The Kaplan-Meier Estimator

Patrick Breheny

September 10

Introduction

- The likelihood construction techniques that we introduced last week can be used to estimate the survival/hazard/distribution functions for any parametric model of survival time
- As I have alluded to in the past, however, the distributions of survival times are often difficult to parameterize
- Our goal for today is to develop *nonparametric* estimates for these distributions, with a particular emphasis on the survival function $S(t)$

Empirical survival function

- In the absence of censoring, estimating $S(t)$ would be straightforward
- We could simply use the empirical survival function

$$\hat{S}(t) = \frac{\#\{i : t_i > t\}}{n}$$

- With censored observations, however, we don't always know whether $\widetilde{T}_i > t$ or not

Nonparametric likelihood

- As we discussed last week, likelihood provides a natural way to proceed with inference in the presence of censoring
- The likelihood of a survival function S given observed, right-censored data is

$$\begin{aligned} L(S|\text{Data}) &= \prod_{i=1}^{n} \mathbb{P}(T_i = t_i)^{d_i} \mathbb{P}(T_i > t_i)^{1-d_i} \\ &= \prod_{i=1}^{n} \{S(t_i^-) - S(t_i)\}^{d_i} S(t_i)^{1-d_i} \end{aligned}$$

- This expression is a bit different from the likelihoods we saw last week

Nonparametric likelihood (cont'd)

- In particular, it is not the likelihood of a parameter, but of a generic survival function S
- The set of possible values we must consider is not just an interval of parameter values, but rather the entire set of all possible survival functions
- This is the basic idea of *nonparametric* statistics: rather than specify a parametric form for $S(\cdot|\theta)$ and carry out inference concerning θ, we adopt procedures that deal directly with S itself

Estimating S

- For today, we will focus on the question of estimating S
- A natural estimate is to choose the value of S that maximizes $L(S)$; this is the nonparametric maximum likelihood estimator
- In other words, we must determine, out of the set of all possible survival functions, which function maximizes $L(S)$

Estimating S: First steps

- This might sound daunting, but it turns out to be easier than you would think
- Let's begin by making two observations that greatly restrict the possible values of S that we must consider
  - In order to maximize the likelihood, S must put positive point mass at any time t at which a subject was observed to fail (otherwise $S(t_i^-) - S(t_i)$ would be zero)
  - In order to maximize the likelihood, S cannot put any probability at times other than those at which subjects were observed to fail (redistributing that probability to next failure time would always increase the likelihood)
- Thus, we really only need to determine how much point mass to put at each observed failure time

Rewriting in terms of observed failure times

- Since the observed failure times are so critical here, let's rewrite the problem in terms of the observed failure times $0 = t_0 < t_1 < t_2 < \cdots < t_J < t_{J+1} = \infty$, and let

$$\begin{aligned} d_j &\equiv \# \text{ of failures at time } t_j \\ n_j &\equiv \# \text{ at risk at time } t_j^- \\ c_j &\equiv \# \text{ censored during the interval } [t_j, t_{j+1}) \end{aligned}$$

- In terms of this new notation, we can rewrite the earlier likelihood as

$$L(S) = \prod_{j=1}^{J} \{S(t_j^-) - S(t_j)\}^{d_j} S(t_j)^{c_j}$$

Solving for $\hat{\boldsymbol{\lambda}}$

- Next, let's rewrite the likelihood in terms of the hazard components, $\hat{\lambda}_1, \ldots, \hat{\lambda}_J$
- Doing so yields

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \prod_j \left\{ \lambda_j^{d_j} \prod_{k=1}^{j-1} (1-\lambda_k)^{d_j} \prod_{k=1}^{j} (1-\lambda_k)^{c_j} \right\} \\ &= \prod_j \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j} \end{aligned}$$

Solving for $\hat{\boldsymbol{\lambda}}$ (cont'd)

- Thus, the joint likelihood for $\boldsymbol{\lambda}$ consists of j separate components in which λ_j appears only in the jth component
- Furthermore, each component is equivalent to a binomial likelihood, so

$$\hat{\lambda}_j = d_j / n_j$$

and

$$\begin{aligned}\hat{S}(t) &= \prod_{t_j \leq t} (1 - \hat{\lambda}_j) \\ &= \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}\end{aligned}$$

Kaplan-Meier estimator

- The estimator on the previous slide was originally proposed by Kaplan and Meier in 1958, and is known as the Kaplan-Meier estimator (or product limit estimator, which is the name Kaplan and Meier proposed)
- This approach has come to be – by far – the most common way of estimating and summarizing survival curves
- The approach is so widespread, in fact, that Kaplan & Meier's original paper is the most highly cited paper in the history of statistics, and the 11th most highly cited paper in all of science

GVHD study

- To illustrate how Kaplan-Meier estimation works, let's apply it to a study involving graft-versus-host disease (GVHD) in bone marrow transplant recipients
- The patients in the study have a condition called severe aplastic anemia, in which the bone marrow produces an insufficient number of new blood cells
- These patients were given a bone marrow transplant from a compatible family member
- A serious complication of bone marrow transplantation is GVHD, in which the immune cells produced by the new bone marrow recognize the recipient as a foreign body and mount an attack

GVHD study (cont'd)

- To ward off GVHD, the recipients were randomized to receive one of two drug combinations:
  - Methotrexate (MTX)
  - Methotrexate and cyclosporine (MTX + CSP)
- The goal of the study is to determine whether treatment affected the occurrence of GVHD and if so, which treatment is superior

Data (by subject)

- Like elsewhere in statistics, survival data is typically organized with each individual subject occupying a row and the outcome and various covariates occupying the columns of the data set
- One difference, however, is that in survival analysis, two columns are required to denote the outcome (t_i and d_i):

Group	Time	Status
MTX+CSP	3	No
MTX+CSP	8	Yes
MTX+CSP	10	Yes
MTX+CSP	12	No
MTX+CSP	16	Yes
	...	

Data (by time)

As we saw in the derivation of the KM estimator, however, for the purposes of analysis it is often helpful to re-express the data in terms of the observed failure times:

Therapy	Time	GVHD
MTX+CSP	3	No
MTX+CSP	8	Yes
MTX+CSP	10	Yes
MTX+CSP	12	No
MTX+CSP	16	Yes
	...	

t	$n(t)$	$d(t)$
0	32	0
3	32	0
4	31	0
8	31	1
9	30	0
10	30	1
16	28	1
	...	

MTX alone group

In the MTX alone group,

Therapy	Time	GVHD
MTX	9	Yes
MTX	11	Yes
MTX	12	Yes
MTX	20	Yes
MTX	20	Yes
MTX	22	Yes
MTX	25	Yes
MTX	25	Yes
MTX	25	No
MTX	28	Yes
MTX	28	Yes
	...	

t	$n(t)$	$d(t)$
0	32	0
9	32	1
11	31	1
12	30	1
20	29	2
22	27	1
25	26	2
28	23	2
	...	

$\hat{S}(t)$: MTX + CSP

t	$n(t)$	$d(t)$
0	32	0
8	31	1
10	30	1
16	28	1
	...	

t	$\frac{n(t)-d(t)}{n(t)}$	$\hat{S}(t)$
0	1	1
8	$30/31$	.968
10	$29/30$	.935
16	$27/28$	.902
	...	

$\hat{S}(t)$: MTX alone

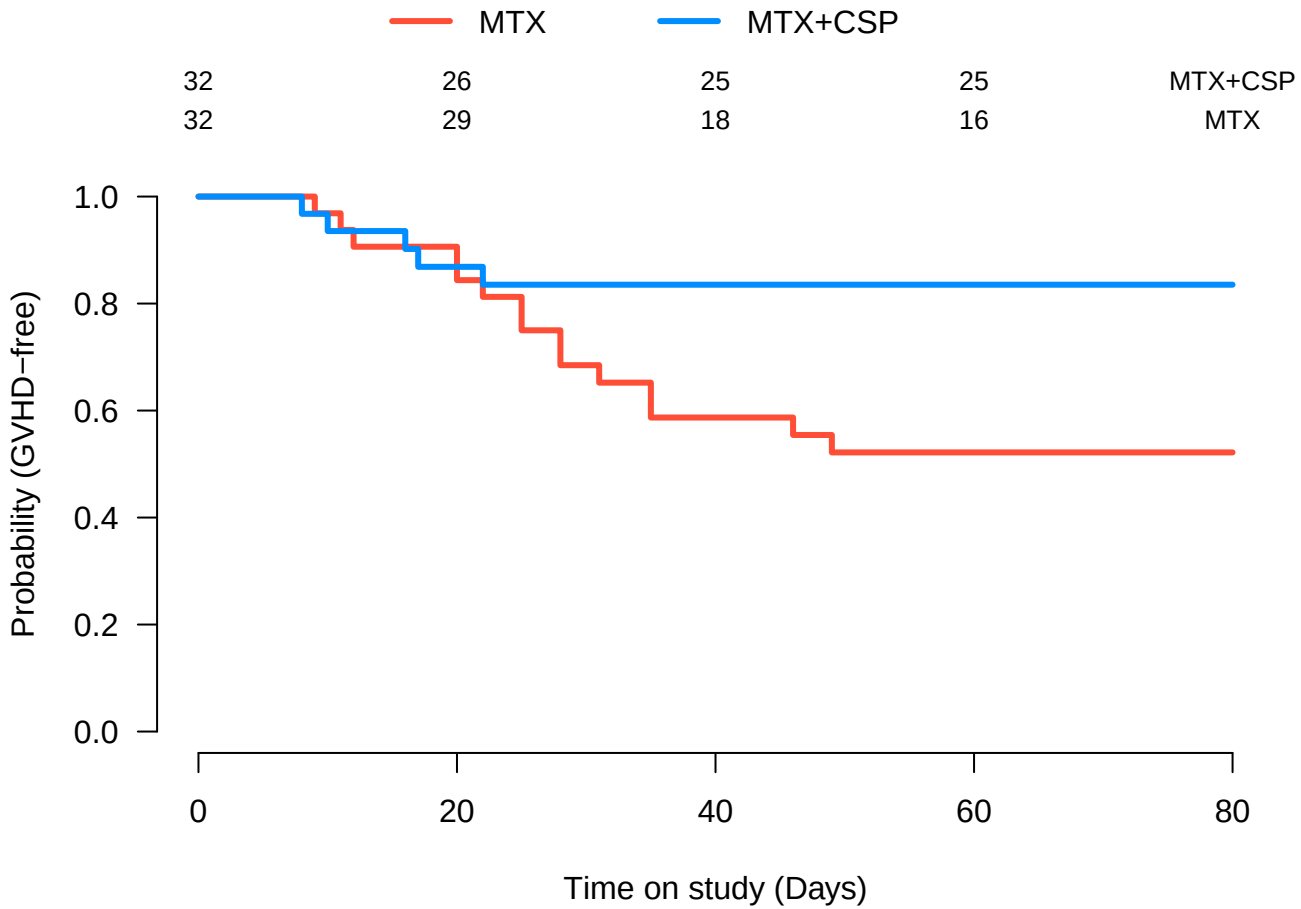
In the MTX group,

t	$n(t)$	$d(t)$
0	32	0
9	32	1
11	31	1
12	30	1
20	29	2
22	27	1
25	26	2
28	23	2
	...	

t	$\frac{n(t)-d(t)}{n(t)}$	$\hat{S}(t)$
0	1	1
9	31/32	.969
11	30/31	.938
12	29/30	.906
20	27/29	.844
22	26/27	.812
25	24/26	.750
28	21/23	.685
	...	

Kaplan-Meier curve: GVHD

The result of all these calculations is usually summarized in a plot called a *Kaplan-Meier curve*:



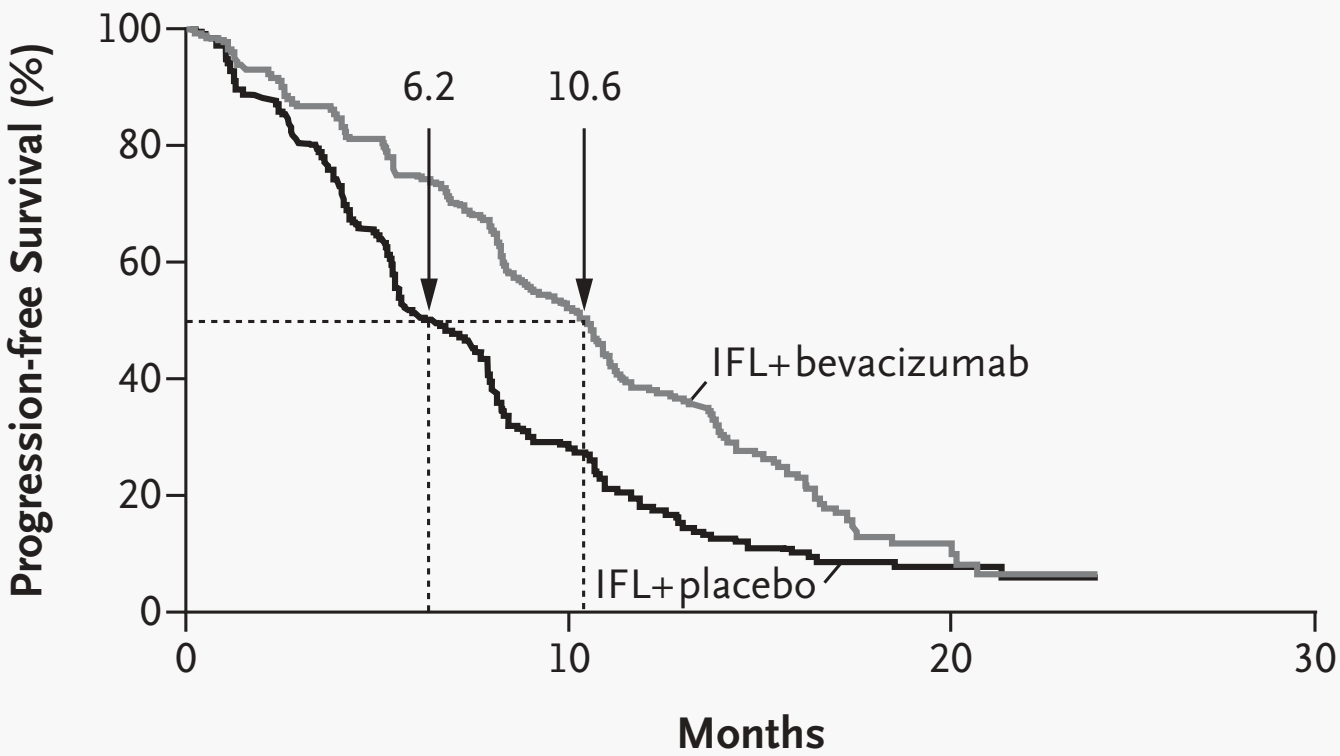
Summary statistics

- Summary statistics for time-to-event data are typically derived from the Kaplan-Meier estimates
- For example, in this study we might report estimates of the probability of remaining GVHD-free at 60 days of 84% in the MTX+CSP group and 52% in the MTX alone group
- This can be obtained simply by reading the Kaplan-Meier curve "vertically"
- One can also read the Kaplan-Meier curve "horizontally" to obtain estimates of quantiles

Median survival times

- One quantile of particular interest is the median; i.e., the time at which the survival function drops below 0.5
- In the case where death is the outcome, this is known as the *median survival time* and is almost always reported (if it can be estimated)
- For our GVHD example, the median time to event cannot be estimated since $\hat{S}(t)$ never reaches 0.5; to see what the idea, though, let's briefly turn to data from a clinical trial of a cancer drug called Avastin

Kaplan-Meier curve: Avastin study



No. at Risk

IFL+bevacizumab	402	269	143	36	6	0
IFL+placebo	411	225	73	17	8	0

The survival package

- Finally, let's discuss the R functions for constructing the Kaplan-Meier estimate and plotting KM curves
- In this course, we will make extensive use of the survival package in R
- The package is bundled by default with R, meaning that you do not have to install it, although you will have to load it with

```
library(survival)
```

before you can use it

Surv objects

The survival package has a construct called a Surv object to handle survival outcomes, which are one entity but with two components (t_i and d_i):

```
> S <- with(Data, Surv(Time, Status))
> class(S)
[1] "Surv"
> head(S)
[1]  3+  8  10  12+ 16  17
> head(S[,1])
[1]  3  8 10 12 16 17
> head(S[,2])
[1] 0 1 1 0 1 1
```

survfit

- The function in survival for constructing Kaplan-Meier estimates is called survfit:

```
fit <- survfit(S ~ Data$Group)
```

  where S is a Surv object

- S does not have to be constructed ahead of time; this also works (and is probably better coding practice):

```
fit <- survfit(Surv(Time, Status) ~ Group, Data)
```

Summarizing the survfit object

By printing the object, we get a rough summary of each group, although the summary revolves around the median, which in our case cannot be estimated:

```
> fit
                n events median 0.95LCL 0.95UCL
Group=CSP      32     15     NA      35      NA
Group=CSP+MTX  32      5     NA      NA      NA
```

Provided they can be estimated, we would see the median survival time in each group, along with upper and lower 95% confidence interval bounds (we'll discuss how those are calculated in the next lecture)

Summarizing the survfit object (cont'd)

To find out more about the KM estimates at specific times, we can use the summary function:

```
> summary(fit, time=40)
            Group=CSP
  time   n.risk   n.event   survival   std.err   95%LCL   95%UCL
40.000   18.000   13.000      0.587     0.088   0.437    0.788

            Group=CSP+MTX
  time   n.risk   n.event   survival   std.err   95%LCL   95%UCL
40.000   25.000   5.0000     0.8353    0.0674   0.7131   0.9784
```

plot.survfit

- Once the Kaplan-Meier curve has been estimated, it can be plotted in a straightforward manner:

```
plot(fit)
```

- Some useful options to be aware of are
  - `mark.time`: Marks the times at which observations were censored (default: `TRUE`)
  - `xmax`: Maximum time at which to plot $\hat{S}(t)$
  - `xscale`: Set this to 365.25 to get curves displayed in years instead of days, and so on

Some limitations of the survival package

- The survival package is very mature, well-maintained, and well-developed software with expansive support for almost all of the models we will discuss in this course
- However, it is also old (originally written in S then ported to R) and has a few shortcomings that make it sometimes inconvenient to work with in practice:
  - Its plots are rather . . . basic
  - The internal data structure of a survfit object is inconvenient to work with

Option #1: Wrapper functions

- One option is just to write a wrapper function that calls / works with the functions in `survival` to improve the appearance of plots, add options, etc.
- For example, I'm providing (in `fun.R`) the `Plot()` and `nrisk()` functions:

```
Plot(fit)
nrisk(fit)
```

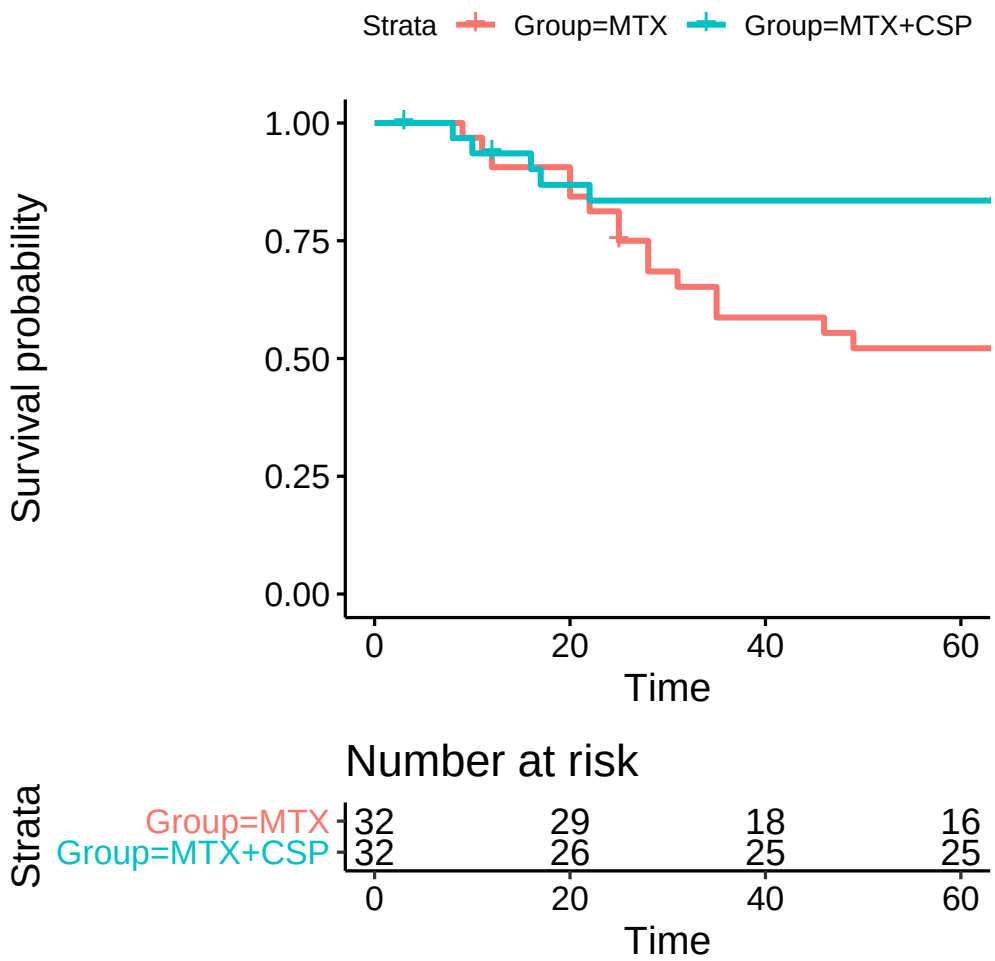
  which I used to make the plot on slide 19; note that you may need to adjust the margins to avoid the number at risk conflicting with the legend
- I've also provided `median(fit)` to return the median survival times

Option #2: Other packages

- Another option is to use a different package for plotting
- Several exist, but I'll discuss one called `survminer`, which is focused on constructing survival plots using `ggplot2`
  - If you've used `ggplot2` before, you'll likely find its usage straightforward
  - If not, you may prefer option #1
- The package has lots and lots of options; I'll only cover its most basic usage – for more information, check out https://rpkgs.datanovia.com/survminer/index.html

ggsurvplot()

```
ggsurvplot(fit, Data, risk.table=TRUE)
```



surv_summary()

Other useful functions survminer provides are surv_median() and surv_summary(), which organizes the survfit object into a data.frame, as opposed to a list of objects that need to be cross-referenced with each other:

```
surv_summary(fit, Data)
```

time	n.risk	n.event	n.censor	surv	std.err	upper	lower	strata	Group
9	32	1	0	0.969	0.032	1.000	0.910	Group=MTX	MTX
11	31	1	0	0.938	0.046	1.000	0.857	Group=MTX	MTX
12	30	1	0	0.906	0.057	1.000	0.811	Group=MTX	MTX
20	29	2	0	0.844	0.076	0.979	0.727	Group=MTX	MTX
22	27	1	0	0.812	0.085	0.960	0.688	Group=MTX	MTX
...									

Next time, we'll discuss confidence bands for Kaplan-Meier curves