text. If the value indicating censoring is not specified Minitab assumes the lower of the two values (1 in our case) as indicating censoring and the higher value (2 in our case) as failure (i.e. death)

The results of the analysis are given below

```
Variable: Time1 (Linoleic acid)

Censoring Information          Count
Uncensored value              15
Right censored value          10
Censoring value:  EvntLinol = 1

Nonparametric Estimates
```

Characteristics of Variable

Mean(MTTF)	Standard Error	95.0% Lower	95.0% Normal CI Upper
23.7425	3.0506	17.7634	29.7216

Median = 24.0000

IQR = 23.0000 Q1 = 13.0000 Q3 = 36.0000

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Lower	95.0% Normal CI Upper
1.000	25	1	0.9600	0.0392	0.8832	1.0000
5.0000	24	1	0.9200	0.0543	0.8137	1.0000
9.0000	21	1	0.8762	0.0671	0.7447	1.0000
10.0000	20	1	0.8324	0.0767	0.6821	0.9827
12.0000	17	1	0.7834	0.0864	0.6140	0.9528
13.0000	12	1	0.7181	0.1009	0.5204	0.9159
15.0000	11	1	0.6528	0.1109	0.4356	0.8701
16.0000	10	1	0.5876	0.1174	0.3574	0.8177
20.0000	9	1	0.5223	0.1212	0.2848	0.7598
24.0000	8	1	0.4570	0.1224	0.2172	0.6968
27.0000	6	1	0.3808	0.1234	0.1389	0.6227
34.0000	4	1	0.2856	0.1240	0.0427	0.5286
36.0000	3	2	0.0952	0.0880	0.0000	0.2678
44.0000	1	1	0.0000	0.0000	0.0000	0.0000

Empirical Hazard Function

Time	Hazard Estimates
1.000	0.04000
5.0000	0.04167
9.0000	0.04762
10.0000	0.05000
12.0000	0.05882
13.0000	0.08333
15.0000	0.09091
16.0000	0.1000
20.0000	0.1111
24.0000	0.1250
27.0000	0.1667
34.0000	0.2500
36.0000	0.5000
44.0000	1.000

Variable: Time2 (control)

Censoring Information	Count
Uncensored value	12
Right censored value	12

Nonparametric Estimates

Characteristics of Variable

Characteristics of Variable

Mean(MTTF)	Standard Error	95.0% Lower	95.0% Normal CI Upper
24.1185	2.0656	20.0699	28.1671

IQR = 12.0000 Q1 = 18.0000 Q3 = 30.0000

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Lower	95.0% Normal CI Upper
3.0000	24	1	0.9583	0.0408	0.8784	1.0000
12.0000	17	1	0.9020	0.0668	0.7710	1.0000

15.0000	14	1	0.8375	0.0878	0.6655	1.0000
16.0000	13	1	0.7731	0.1020	0.5733	0.9729
18.0000	12	2	0.6443	0.1189	0.4112	0.8773
22.0000	9	1	0.5727	0.1254	0.3269	0.8185
28.0000	7	3	0.3272	0.1289	0.0747	0.5798
30.0000	4	1	0.2454	0.1198	0.0105	0.4803
33.0000	2	1	0.1227	0.1055	0.0000	0.3294

Empirical Hazard Function

Time	Hazard Estimates
3.0000	0.04167
12.0000	0.05882
15.0000	0.07143
16.0000	0.07692
18.0000	0.09091
22.0000	0.1111
28.0000	0.2000
30.0000	0.2500
33.0000	0.5000

Comparison of Survival Curves

Log-Rank Statistic

Variable	1	2
	0.6397	-0.6397

Variance/Covariance of Log-Rank Statistic

Variable	1	2
1	0.1707	-5.8582
2	-5.8582	5.8582

Wilcoxon Statistic

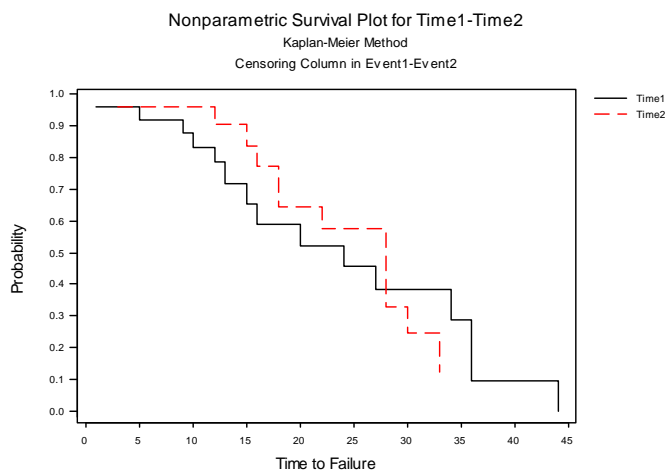
Variable	1	2
	55.0000	-55.0000

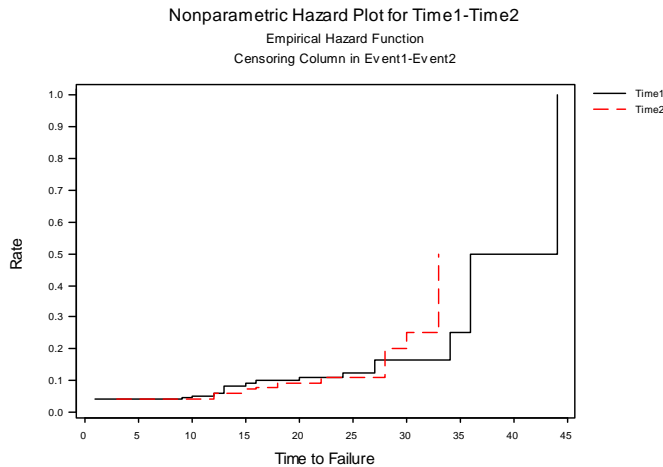
Variance/Covariance of Wilcoxon Statistic

Variable	1	2
1	0.000233	-4291.942
2	-4291.942	4291.942

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	0.06986	1	0.7915
Wilcoxon	0.7048	1	0.4012





Interpreting the results

The first portion of the output gives survival times for those subjects who were on γ -linoleic acid. For the variable γ linoleic acid the mean survival time is 23.7 months (C.I. 17.7 to 29.7).

The different survival estimates are displayed in Kaplan-Meier Estimates, together with the numbers at risk and the number failing. Survival probability for Time = 12.00 (one year) is 78% and for Time = 24.00 (2 years) it is 45 per cent.

Hazard estimates are a measure of the 'hazard function' at the mid-point of the interval. Hazard function is a measure of the failure rate (i.e. death) for each time t .

Similar results follow for the control group. Mean survival time for this group is 24.1 months.

Are the survival curves for γ linoleic acid and control treatment different? In the table of test statistics the P value for both tests (Log-Rank and Wilcoxon) are not significant. P value for Log-rank test is 0.7915 and for the Wilcoxon test it is 0.4012. This table contains measures that tell one if the survival curves for various groups are significantly different.

The log rank test is a form of Chi-square test. It calculates a test statistic for testing the null hypothesis that the survival curves are the same for all groups, in other words, to test the null hypothesis that there is no difference between the populations in the probability of an event (here death) at any time point. For each time point the observed number of deaths in each group and the number expected if there was no difference are calculated. The number of expected is calculated as the proportion of subjects who are at risk at a given time point multiplied by the total number of events at that point.

$$\text{The log rank statistic} = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Variance of (observed} - \text{expected)}}$$

The log rank test is based on the same assumptions as the hazard ratio viz. the survival probabilities are the same for subjects early and late in the study, and the events happened at the time specified. The test is more likely to detect a difference between groups when the risk of an event is consistently greater for one group than another. It is unlikely to detect a difference when survival curves cross. Hence it is useful to plot survival curves when analysing survival data.

The Wilcoxon statistic (also known as the Breslow test) is similarly based on computing the weighted difference between the observed and the expected number of deaths at each of the time point.

Comments

An important distinction between modelling methods is the type of outcome variable. In survival analysis the outcome variable is “time to an event”, and additionally there is censored data.

In linear regression modelling the outcome variable is continuous. In logistic regression the outcome variable is dichotomous. As with linear and logistic modelling, one goal of survival analysis is to obtain a measure of effect that describes the exposure-outcome relationship adjusted for relevant extraneous variables. In linear regression the measure of effect is usually a regression coefficient β . In logistic modelling the measure of effect is an odds ratio measured in terms of an exponential of a regression coefficient i.e. e^β .

In survival analysis the measure of effect is the hazard ratio. It is interpreted in the same way as the odds ratio. A hazard ratio of 1, like an odds ratio of 1, means no effect. A hazard ratio of 3 means that the exposed group has 3 times the hazard of the unexposed group. The higher the hazard ratio the lower is the survival probability, and vice versa. If for an exposed group the hazard ratio is high, the survival probability would be equivalently low. If $h(t)$ is probability of death at time t for subjects that have survived to time t then

$$\log_e \{ h(t)/h(0) \} = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

$$\text{And } h(t) / h(0) = e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

Since the exposure variables (the X 's) are coded as 0 and 1, exposed / not exposed the hazard ratio becomes $e^{\beta_1 + \beta_2 + \dots + \beta_p}$.

The assumptions in Cox's regression of proportional hazards are:

1. Hazard ratio is the same all the way through the study.
2. Any factor which affects death does so by the same ratio at all times.
3. $h(t)/h_0(t)$ is the hazard ratio and is the relative risk of death at time t . The regression coefficients are log hazard ratios. Recall that $h_0(t)$ is the baseline hazard.