A Quick and Gentle Introduction to Survival Analysis

Cancer Prevention and Control Journal Club

June 11, 2009

Introduction

Survival analysis is a general term describing the analytic techniques for data in which the endpoint is the time measured from a defined beginning to the occurrence of a specified event. The endpoint is the *event time*.

- In a cancer clinical trial, the outcome of interest is the survival time of patients from the start of treatment until death.
- In a study of married couples, the outcome of interest is the time from the wedding until the birth of the first child.
- In a study of the carcinogenicity of a chemical, rats are exposed to the chemical and the outcome of interest is the time until a tumor develops.

Censoring

Event-time data are subject to *censoring*. Censoring occurs when the event of interest is not observed for some of the subjects in the study. The most common causes of censoring are

- The subject has not yet had the event when the study is terminated.
- The subject is lost to follow-up or withdrawn from the study.
- The subject dies from causes not relevant to the study.

In general, we assume that censoring is *non-informative*. That is to say, censoring should not convey information about the patient's outcome (event versus non-event).

Types of censoring

- Right-censoring
  - ○ Event occurs *after* a given time point
  - ○ E.g. In a clinical trial with overall survival as the primary endpoint, the subject has not had the event when the study is terminated. For this subject the event is assumed to occur *after* the study's termination.
- Left-censoring
  - ○ Event occurs *before* a given time point
  - ○ E.g. Three months following surgical removal of the primary tumor, patients are examined to see if cancer has recurred. Events for these patients are assumed to have occurred prior to three months post surgery.

Types of censoring (cont.)

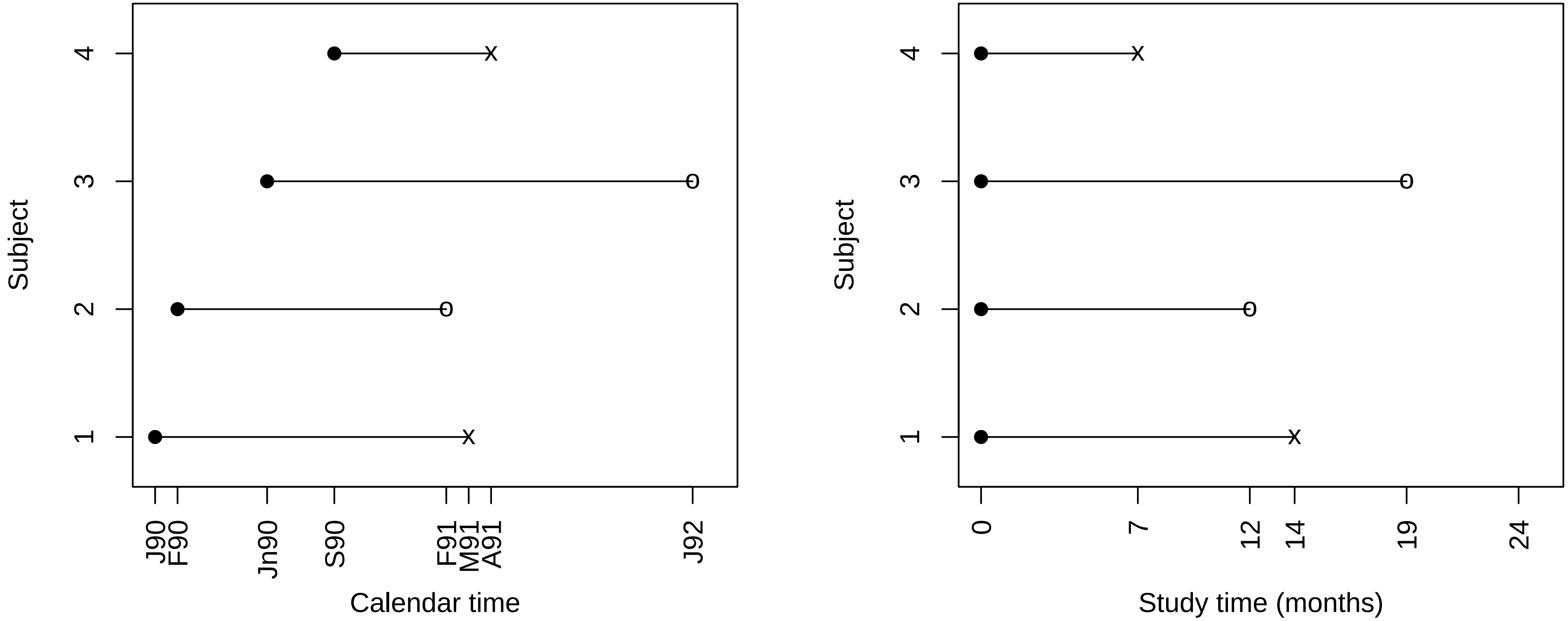
- Interval-censoring
  - Event occurs *between* two time points
  - E.g. In a clinical trial with progression-free survival as the primary endpoint, subjects are examined every three months to determine if cancer has progressed from baseline. At a six-month follow-up visit, a patient has not progressed, but at the nine-month follow-up visit, progression has occurred. The event for this subject is assumed to have occurred between six and nine months.

In this presentation, we will focus on right-censored data.

Event-time data

ID	Entry	End	Time (mos)	Event
1	01/01/90	03/01/91	14	Death
2	02/01/90	02/01/91	12	Lost to FU
3	06/01/90	12/31/91	19	Study ended
4	09/01/90	04/01/91	7	Death

Event-time data depiction



Data issues

- Distribution of event times tends to be positively skewed
  - Some observations have much longer event times than others
  - Non-normal distribution
- Censoring
  - Event times only partially observed
  - Comparison of mean event time between groups not appropriate

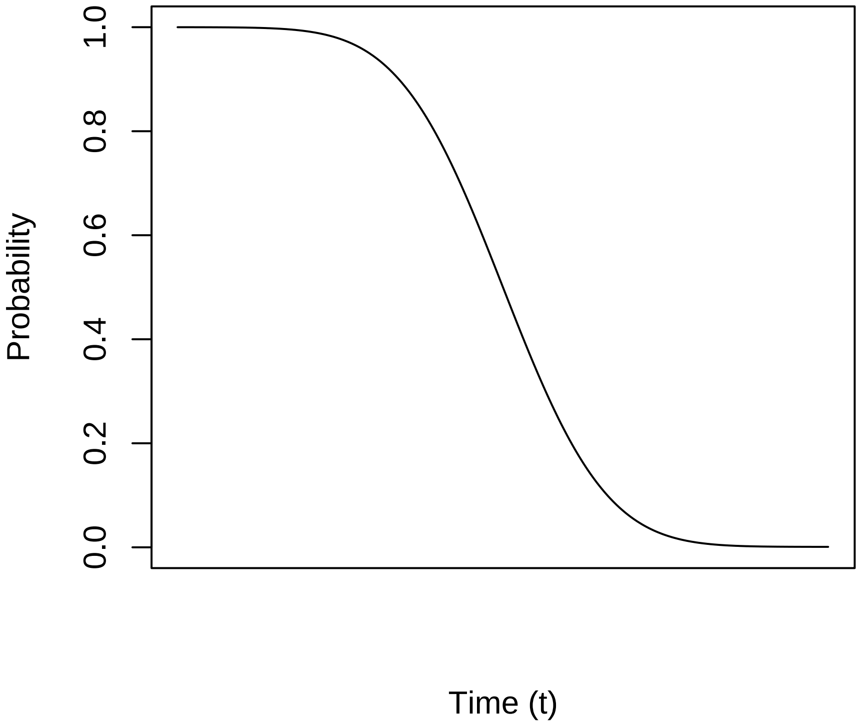
Terminology and notation

- T is the time to the specified event and its observed value for a given subject is denoted as t.
- The *survival function*, $S(t)$, expresses the probability of surviving at least t time units. For example, a “five year survival rate” is simply the probability of surviving at least five years (from some pre-defined time point). The definition of the survival function is

$$S(t) = \text{Prob}(T > t).$$

Properties of $S(t) = \text{Prob}(T > t)$

- Non-increasing function of t
- $S(0) = 1$. In words, at the beginning of observation, no subject has had the event of interest.
- $S(\infty) = 0$. In words, if subjects were observed forever, everyone would eventually experience the event.



Estimation of $S(t)$

The most common estimator of the survival function is the *Kaplan-Meier estimator*, also known as the *Product-limit estimator*. It is a non-parametric estimator of survival, which means that it requires no distributional assumptions about the event times.

We first introduce the following useful terminology and notation.

Kaplan-Meier estimator of $S(t)$

- Let k index the ordered (from smallest to largest) non-censored event times in the data. These event times are represented as t_k.
- The *risk set* at event time t_k refers to the collection of subjects at risk of failure just before time t_k.
- n_k is the size of the risk set associated with event time t_k.
- d_k is the number of events at event time t_k.

Then the Kaplan-Meier estimator of $S(t)$ is

$$\hat{S}_{KM}(t) = \prod_{\{k : t_k \le t\}} \left(1 - \frac{d_k}{n_k}\right).$$

KM estimation - example

Consider the following *ordered* event times for ten subjects. The variable CENSOR is equal to 1 if an event is observed and 0 if the event time is censored.

ID	1	2	3	4	5	6	7	8	9	10
t	3	5	5	7	7	8	10	11	13	13
CENSOR	1	1	0	1	1	0	1	0	1	0

The five observed event times are summarized below.

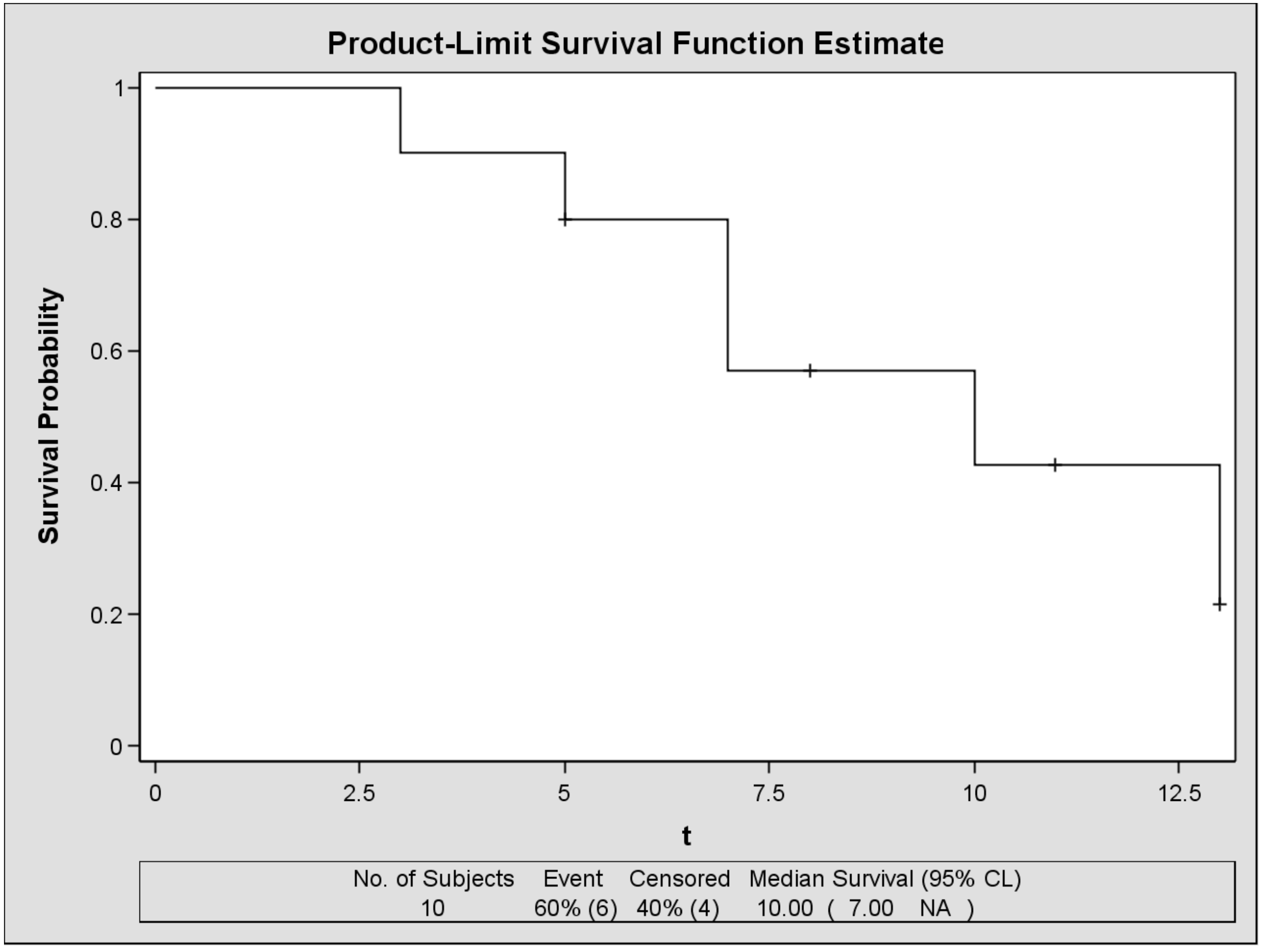
k	t_k	n_k	d_k	d_k/n_k	$1 - d_k/n_k$
1	$t_1 = 3$	10	1	1/10	9/10
2	$t_2 = 5$	9	1	1/9	8/9
3	$t_3 = 7$	7	2	2/7	5/7
4	$t_4 = 10$	4	1	1/4	3/4
5	$t_5 = 13$	2	1	1/2	1/2

KM estimation - example (cont.)

$$\hat{S}_{KM}(t) = \prod_{\{k:t_k \le t\}} \left(1 - \frac{d_k}{n_k}\right),$$

t	$\{k : t_k \le t\}$	$\hat{S}_{KM}(t)$
$t \in [0,3)$	none	1
$t \in [3,5)$	$k = 1$	$9/10 = 0.9$
$t \in [5,7)$	$k = 1,2$	$(9/10)\,(8/9) = 0.8$
$t \in [7,10)$	$k = 1,2,3$	$(9/10)\,(8/9)\,(5/7) \doteq 0.57$
$t \in [10,13)$	$k = 1,2,3,4$	$(9/10)\,(8/9)\,(5/7)\,(3/4) \doteq 0.43$
$t = 13$	$k = 1,2,3,4,5$	$(9/10)\,(8/9)\,(5/7)\,(3/4)\,(1/2) \doteq 0.21$

Graphing KM estimator



Product-Limit Survival Function Estimate

No. of Subjects	Event	Censored	Median Survival (95% CL)
10	60% (6)	40% (4)	10.00 (7.00 NA)

Note $\hat{S}_{KM}(t)$ changes at observed event times and remains constant between observed event times.

German Breast Cancer Study

The German Breast Cancer Study was a randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. The data are described in the handout provided.

Comparing survival functions

Suppose we want to compare survival experiences for subjects with zero or one nodes involved to those with two or more nodes involved. The most common method to compare survival functions is the *log-rank test*.

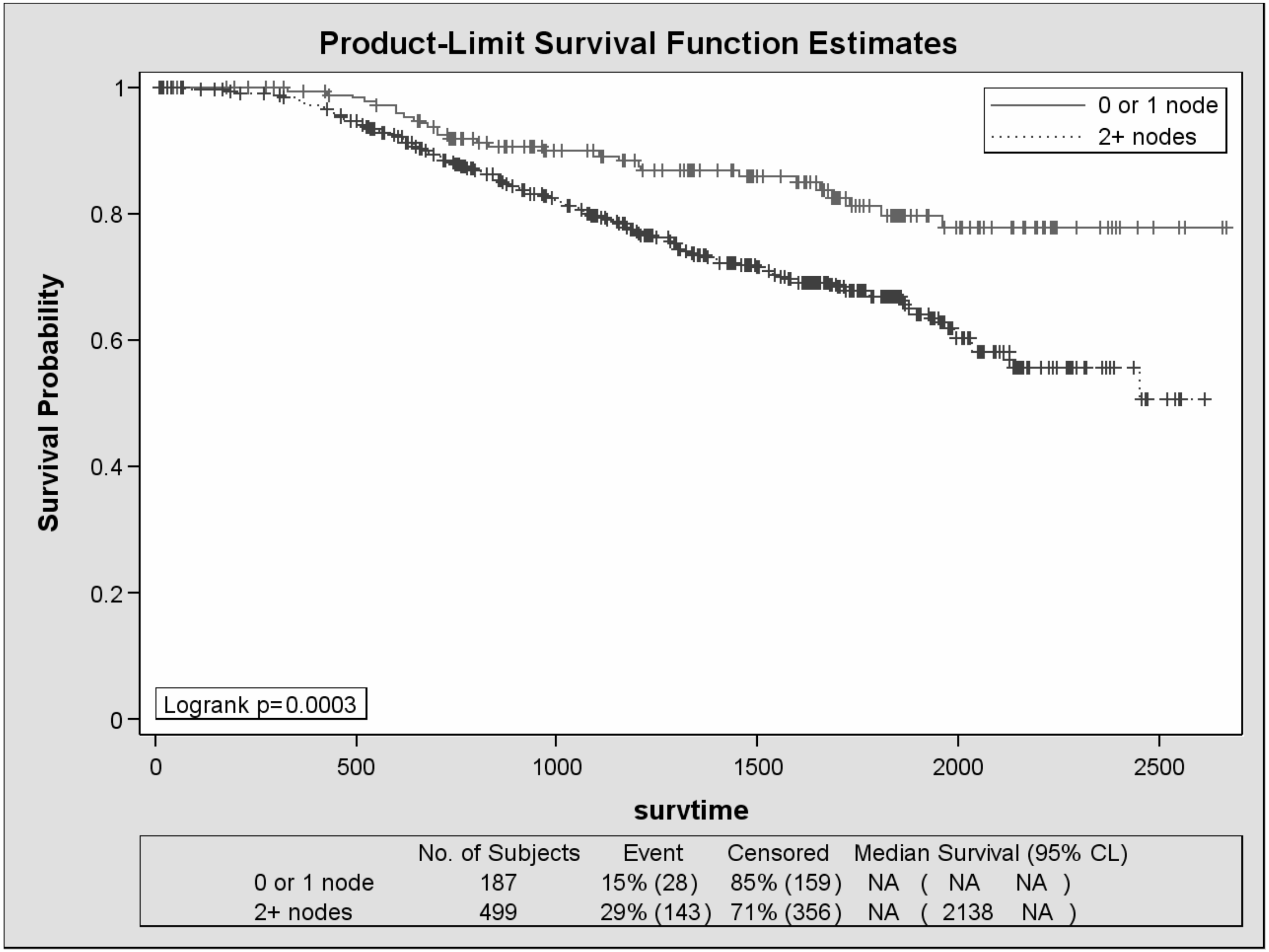
The Log-Rank test

Let $S_1(t)$ be the survival function for Group 1.
Let $S_2(t)$ be the survival function for Group 2.
The log-rank test tests the following null versus alternative hypothesis:

$$\begin{aligned} H_0 : S_1(t) &= S_2(t) \text{ for all } t \\ H_A : S_1(t) &\neq S_2(t) \text{ for at least one } t \end{aligned}$$

The idea behind the log-rank test is to construct a 2×2 contingency table of group membership versus survival for each event time, t. The data from the sequence of tables are accumulated using the Mantel-Haenszel test statistic.

The Log-Rank test (cont.)



	No. of Subjects	Event	Censored	Median Survival (95% CL)
0 or 1 node	187	15% (28)	85% (159)	NA (NA NA)
2+ nodes	499	29% (143)	71% (356)	NA (2138 NA)

Some words of caution

The log-rank test is the most powerful test for the specific alternative

$$H_A : S_1(t) = [S_2(t)]^c,\ c \neq 1.$$

It is not as powerful for other alternatives for which $S_1(t)$ is different from $S_2(t)$. This means that failing to detect a significant difference between the survival functions for two groups can be attributed to any of the following:

1. H_0 is true
2. Lack of power because of inadequate sample size
3. Lack of power due to departure from the assumption of the alternative for which the log-rank test is most powerful.

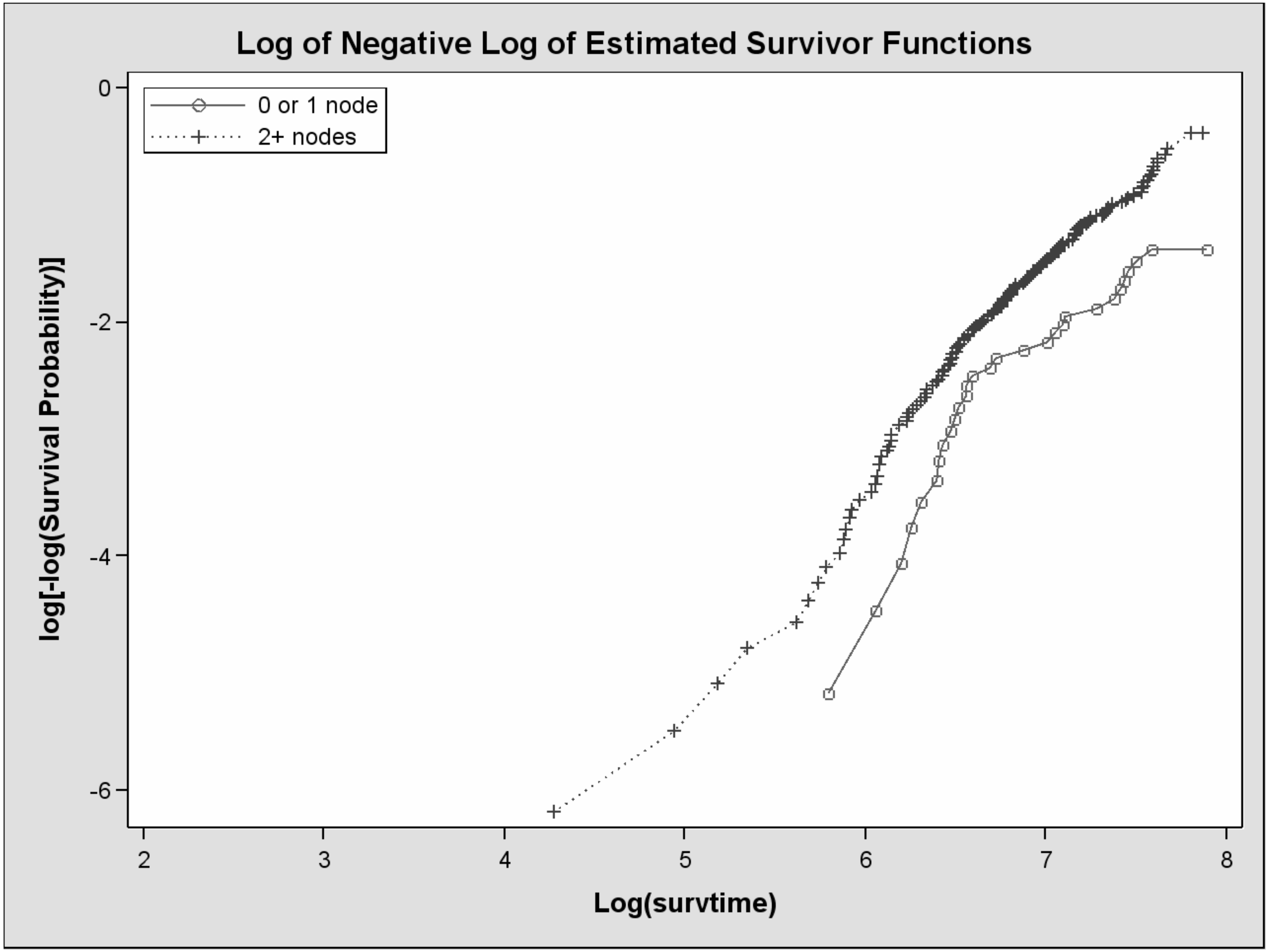
Checking for proportional hazards

$S_1(t) = [S_2(t)]^c,\ c \neq 1$ is known as the *proportional hazards assumption* (more on this later). To assess the validity of this assumption, we use the following fact.

$$\begin{aligned} \log S_1(t) &= c \log S_2(t) \\ \Longleftrightarrow\ -\log S_1(t) &= c(-\log S_2(t)) \\ \Longleftrightarrow\ \log(-\log S_1(t)) &= \log c + \log(-\log S_2(t)) \end{aligned}$$

If we plot $\log\left[-\log[S_1(t)]\right]$ on the same graph with $\log\left[-\log[S_2(t)]\right]$ we should see two curves that are separated by a constant distance, $\log c$. We can construct this plot directly in SAS.

Log-minus-log-survival plots



The hazard function

While the survival function addresses a patient's question about how long they have to live following a diagnosis, the hazard function addresses the question, "How likely am I to die right now?"

The *hazard function*, $h(t)$, is the instantaneous risk of failure at time t for an individual surviving to time t. It is sometimes referred to as the *intensity rate* or the *force of mortality*.

Key relationships

1. $S(t) = \text{Prob}(T > t)$
2. $h(t) = \lim_{\Delta t \to 0} \left[\frac{\text{Prob}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right]$
3. $h(t) = \frac{-dS(t)/dt}{S(t)}$
4. $H(t) = \int_0^t h(u) du$
5. $H(t) = -\ln S(t)$

Proportional hazards revisited

Recall the log-rank test is most powerful for the alternative

$$S_1(t) = [S_2(t)]^c, \ c \neq 1.$$

We said that this assumption was called the *proportional hazards assumption*. We now have the tools to demonstrate where this name comes from.

Proportional hazards revisited (cont.)

As a quick review, recall that two quantities are *proportional* if their ratio is a constant. That is, X is proportional to Y (and vice versa) if $X/Y = c$, where c is some constant.

$$\begin{aligned}
& S_1(t) = [S_2(t)]^c \\
\iff \quad & \ln S_1(t) = c \ln S_2(t) \\
\iff \quad & -\ln S_1(t) = c(-\ln S_2(t)) \\
\iff \quad & H_1(t) = cH_2(t) \\
\iff \quad & \frac{H_1(t)}{H_2(t)} = c
\end{aligned}$$

Proportional hazards revisited (cont.)

Therefore, when we plot $\ln(-\ln S_1(t))$ and $\ln(-\ln S_2(t))$ on the same set of axes, we are actually plotting $\ln(H_1(t))$ and $\ln(H_2(t))$. It follows from properties of logarithms that if $H_1(t)$ and $H_2(t)$ are proportional to one another, then their logarithms should differ by a constant.

Regression models for survival data

We approach the problem of modeling event time via the hazard function. We impose a regression model-type structure on the hazard function that is the product of two components. One factor captures the effect of the event time on the hazard and the second expresses the effects of covariates associated with survival, such as age, race, sex, etc. The form of the model is

$$h(t|X_1, \ldots, X_k) = h_0(t)e^{\beta_1 X_1 + \ldots + \beta_k X_k}.$$

Regression models for survival data (cont.)

$$h(t|X_1, \ldots, X_k) = h_0(t)e^{\beta_1 X_1 + \ldots + \beta_k X_k}$$

- $h_0(t)$ is called the *baseline hazard*. It characterizes how the hazard function changes as a function of time.
- $e^{\beta_1 X_1 + \ldots + \beta_k X_k}$ characterizes how the hazard function changes as a function of covariates.
- $h(t|\mathbf{X})$ is referred to as the *Cox model* or *Cox proportional hazards model* or simply the *proportional hazards model*.
- $h(t|\mathbf{X})$ is linear in the covariates on the log scale. That is, $\ln h(t|\mathbf{X}) = \ln h_0(t) + (\beta_1 X_1 + \ldots + \beta_k X_k)$

Why “proportional hazards”?

The proportional hazards assumption implies that the ratio of the instantaneous failure rates for two subjects is a constant. To see why this assumption is implicit in the form of the model, consider two subjects, A and B with covariates $\mathbf{X}_A$ and $\mathbf{X}_B$, respectively. Then

- $h(t|\mathbf{X}_A) = h_0(t)e^{\mathbf{X}'_A\beta}$
- $h(t|\mathbf{X}_B) = h_0(t)e^{\mathbf{X}'_B\beta}$

so that

$$\frac{h(t|\mathbf{X}_A)}{h(t|\mathbf{X}_B)} \quad = \quad \frac{h_0(t)e^{\mathbf{X}'_A\beta}}{h_0(t)e^{\mathbf{X}'_B\beta}} = \frac{e^{\mathbf{X}'_A\beta}}{e^{\mathbf{X}'_B\beta}}$$

Why “proportional hazards”? (cont.)

Since $\frac{e^{\mathbf{x}'_A \beta}}{e^{\mathbf{x}'_B \beta}}$ is just a constant, the hazards for the two subjects are proportional to one another. Notice that the ratio of their hazards does not depend on time. The proportional hazards assumption means we assume that the ratio of the hazards is constant over time.

Multivariable models

Consider modeling the hazard of death as a function of

- hormone treatment (Yes or No)
- nodal involvement (0,1 versus 2+)
- tumor grade (1,2,3)

The model is

$$h(t|\text{hormone, node, grade}) = h_0(t)e^{\beta_1 \text{HM} + \beta_2 \text{ND} + \beta_3 \text{G2} + \beta_4 \text{G3}}.$$

(*Segue to fitting in SAS*)

Interpretation of parameter estimates

The fitted model is

$$h(t|\text{hormone, node, grade}) = h_0(t)e^{-0.25\text{HM}+0.74\text{ND}+1.23\text{G2}+1.84\text{G3}}.$$

We can use the fitted model to obtain estimates of hazard ratios. Specifically, suppose we want to compare the hazard of failure for subjects with and without hormone treatment, controlling for nodal involvement and tumor grade.

- For the subject randomized to hormone treatment, $h(t|\text{HM=1, node, grade}) =$ $h_0(t)e^{-0.25\times 1+0.74\text{ND}+1.23\text{G2}+1.84\text{G3}}$.
- For the subject randomized to no hormone treatment, $h(t|\text{HM = 0, node, grade}) =$ $h_0(t)e^{-0.25\times 0+0.74\text{ND}+1.23\text{G2}+1.84\text{G3}}$.

Interpretation of parameter estimates

- The hazard ratio is

$$\frac{h(t|\text{HM = 1, node, grade})}{h(t|\text{HM = 0, node, grade})} = \frac{h_0(t)e^{-0.25+0.74\text{ND}+1.23\text{G2}+1.84\text{G3}}}{h_0(t)e^{0.74\text{ND}+1.23\text{G2}+1.84\text{G3}}}$$

$$= e^{-0.25} \doteq 0.78.$$

In words, there is a 22% reduction in the hazard of death for subjects randomized to hormone treatment relative to those randomized to the no hormone treatment arm, after controlling for nodal involvement and tumor grade. However, the p-value corresponding to the chi-square test for the parameter is non-significant (p = 0.13), so this reduction in the hazard is not statistically significant.

Hazard ratios for categorical predictors

In general, let X be a k-level categorical variable. Let $Z_1, Z_2, \ldots, Z_{k-1}$ be the $k-1$ dummy variables (reference cell coding) associated with X. Assume β_j is the regression coefficient of Z_j obtained from a Cox-PH regression model, where $j = 1, \ldots, k-1$. Then e^{β_j} is the hazard ratio comparing the jth level of X to the reference level, controlling for the other variables in the model.

Confidence intervals for hazard ratios

A 95% confidence interval for the HR is simply

$$e^{\hat{\beta} \, \pm \, 1.96 \, \widehat{SE}(\hat{\beta})}.$$

Therefore, from the SAS output, a 95% CI for the HR of death for those randomized to hormone treatment relative to those not randomized to hormone treatment, is

$$e^{-0.24563 \, \pm \, 1.96 \times 0.16380} \doteq (0.57, 1.08).$$

Because the CI contains the null value of 1, we conclude the difference in the hazard of death for subjects in these two arms is not statistically significant.

HRs and CIs for continuous model covariates

Consider inclusion of the variable SIZE. The default HR and corresponding 95% CI are 1.02, 95% CI = 1.01 to 1.03, with p = 0.0002. Clearly tumor size is strongly associated with the hazard of death, but the HR and 95% CI do not convey this. This is because the default HR indicates the change in the hazard for a *unit increase* in tumor size, a change that is not likely to have any clinical value.

Constructing other HRs and CIs

To compute the hazard ratio and corresponding 95% CI comparing subjects with values for a continuous covariate that differ by a fixed amount, say Δx, we use the following principle:

$$\widehat{\text{HR}} = e^{\Delta x \hat{\beta}}$$

and the 95% CI is

$$(e^{\Delta x \hat{\beta} - 1.96 \times |\Delta x| \times \text{SE}(\hat{\beta})}, e^{\Delta x \hat{\beta} + 1.96 \times |\Delta x| \times \text{SE}(\hat{\beta})}).$$