Notes for Tuesday March 31, 1998

Survival Analysis

The term *survival analysis* originated in the study and analysis of times to death (that is, survival times) for medical patients diagnosed with some fatal disease. Survival analysis is now a well-developed field of statistical research and methodology pertaining to modeling and testing hypotheses of failure time data for humans as well as animals, machines, electronic equipment, automobile components, etc. Hence, the methodology is far more general than the analysis of survival times. In fact, fields of study other than medicine have given other names to the identical methodology discussed here, such as:

- Analysis of Failure Time Data
- Reliability Analysis
- Event History Analysis

However, the focus of most of the examples and the history of the development of this material is on medical and actuarial problems, so we will use the *survivial analysis* terminology (patients, people or items and the times at which they die/fail/etc.). This helps to simplify the presentation.

Modeling of survival times is based on two distinct approaches: *parametric* and *nonparametric*. Nonparametric methods are widely used in clinical trials in medicine, and we will concentrate on these, especially Kaplan-Meier estimates of survival and Cox proportional hazards regression models. Miller (1981) and Kalbfleisch and Prentice (1980) are excellent references; important extensions are due to Andersen and Gill (1982).

Some parametric models can be fitted with the `glm()` function that we have already become familiar with from log-linear and logistic regression models; other parametric models require special modifications of the `glm` approach. We will illustrate these methods only briefly.

Some useful references

- Rupert G. Miller Jr., *Survival Analysis*, Wiley 1981.
- John Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley, 1980. [very readable]
- Miller, Efron, et. al, *Biostatistics Casebook*, Wiley, 1980.
- Venables and Ripley, Chapter 12. [Not bad, on this topic!]
- *Encyclopedia of Statistical Sciences* [look for topics like "Kaplan-Meier estimate", "Cox regression model", "Greenwood formula", etc.]
- The *Splus 4.0 Guide to Statistics* has several chapters devoted to survival analysis.

The Problem

We observe a **positive valued** random variable (the survivor time or failure time) and some covariates. Typically the value of the random variable is the time to failure of some physical component or the time to death of a biological unit (or the time to relapse of some disease, etc.). There are all sorts of examples of waiting times which can be similarly analyzed—e.g., the length of time required to learn a new skill.

Early work in this area was some of the first (modern) statistics ever recorded. Graunt and Edmund Halley were responsible for the first useful **life tables** (used to estimate the *force of mortality*, or instantaneous risk of dying) ca. 1660. Medical and insurance applications carried the work steadily forward.

The modern era was ushered in by the need to test mechanical components for reliability. Rapid industrial expansion in the early 1940's was a major impetus.

The most novel feature of failure-time data is the notion of *censoring*—not all sample units (e.g., patients in a medical trial) will be observed to failure (e.g., death), yet even the "partially" observed data provide useful information on the lifetime distribution.

Example *Carcinogenesis.* Two groups of rats were given pre-treatments then exposed to a carcinogen. We would like to compare the pre-treatments. Presumably, we could transform the data and perform a two-sample t-test, except there is some censoring in the data. Some of the rats died of causes unrelated to the cancer, or survived the study.

Days to Vaginal Cancer Mortality in Rats

Group 1	143	164	188	188	190	192	206	209	213	216
	220	227	230	234	246	265	304	216*	244*	
Group 2	142	156	163	198	205	232	232	233	233	233
	233	239	240	261	280	280	296	296	323	204*
	344*									

The items with (*) are those that are censored. This is like a form of missing data except the "missing-ness" of the data contains information about the questions of interest.

Example *Accelerated Life Test.* Many physical components are **designed** to be very reliable and have long lifetimes. (and then there are the other components). It is still of interest to know something about the expected lifetime of a component or a comparison between differing designs. When the lifetime is very long, it is often difficult to actually test the component. (Think about a computer chip. If you tested the lifetime under normal wear-and-tear, the chip would be obsolete before you saw any failures!).

In these circumstances and "accelerated life test" is often performed. This consists of an experiment where the components are subjected to *more* stress than would be expected under normal operating conditions with the aim of extrapolating or comparing.

For a certain type of motor designed to operate at 130 degrees C, the following severely censored data were obtained.

150C All 10 motorettes without failure at 8064 hours

170C 1764, 2772, 3444, 3542, 3780, 4860, 5196, 3 motorettes without failure at 5448 hours

190C 408, 408, 1344, 1344, 1440, 5 motorettes without failure at 1680 hours

220C 408, 408, 504, 504, 504, 5 motorettes without failure at 528 hours

How would you analyze these data? Suppose you wanted to compare two types of motor?

Example *African Children.* A Stanford psychiatrist wanted to know the age at which a group of African children learned to perform a particular task—so he wrote a grant proposal and went to Africa. When he arrived in the village some of the children could already perform the task, some children learnt while he was there and some hadn't learnt when he left. In this case there is both left and right censoring. For some individuals we know an age by which they could perform the task (left censoring); for some we know an age at which they could not perform the task, and for the rest we know the age at which they learned the task.

Survival Functions and Hazard Rates

Let the random variable $T > 0$ denote survival time. Let T have density $f(t)$ and distribution

$$F(t) = \text{Prob}(T \leq t)$$

(Note that some books use $F(t) = \text{Prob}(T > t)$ which makes sense but is not standard).

The survival function is

$$S(t) = 1 - F(t) = P(T > t)$$

and the hazard rate (function), also called the instantaneous risk, or force of mortality, is

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

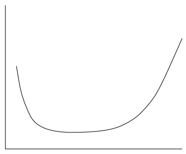
which can be interpreted as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{1 - F(t)}$$

Note that the numerator is

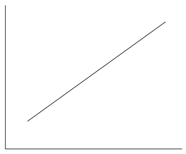
$$\text{Prob}\{\text{ dying in } (t, t + \Delta t) \mid \text{lived to time } t\ \}$$

Here are some typical shapes for hazard rate functions:



(a)

(b)



(c)

Figure (a) shows the hazard of dying as a function of time, for humans. Figure (b) shows a constant hazard rate; this model is often used for the lifetimes of electronic parts, well-maintained machinery, and the like. It leads to a the *exponential distribution* for failure times. Figure (c) shows an increasing hazard rate. Events that are more likely as time goes on have this hazard rate.

Now, notice that if we integrate $\lambda(t)$ we get

$$\begin{aligned}\int_0^t \lambda(u)du &= \int_0^t \frac{f(u)}{1-F(u)}du \\ &= -\log\left[1-F(u)\right]_0^t \\ &= -\log\left(S(t)\right)\end{aligned}$$

Thus

$$S(t) = \exp\{-\int_0^t \lambda(u)du\}$$

The function $\Lambda(t) = -\log S(t) = \int_0^t \lambda(u)du$ is called the *cumulative hazard function*; some estimates of $S(t)$ are based on estimating $\Lambda(t)$ first and then exponentiating.

These various forms, definitions and **ideas** are very important. If you don't understand them now, stop me.

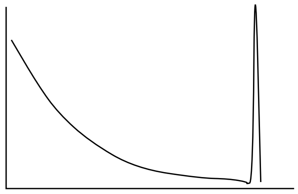
Types of Censoring

Let $T_1, \ldots, T_n$ be iid F.

- Type I Censoring. Let t_c be some preassigned fixed number, the fixed censoring time. Instead of seeing $T_1, \ldots, T_n$ we get to see

  $$Y_i = \min\left(T_i, t_c\right)$$

  so the density of Y might look like



- Type II Censoring. Let $r < n$ be fixed. All we observe are $T_1, \ldots, T_r$.

Both types of censoring are common—either we have only a certain amount of time in which to experiment (I) or we don't know how long we need to go but know how what percentage of failures we need to wait for (II).

- Type III, or Random Censoring. Let $C_1, \ldots, C_n$ be iid G were C_i is the censoring time associated with T_i. We observe

  $$Y_i = \min(T_i, C_i)$$

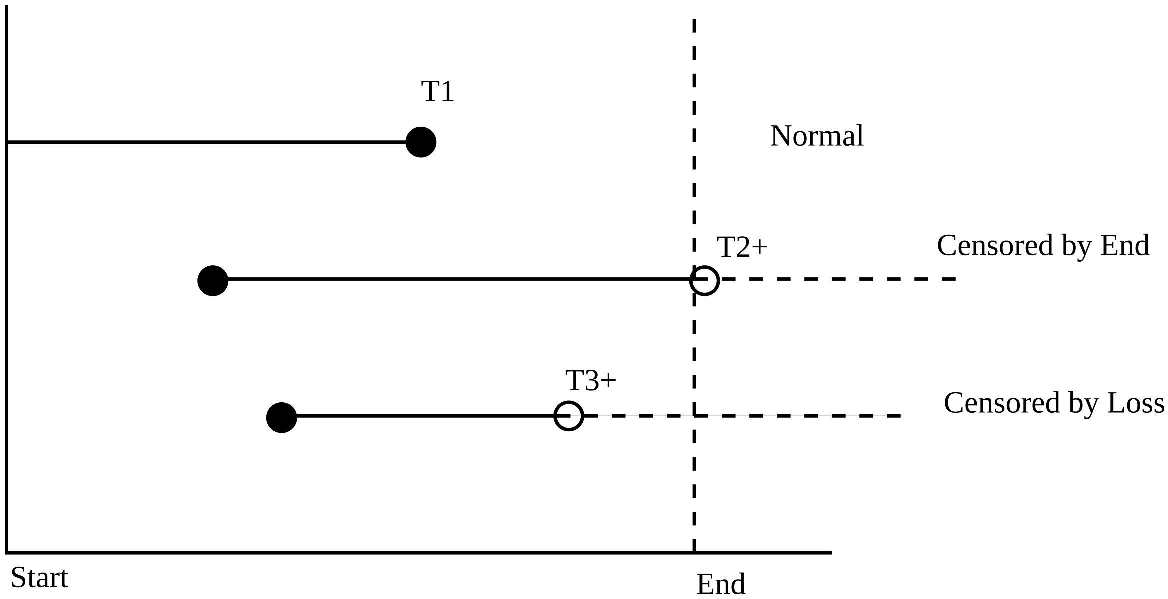
  and

  $$\delta_i = 1_{(T_i \le C_i)} = \begin{cases} 1 & \text{if } T_i \text{ not censored} \\ 0 & \text{if } T_i \text{ is censored} \end{cases}$$

Many medical examples have people enter the study at different times and then exit the study because of

1. loss to follow up (they move, etc., without telling us, and we can't find them)
2. drop out (they leave the study due to side effects, contrariness, etc., and they tell us so)
3. end of study.

The following picture shows what might happen in a study.



One might also imaging left censoring, where the start date is only know to be after a certain date.

Some distributions for Survival Analysis

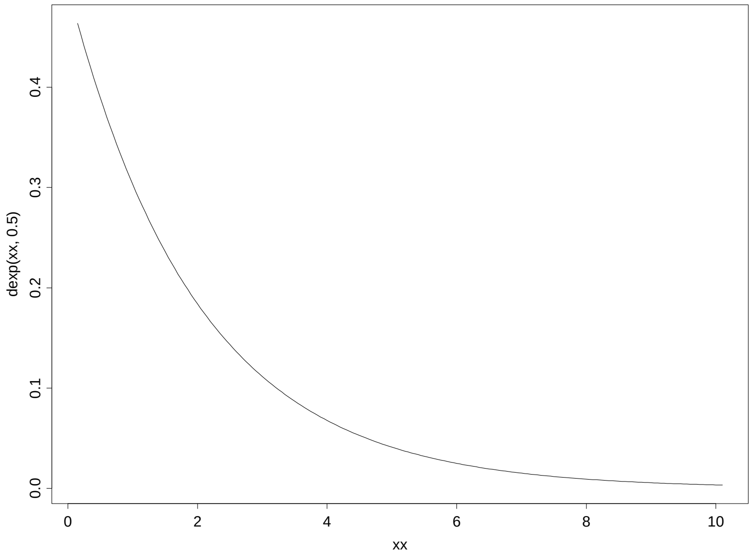
Exponential This is the basic model with constant risk, $\lambda(t) = \lambda > 0$. Thus

$$\int_0^t \lambda(u)du = \lambda \times t$$

and so

$$\begin{aligned} S(t) &= \exp \int_0^t \lambda(u)du = e^{-\lambda t} \\ F(t) &= 1 - e^{-\lambda t} \\ f(t) &= \lambda e^{-\lambda t} \end{aligned}$$

and thus $E(T) = 1/\lambda$ and $\mathrm{Var}(T) = 1/\lambda^2$. Here is a plot of the density.



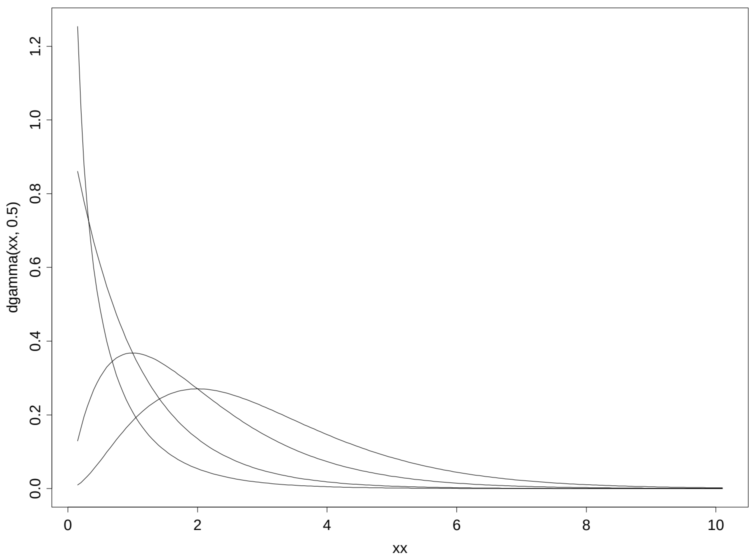
If you have an estimate $\hat{S}(t)$ of the survival function, a reasonable way to check to see if the exponential distribution applies might be to plot $\log(\hat{S}(t))$ against t. It should be a straight line.

Gamma One possible generalization of the exponential.

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{\lambda t} \quad \alpha > 0 \quad \lambda > 0$$

$$\begin{aligned} F(t) &= \text{Incomplete Gamma Integral, no closed form} \\ S(t) &= 1 - F(t) \\ \lambda(t) &= 1 - e^{-\lambda t} \end{aligned}$$

$E(T) = \alpha/\lambda$ and $\mathrm{Var}(T) = \alpha/\lambda^2$. Here is a plot of the density, for $\alpha = 0.5, 1, 2, 3$



You really already know a lot about the Gamma distribution because

$$\chi^2_k = \text{Gamma}(k/2, 1/2)$$

that is, $\alpha = k/2$ and $\lambda = 0.5$.

Weibull Another generalization of the exponential.

$$S(t) = e^{-(\lambda t)^\alpha} \quad \alpha > 0 \quad \lambda > 0$$

This can be forced to look a LOT like a gamma distribution. The Weibull is very important, especially in *accelerated life testing*, but we will not dwell on it.

With complete (i.e., uncensored) data estimation of any of these distributions is relatively easy. Just write out the log-likelihood and differentiate. For some of the models there are even closed-form estimates. We will defer looking at the parametric ML estimates until we have looked at a non-parametric estimate of the survivor function.

Nonparametric Estimation of the Survivor Function

Remember what the survivor function is, just the probability of surviving to be older than time t. A convenient and familiar object is the sample cumulative distribution function (cdf), written as

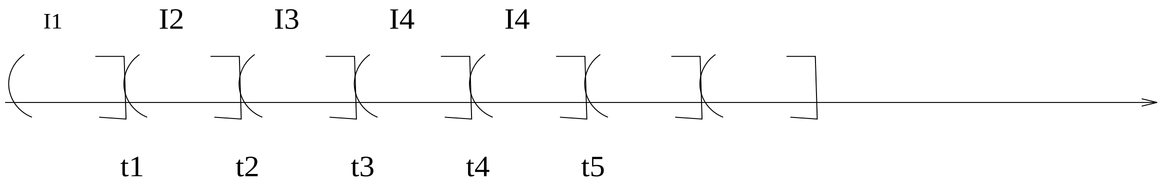
$$\bar{F}_n(x) = \frac{\#\text{ of obs} \leq x}{n}$$

Estimates of the sample survival function are obviously of the form, $1 - \bar{F}_n(x)$. In the analysis of survival data it is often useful to summarize the survival experience in terms of $\bar{F}_n(x)$ or $\bar{S}_n(x)$. (Why? Well it might direct us to an appropriate parametric model, or it might be interesting in its own right). When the data is censored we need to think of something other than the simple cdf.

Life Tables The classical method of estimating $S(t)$ is the so-called actuarial method which depends on the life table.

Let time be partitioned into a fixed sequence of intervals, $I_1, \ldots, I_k$ mostly of equal length. The intervals should be open at the left and closed at the right.

Consider the line



with

$$
\begin{aligned}
n_i &= \#\text{ alive at beginning of } I_i \\
d_i &= \#\text{ died in } I_i \\
l_i &= \#\text{ lost to follow-up in } I_i \\
w_i &= \#\text{ withdrawn in } I_i \\
p_i &= P(\text{ surviving through } I_i |\text{ alive at start of } I_i)
\end{aligned}
$$

As an example, look at the data below, where we would like to estimate $S(5)$ say.

Years after diag	Alive at start	Died	Lost to Followup	Withdrawn
0–1	126	47	4	15
1–2	60	5	6	11
2–3	38	2	—	15
3–4	21	2	2	7
4–5	10	—	—	6

Reduced Sample Method To estimate $S(\tau_k)$ we only use those subjects who are at risk for the entire period $(0, \tau_k]$. That is then

$$
\begin{aligned}
n &= n_1 - \textstyle\sum_{i=1}^k l_i - \sum_{i=1}^k w_i & \qquad n &= 126 - 12 - 54 = 60 \\
d &= \textstyle\sum_{i=1}^k d_i & d &= 56 \\
\text{and } S(\tau_k) &= 1 - d/n & S(5) &= 1 - \tfrac{56}{60} = 0.067
\end{aligned}
$$

What (if anything) is wrong with this? It throws away information, all the information in l_i and w_i, and it is thus an inefficient (and biased) estimate.

Actuarial Method The survival probability $S(\tau_k)$ can be broken into a product of probabilities,

$$\begin{aligned} S(\tau_k) &= P(T > \tau_k) \\ &= P(T > \tau_1) \cdot P(T > \tau_1 | T > \tau_2) \cdots P(T > \tau_k | T > \tau_{k-1}) \\ &= p_1 \cdot p_2 \cdot p_3 \cdots p_k \\ \text{where} & \\ p_i &= P(T > \tau_i | T > \tau_{i-1}) \end{aligned}$$

The actuarial method estimates each p_i separately and multiplies to get $S(\tau_k)$. To estimate p_i one could use

$$1 - \frac{d_i}{n_i}$$

but this ignores losses and withdrawals in interval I_i. A sensible(?) approach might be to assume that on average those that were lost or withdrew were at risk for 1/2 of the interval. Thus one might define the effective sample size as

$$n'_i = n_i - 1/2 \cdot (l_i + w_i)$$

with

$$\begin{aligned} \hat{q}' &= \frac{d_i}{n'_i} \\ \hat{p}_i &= 1 - q_i \\ \text{and } \hat{S}(\tau_k) &= \prod_{i=1}^{k} \hat{p}_i \end{aligned}$$

For the above data, this yields,

$$\hat{S}(5 \text{ years}) = 0.44$$

Variance of $\hat{S}(\tau_k)$ By assuming that

$$n'_i \hat{p}_i \approx \text{Binomial}(n'_i, p_i)$$

and some manipulations an estimate of the variance of the estimate can be found as

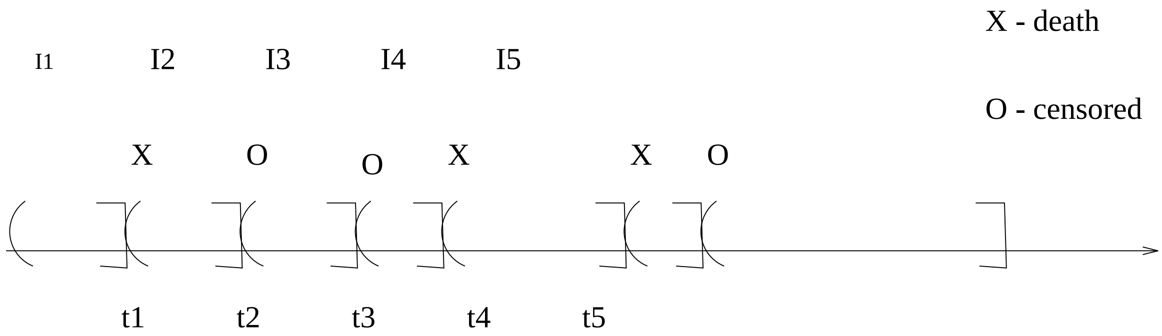
$$\widehat{\text{Var}}(\hat{S}(\tau_k)) = \hat{S}^2(\tau_k) \sum_{i=1}^{k} \frac{d_i}{n'_i(n'_i - d_i)}$$

This is known as Greenwood's formula. By this formula,

$$\sqrt{\text{Var}(\hat{S}(5))} \approx 0.06$$

The Kaplan-Meier Product Limit Estimate

This is very similar to the actuarial method we just saw, except that the intervals I_i are not fixed in advance. In fact the intervals are determined by when an event occurs. So a revised picture might be,



Recall that we observe

$$\begin{aligned} Y_1, \ldots, Y_n &= \text{the survival times, and} \\ \delta_1, \ldots, \delta_n &= \begin{pmatrix} 1 & \text{censored} \\ 0 & \text{not censored} \end{pmatrix} \end{aligned}$$

Order the Y_j into

$$\begin{gathered} Y_{(1)}, \ldots, Y_{(n)} \\ \delta_{(1)}, \ldots, \delta_{(n)} \end{gathered}$$

(the $\delta_{(j)}$ are the δ_j written in the same order as $Y_{(j)}$). Let $\mathcal{R}(t)$ be the risk set at t and

$$\begin{aligned} n_j &= \#\text{ in } \mathcal{R}(Y_{(j)}) = \#\text{ alive just before } Y_{(j)} \\ d_j &= \#\text{died at } Y_{(j)} \\ p_j \text{ and } q_j &= \text{same as before—conditional probability of surviving} \end{aligned}$$

Then

$$\hat{p}_j = \begin{cases} 1 - 1/n_j & \text{if } \delta_{(j)} = 1 \\ 1 & \text{if } \delta_{(j)} = 0 \end{cases}$$

Thus, without ties, the PL estimate is

$$\begin{aligned} \hat{S}(t) &= \prod_{y(j) \leq t} \hat{p}_j = \prod (1 - 1/n_j)^{\delta(j)} \\ &= \prod_{y(j) \leq t} \left[1 - \frac{1}{(n - j + 1)} \right]^{\delta(j)} \end{aligned}$$

A View of the PL estimate as an MLE

The PL estimate would not be nearly so popular if it could not be interpreted in other ways.

Let $t_1, \ldots, t_k$ be the observed failure times in a sample of size $n_0 = n$ with survivor function, $S(t)$. Suppose d_j items fail at time t_j and m_j items are censored in $[t_j, t_{j+1})$ at known times $t_{j_1}, t_{j_2}, \ldots, t_{j_{m_j}}$. Let $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ be the number of items at risk at time just before t_j.

The probability of failure just exactly at time t_j is

$$S(t_j) - S(t_j + 0).$$

If the only information that censoring at time t_{j_l} gives us is that the failure time for that item (death of that patient...) was greater than t_{j+l} then we may assume that $P(T > t_{j_l}) = S(t_{j_l} + 0)$.

Thus, the probability of seeing failures at exactly the times t_j and censored observations at exactly the times t_l should be

$$\mathcal{L} = \prod_{j=0}^{k} \left\{ [S(t_j) - S(t_j + 0)]^{d_j} \cdot \prod_{l=1}^{m_j} S(t_{j_l} + 0) \right\}$$

This can be viewed as a likelihood function on the space of all survivor functions, $S(t)$. There are problems with this definition of the MLE, but the intuition is good.

Clearly if $S(t)$ is *continuous*, $\mathcal{L} \equiv 0$. So to find a useful MLE one needs an S which is *discontinuous* at the observed failure times.

Nothing about the data or the likelihood gives any information about the values of $S(t)$ *between* t_j's, so we might as well assume it is constant on these intervals. Thus, we seek an estimator of $S(t)$ of the form

$$\hat{S}(t_j) = \prod_{l=1}^{j-1} (1 - \hat{q}_l) \qquad \text{and} \qquad \hat{S}(t_j + 0) = \prod_{l=1}^{j} (1 - \hat{q}_l)$$

where the $q_i = P[\text{ death } \leq t_i \mid t_{i-1} < \text{death }]$ are the constant hazards of death between t_{i-1} and t_i. It then follows that

$$S(t_j) - S(t_j + 0) \;=\; \prod_{l=1}^{j-1} (1 - q_l) - \prod_{l=1}^{j} (1 - q_l) \;=\; q_j \prod_{l=1}^{j-1} (1 - q_l)$$

Substituting these expressions into $\mathcal{L}$ above, we should choose the q's so as to maximize the function

$$\begin{aligned} \mathcal{L} \;&=\; \prod_{j=1}^{k} \left\{ \left[\prod_{l=1}^{j-1} (1 - q_l)^{d_j} \right] q_j^{d_j} \left[\prod_{l=1}^{j} (1 - q_l)^{m_j} \right] \right\} \\ &=\; \prod_{j=1}^{k} q_j^{d_j} (1 - q_j)^{n_j - d_j} \end{aligned}$$

Obviously the maximum is achieved when $\hat{q}_j = \frac{d_j}{n_j}$ and the product limit estimate is

$$\hat{F}(t) = \prod_{j:\ t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:\ t_j < t} \left(1 - \frac{d_j}{n_j} \right)$$

When $d_j \equiv 1$, this reduces to the previous form of the PL estimate. Note that this form also "takes care of" ties naturally:

1. Tied Uncensored observations—do the obvious thing which is $p_j = 1 - \frac{d_j}{n_j}$ where d deaths occur.

2. Censored and uncensored tied data—consider the uncensored data to be just before the censored data.

Interpretation We are in effect making the conditional probability of failure at each t_j agree exactly with the observed conditional relative frequency of failure at t_j.

Problem: Even if we assume that $S(t)$ is continuous, our estimate $\hat{S}(t)$ is always bumpy, which might give you cause for concern. (This can all be fixed by appropriate technical mumbling).

Example

No Censoring Assume that $n = 5$ and $x =$ death and $o =$ censor, with

$$\begin{array}{cccccc} & x & x & x & x & x \\ p_i & (1-\frac{1}{5}) & (1-\frac{1}{4}) & (1-\frac{1}{3}) & (1-\frac{1}{2}) & (1-\frac{1}{1}) \end{array}$$

where p_i is the $p($ survive to age $i \mid$ alive before time $i)$

$$\begin{array}{cccccc} i & = & 1 & 2 & 3 & 4 & 5 \\ S_i & = & \frac{4}{5} & \frac{4}{5}\cdot\frac{3}{4} & \frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3} & \dots & \\ & = & \frac{4}{5} & \frac{3}{5} & \frac{2}{5} & \frac{1}{5} & 0 \end{array}$$

Thus, without censoring the PL estimate is the obvious one

Some censored data Let $n = 7$ and take the data

$$\begin{array}{ccccccc} & x & x & o & x & o & x \\ p_i & \frac{6}{7} & \frac{5}{6} & & \frac{3}{4} & & \frac{1}{2} \end{array}$$

The PL estimate is

$$\begin{array}{cccccccc} i & = & 1 & 2 & 3 & 4 & 5 & 6 & \\ S_i & = & \frac{6}{7} & \frac{6}{7}\cdot\frac{5}{6} & \frac{6}{7}\cdot\frac{5}{6} & \frac{6}{7}\cdot\frac{5}{6}\cdot\frac{3}{4} & \dots & & \\ & = & \frac{6}{7} & \frac{5}{7} & \frac{5}{7} & \frac{15}{28} & \frac{15}{28} & \frac{15}{56} & \frac{15}{56} \end{array}$$

Example: A Leukemia Study

The data presented in the table below are preliminary results from a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML). The study was conducted by Embury et al. (1977) at Stanford University. After reaching a status of remission through treatment by chemotherapy, the patients who entered the study were assigned randomly to two groups. The first group received maintenance chemotherapy; the second, or control, group did not. The objective of the trial was to see if maintenance chemotherapy prolonged the time until relapse.

Data for AML maintenance study. A + indicates a censored value.

Group	Length of complete remission (in weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Non − maintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

This data set is built into SPLUS. It looks like this (note the coding of the censoring variable, `status`).

```
402 > leukemia
   time status        group
 1    9      1    Maintained
 2   13      1    Maintained
 3   13      0    Maintained
 4   18      1    Maintained
 5   23      1    Maintained
 6   28      0    Maintained
 7   31      1    Maintained
 8   34      1    Maintained
 9   45      0    Maintained
10   48      1    Maintained
11  161      0    Maintained
12    5      1 Nonmaintained
13    5      1 Nonmaintained
14    8      1 Nonmaintained
15    8      1 Nonmaintained
16   12      1 Nonmaintained
17   16      0 Nonmaintained
18   23      1 Nonmaintained
19   27      1 Nonmaintained
20   30      1 Nonmaintained
21   33      1 Nonmaintained
22   43      1 Nonmaintained
23   45      1 Nonmaintained
```

The discussion above shows that it is easy to calculate the Kaplan-Meier estimate of $\hat{S}(t)$ by hand. The SPLUS function `survfit` will do this for us also. Note that separate curves are fitted for each of the two groups under study.

```
402 > leukemia.km _ survfit(Surv(time,status) ~ group, data=leukemia,
+ conf.type="plain")
```

The right hand side of the formula is interpreted differently than it would be for an ordinary linear model or a `glm`. The survfit function uses the + operator to specify an interaction.

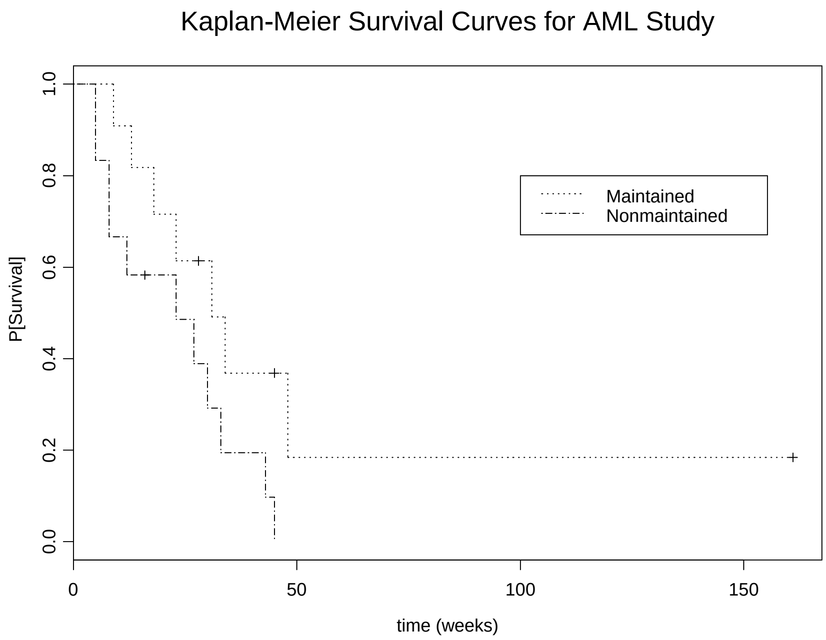
```
402 > summary(leukemia.km)
Call: survfit(formula = Surv(time, status) ~ group, data = leukemia, conf.type = "pl
ain")
                group=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867       0.7392        1.000
   13     10       1    0.818  0.1163       0.5903        1.000
   18      8       1    0.716  0.1397       0.4422        0.990
   23      7       1    0.614  0.1526       0.3145        0.913
   31      5       1    0.491  0.1642       0.1691        0.813
   34      4       1    0.368  0.1627       0.0494        0.687
   48      2       1    0.184  0.1535       0.0000        0.485
```

```
                 group=Nonmaintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     12       2   0.8333  0.1076       0.6225        1.000
    8     10       2   0.6667  0.1361       0.3999        0.933
   12      8       1   0.5833  0.1423       0.3044        0.862
   23      6       1   0.4861  0.1481       0.1958        0.776
   27      5       1   0.3889  0.1470       0.1008        0.677
   30      4       1   0.2917  0.1387       0.0198        0.564
   33      3       1   0.1944  0.1219       0.0000        0.433
   43      2       1   0.0972  0.0919       0.0000        0.277
   45      1       1   0.0000      NA           NA           NA
```

The standard errors are calculated using the Greenwood formula, and the confidence intervals are calculated using a simple normal approximation. Narrower confidence intervals can be obtained by estimating $\lambda(t)$ and its standard error directly, but this takes us beyond the scope of a brief introduction.

We can also get a plot of the survival function and its confidence envelope,

```
402 > plot(leukemia.km,lty=2:3)
402 > legend(100, .8, c("Maintained", "Nonmaintained"), lty =2:3)
402 > title("Kaplan-Meier Survival Curves for AML Study",xlab="time (weeks)",
+ ylab="P[Survival]")
```



Note that the censored observations are marked on the graph with + signs. Did maintenance help? How different are these two curves?