Survival, Hazards, & Censoring

Matthew Austin

Institute for Advanced Analytics
MSA Class of 2019

Introduction to survival analysis:
What's survival analysis?

Overview

- In survival analysis, we are interested in the **time until an event occurs**, or *failure time*
  - Time until transplant patients need a new kidney
  - Time until someone develops a disease
- Originally used in medical studies (hence the name "survival"), but applicable to any "time-to-event" data
  - How long until a piece of equipment/machinery fails?
  - How long until the ~~Cubs~~ Orioles win the World Series?
    - (spoilers: never)
- Survival analysis is the branch of statistics that deals with these kinds of questions

"Time-to-event" data?

- In survival analysis, "time" generally refers to *tenure* or *follow-up/survival time* rather than actual calendar time
- The "event" is some specific outcome of interest
  - Person develops disease
  - Team wins championship
  - Car battery dies

So... basically like logistic regression?

- Similar to logistic regression, we are looking at whether or not an event occurs
  - Logistic regression: "Did it happen?"
- But unlike logistic regression, we are also interested in the *time it takes* for the event to occur
  - Survival analysis: "How long did it take to happen?"

Okay... then like regular linear regression?

- The biggest problem with using OLS for time-to-event data is **censoring**—for some observations, the event may never occur!
- Other problems with using OLS:
  - Even without censoring, "time" is always positive and would rarely satisfy OLS assumptions (not normally distributed, etc.)
  - Risk of failure may change over time
- This is all difficult (but not impossible) to characterize using distributions as you've been accustomed to doing

Why wouldn't an event happen?

Because the Orioles are a really bad baseball team and will never win the World Series again.

Introduction to survival analysis:
Data structure

So is the response continuous or categorical?

- Neither!
- Or maybe both!

Structure of survival data

In survival analysis, the response variable is actually two parts:

1. `time`: The tenure for an observation
2. `event/status`: At the end of that time, what happened?

team	time	event
MIL	3	0
LAD	4	0
ATL	4	1
COL	4	1
CHC	1	1
BOS	4	0
HOU	3	0
CLE	3	1
NYY	5	1
OAK	1	1

Maryland recidivism study

- The `recid` dataset is a recidivism study from the 1970s following 432 men for one year after their release from Maryland state prisons
- Of the $n = 432$ men, 114 (26%) were re-arrested within one year of their release
- We will model the time to recidivism as a function of various predictors

Dataset

- Response: how long until a re-arrest occurs?
  - `week` is the week of arrest; `week = 52` if censored (not arrested)
  - `arrest` indicates whether or not an arrest occurred
- Predictors:
  - `fin`: received financial aid upon release
  - `age`: age (in years) at time of release
  - `race`: indicator for Black
  - `wexp`: indicator of full-time work experience prior to incarceration
  - `mar`: married at time of release
  - `paro`: released on parole
  - `prio`: number of prior convictions

The survival & hazard functions:
The survival function

Summarizing survival data

- We are interested in the event time T
- The nature of survival data presents unique challenges in summarizing information about T
  - Are means/variances useful for skewed distributions such as time?
  - In the presence of censoring, can we even estimate means and variances without actually knowing that T is?
- Pretty much everything in survival analysis is described in terms of two major quantities:
  - The **survival function**
  - The **hazard function**

The survival function

- The **survival function** is the probability of surviving **beyond** time t

$$S(t) = \Pr(T > t)$$

- Properties of the survival function:
  - $S(t)$ is in the interval $[0, 1]$
  - Always starts at 1: $S(0) = 1$
  - $S(t)$ never increases as t increases

Kaplan-Meier estimation

- Notation: since we know `time` and `event`, then we also know how many events occur at each time
  - d_t is the number of events occurring at `time = t`
  - r_t is the number of observations available (the **risk set**) right before `time = t`
- The **Kaplan-Meier** estimate of the survival function is

$$\hat{S}(t) = \prod_{t_m \leq t} \left(1 - \frac{d_m}{r_m}\right)$$

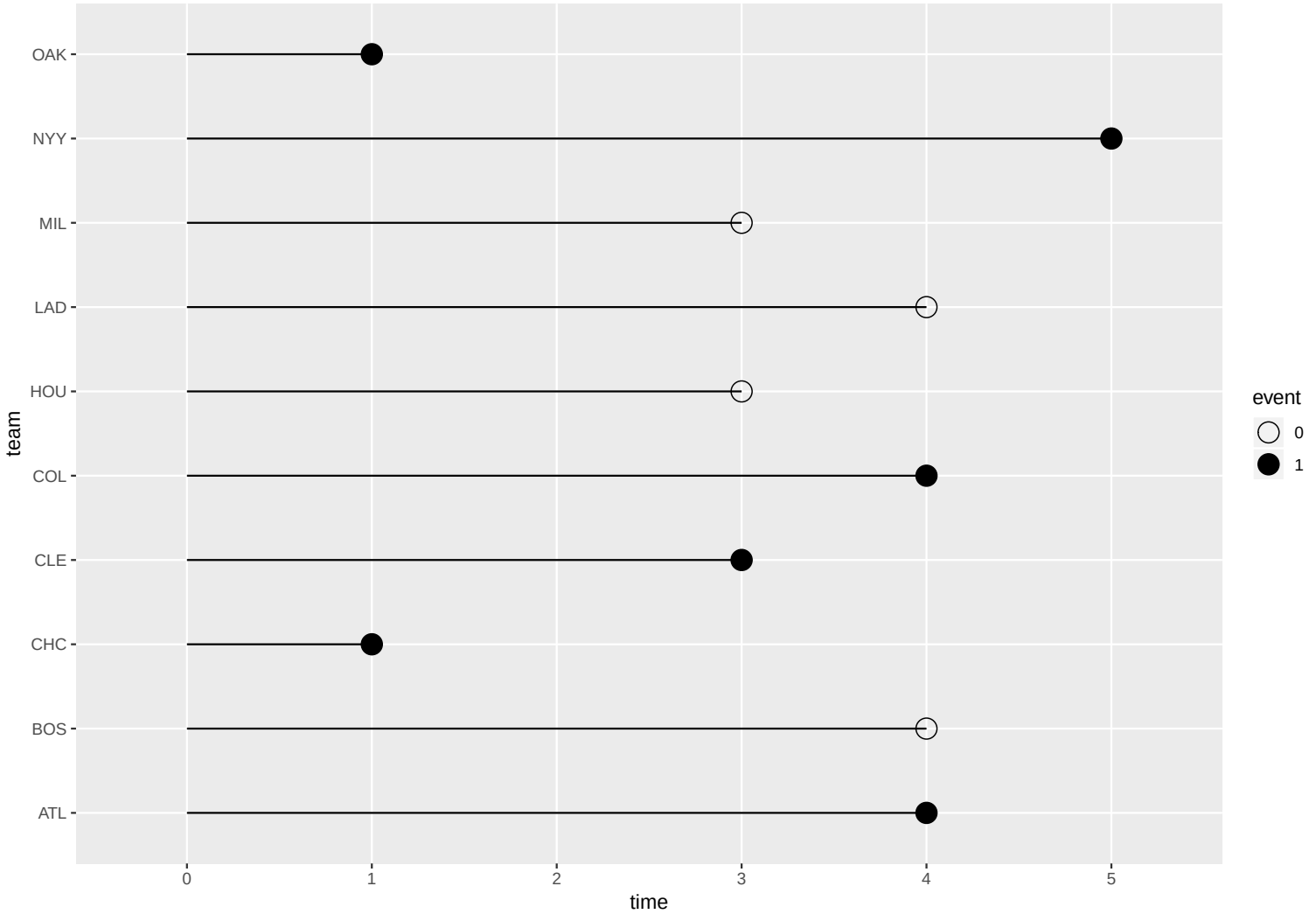
Kaplan-Meier estimation (con't)

- At the beginning ($t = 0$), no events have occurred ($d_0 = 0$), and all observations are at risk ($r_0 = n$), so

$$S(0) = \left(1 - \frac{0}{n}\right) = 1$$

- Basically, we start with $\hat{S}(0) = 1$ and step forward in time, reducing $\hat{S}(t)$ by a factor of $\frac{d_t}{r_t}$ at each time
  - $\hat{S}(0) = 1$
  - $\hat{S}(1) = \hat{S}(0) \times \left(1 - \frac{d_1}{r_1}\right)$
  - $\hat{S}(2) = \hat{S}(1) \times \left(1 - \frac{d_2}{r_2}\right)$
  - $\hat{S}(t) = \hat{S}(t-1) \times \left(1 - \frac{d_t}{r_t}\right)$

Kaplan-Meier estimation (con't)

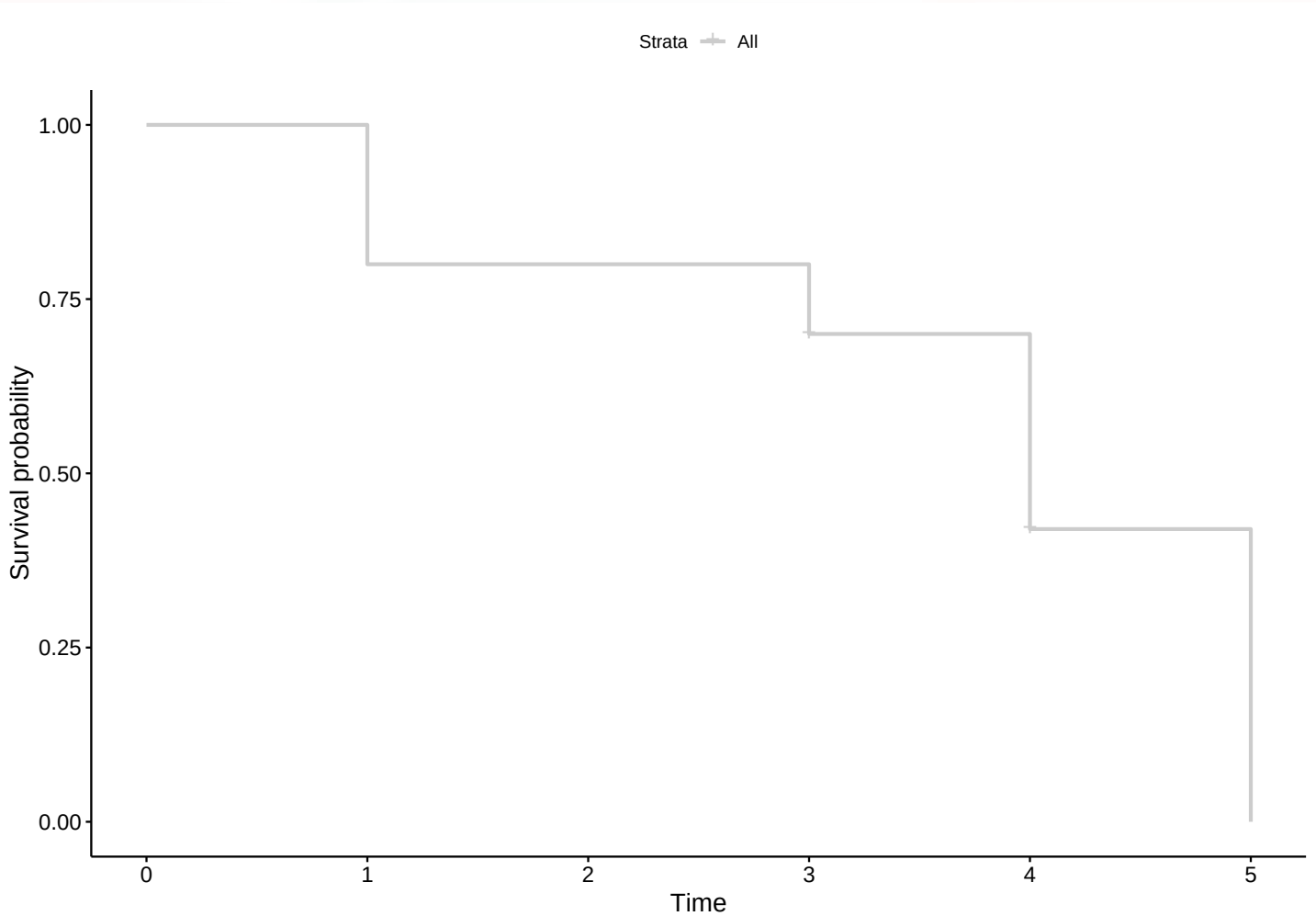


Kaplan-Meier estimation (con't)

- $\hat{S}(0) = 1$
- $\hat{S}(1) = \hat{S}(0) \times \frac{8}{10} = 0.8$
- $\hat{S}(2) = \hat{S}(1) \times \frac{8}{8} = 0.8$
- $\hat{S}(3) = \hat{S}(2) \times \frac{7}{8} = 0.7$
- $\hat{S}(4) = \hat{S}(3) \times \frac{3}{5} = 0.42$
- $\hat{S}(5) = \hat{S}(4) \times \frac{0}{1} = 0$

Survival curves

The **survival curve** is a plot of $\hat{S}(t)$ vs. `time`



Summary statistics

- Due to censoring, the mean is impossible to truthfully estimate
- But the **median** is not; we only need the event to occur for at least half of the sample
- The median (or half-life) is whatever time t at which $\hat{S}(t)$ drops below 0.5
- Interpretation: 50% of observations survive beyond time t

Kaplan-Meier estimation: R syntax

- In R, we'll be using the `survival` package throughout the course
- The function `Surv()` creates an object of survival outcomes, where we need to specify the `time` and `event`

```
library(survival)
Surv(time = lcs$time, event = lcs$event == 1)
```

```
##  [1] 3+ 4+ 4  4  1  4+ 3+ 3  5  1
```

Kaplan-Meier estimation: R syntax (con't)

```
lcs_fit <- survfit(Surv(...) ~ 1, data = lcs)
summary(lcs_fit)
```

time	n.risk	n.event	survival	std.err	lower	upper
1	10	2	0.80	0.158	0.587	1.000
3	8	1	0.70	0.207	0.467	1.000
4	5	2	0.42	0.420	0.184	0.956
5	1	1	0.00	Inf	NA	NA

Kaplan-Meier estimation: R syntax (con't)

```
recid_fit <- survfit(Surv(week, arrest) ~ 1,
                     data = recid)
summary(recid_fit)
```

time	n.risk	n.event	survival	std.err	lower	upper
1	432	1	0.998	0.002	0.993	1
2	431	1	0.995	0.003	0.989	1
3	430	1	0.993	0.004	0.985	1
4	429	1	0.991	0.005	0.982	1
5	428	1	0.988	0.005	0.978	0.999
6	427	1	0.986	0.006	0.975	0.997
...	...	...	...	...	...	...
46	337	4	0.771	0.026	0.732	0.812
47	333	1	0.769	0.026	0.73	0.809
48	332	2	0.764	0.027	0.725	0.805
49	330	5	0.752	0.028	0.713	0.794
50	325	3	0.745	0.028	0.705	0.788
52	322	4	0.736	0.029	0.696	0.779

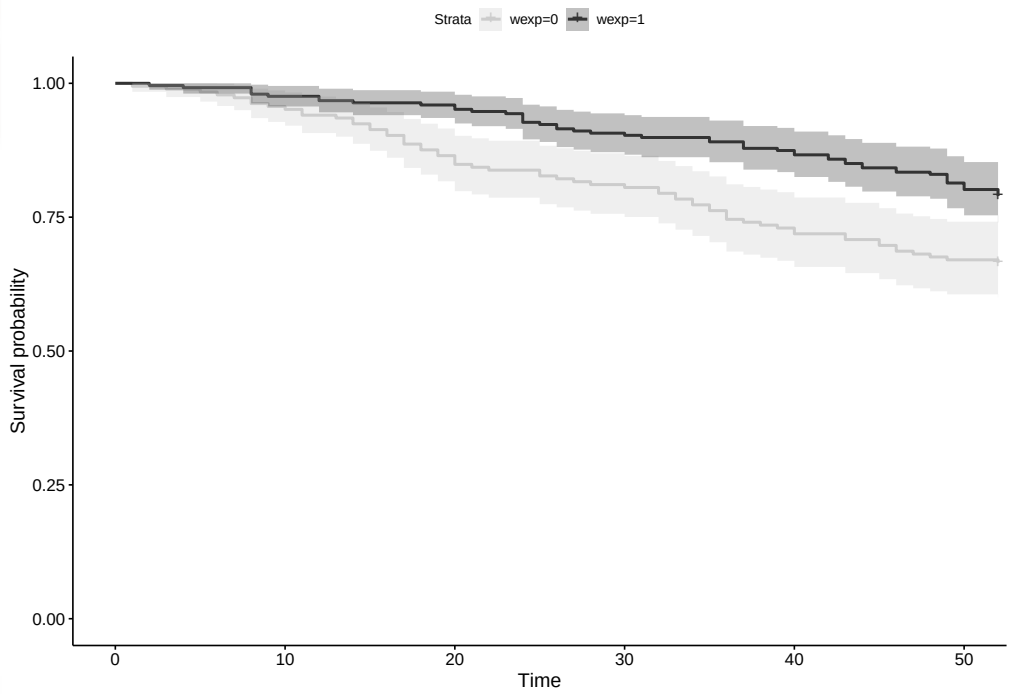
Kaplan-Meier estimation: SAS syntax

```
proc lifetest data=survival.recid plots=s(cl, cb=ep)
  time week*arrest(0);
run;
```

The survival & hazard functions:
Log-rank test

Stratified survival curves

- We can also create separate/stratified curves by group



- And of course, you'll probably want a way to test the differences between curves...

Log-rank tests

- The **log-rank test** combines all of the information from the KM estimate at times where events occur
  - **Null hypothesis**: all curves are equal
- Same idea as CMH tests for association from categorical data, where we have a partial table at each distinct time:

At `time = t`:	#events	#non-events	total
Group 1	d_{1t}	$r_{1t} - d_{1t}$	r_{1t}
Group 2	d_{2t}	$r_{2t} - d_{2t}$	r_{2t}
total	d_t	$r_t - d_t$	r_t

- Weighted log-rank tests can adjust for r_t or $\hat{S}(t)$, both placing a larger emphasis on earlier event times

Log-rank test: R syntax

```
survdiff(Surv(time = week, event = arrest) ~ wexp,
         rho = 0, data = recid)
```

```
## Call:
## survdiff(formula = Surv(time = week, event = arrest) ~ wexp,
##     data = recid, rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## wexp=0 185       62     45.6      5.91      9.91
## wexp=1 247       52     68.4      3.94      9.91
##
##  Chisq= 9.9  on 1 degrees of freedom, p= 0.002
```

Log-rank test: SAS syntax

```
proc lifetest data=survival.recid plots=s(cl, cb=ep)
  time week*arrest(0);
  strata wexp;
run;
```

The survival & hazard functions:
The hazard function

Hazard functions

- In survival analysis, we also use the **hazard function**

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{\Pr\left(t < T < t + \Delta t | T > t\right)}{\Delta t} \\ &= -\frac{d}{dt} \log S(t)\end{aligned}$$

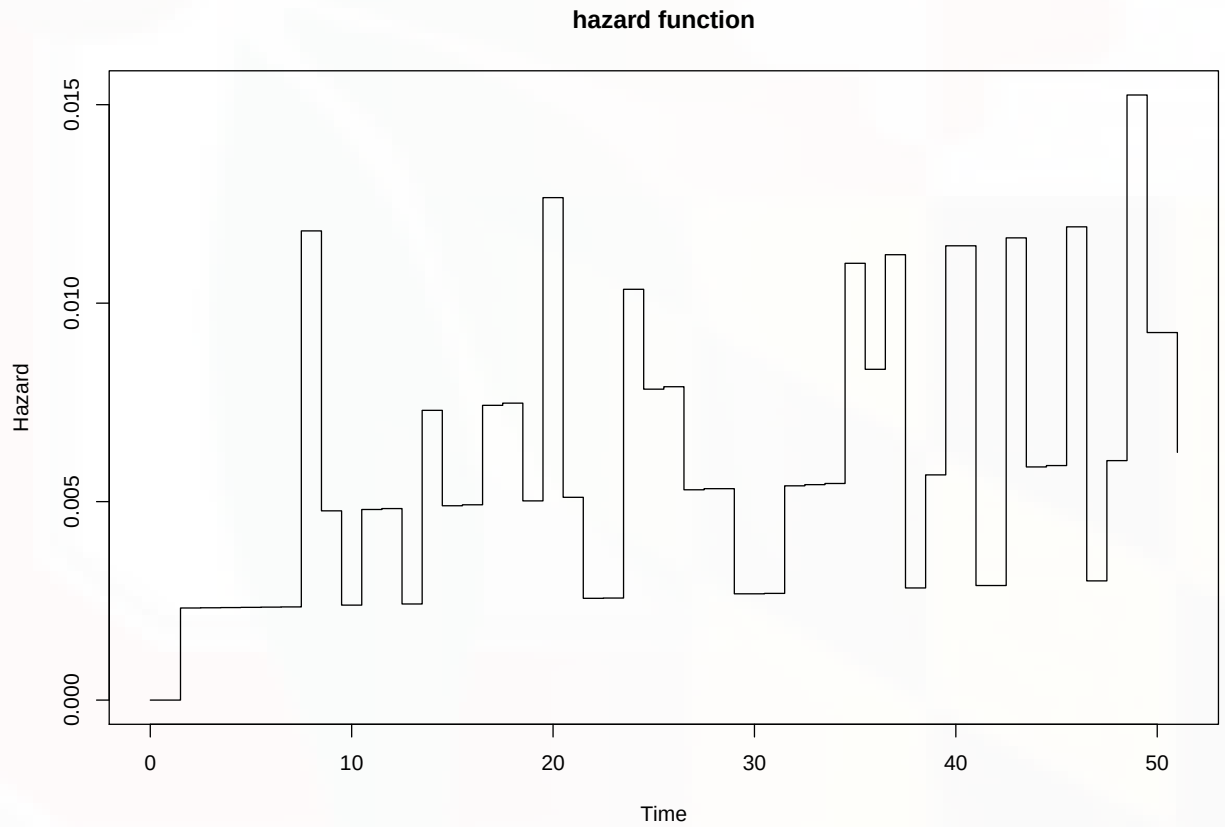
- (Gross)
- Key things to note:
  - Numerator: The hazard is **conditional**
  - Denominator: The hazard is a **rate**, not a probability

Meaning of the hazard function

- The hazard is the **instantaneous event rate** for the risk set at time t
  - Essentially, given survival up to a certain time t, it is the rate of events in the interval $[t, t + \Delta)$
- Think of it like a measure of risk: a smaller/larger hazard between times t and $t + \Delta$ indicates a lower/higher risk of failure in this interval

Hazard functions: R syntax

```
recid_haz <- muhaz::kphaz.fit(recid$week2, recid$arrest)
muhaz::kphaz.plot(recid_haz)
```



Hazard functions: SAS syntax

```
proc lifetest data=recid2 method=life plots=h width=1;
  time week*arrest(0);
run;
```

Cumulative hazard

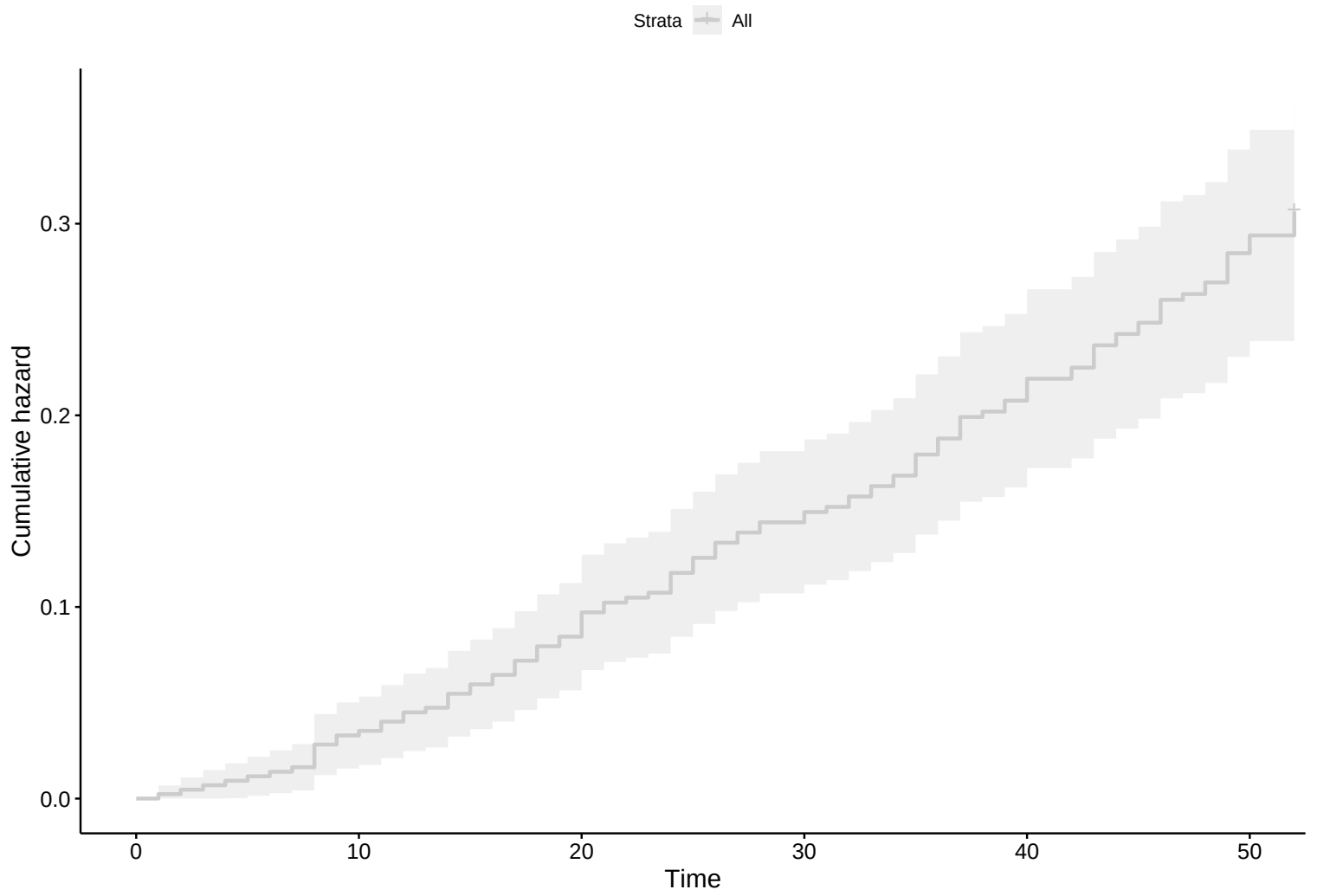
- The **cumulative hazard** is just the total hazard up until time t

$$\Lambda(t) = \int_0^s \lambda(s)ds$$

- Like the hazard, this is also a measure of risk, not a probability: a smaller/larger cumulative hazard at `time = t` indicates a lower/higher risk of failure by `time = t`
- The Nelson-Aalen estimate of the cumulative hazard is

$$\hat{\Lambda}(t) = \sum_{t_m \leq t} \hat{\lambda}(t_m) = \sum_{t_m \leq t} \frac{d_m}{r_m}$$

Cumulative hazard plot



Relationship between survival and hazards

- The survival, hazard, and cumulative hazard functions are all directly related
  - $\Lambda(t) = -\log S(t)$
  - $S(t) = e^{-\Lambda(t)}$
  - $\lambda(t) = -\frac{d}{dt} \log S(t) = \frac{f(t)}{S(t)}$
    - ($f(t)$ is called the *density* function)
- These three quantities are all different ways of describing the same distribution; if you know one of them, you can compute the others

Time & censoring:
Origin of time

The meaning of time

- Survival analysis has a few things that set it apart from any kind of statistical modeling you've seen before
  - Time to an event
  - Censoring
- Like with any kind of analysis, how we approach survival data depends on a whole bunch of assumptions that we make—both explicitly and implicitly
- As such, it's worth considering what we mean by "time" and by the concept of "censored" data

When does time start?

- In survival analysis, we often create an artificial world in which everyone "starts" at the same time
- However, the choice of that starting point isn't always obvious
  - Birth/age
  - Some fixed point in time
  - Time since exposure to disease vs. developing disease
  - Time since diagnosis vs. surgery vs. treatment
  - Time since some other event
- When multiple time information is available, whatever isn't chosen as the origin is usually included in models as a predictor
  - Time since production, purchase, last repair, etc. are all reasonable choices of origin for "time until a car dies"

Time & censoring:
Censored responses

Observed time & status

- Recall that we are interested in T, the time to an event, but observations can be censored
- The "`time`" we actually observe for each observation is $\min\{T_i, C_i\}$
  - T_i is the time until the event
  - C_i is the censoring time
- And the `status` is an indicator variable taking the value 1 if the event occurred ($T_i \leq C_i$) or 0 if the observation was censored ($T_i > C_i$)
- So our response in survival analysis is `(time, status)` together

Censored ≠ missing

- If an observation is censored, then we never observe the event and cannot know what T actually is
- But we *do* know that for some amount of time, the event had not occurred. Thus, the censoring time C does give us *some* information about T, just not all of it
- Censored data should be treated as **incomplete** rather than missing outright—ignoring censored observations would be acting as if we know absolutely nothing about T (which is false!)
- **Main idea**: every subject contributes to our estimates for some amount of time, and censoring tells us how long that contribution lasts for

Time & censoring:
Types of censoring

Type I & Type II censoring

- In **Type I censoring**, there is a specific end time c where any subject that hasn't had the event by `time = c` is censored
- In **Type II censoring**, the time goes until a certain [pre-specified] number of events have occurred, and any subjects who haven't had the event by that time are censored

Right-censoring

- If a subject is **right-censored**, then all we know is $T_i > c$
  - A person is arrested sometime after 52 weeks
  - A pump fails sometime after 48 hours
- The most common form of censored data you will encounter is right-censored; specifically, Type I right-censoring
- Right-censoring can occur for many reasons:
  - The subject actually does not experience the event before the study ends
  - The subject is lost due to moving away, death (if death is not the event of interest), etc.
  - The subject withdraws from the study for whatever reason

Fixed and random censoring

- Notice that some of the reasons at the end of the previous slide have some subjects being censored before the end of the study
- In **fixed censoring**, censoring only occurs at the end of the study; i.e., $C_i = c$ is known in advance
- In **random censoring**, C_i may vary between subjects for reasons beyond the investigator's control
- Recidivism example:
  - Fixed: A man not arrested after 52 weeks is censored by design, because that's when the study ended
  - Random: A man hasn't been arrested after 30 weeks, but we lose contact with them after that

Random censoring

- Regardless, we assume T and C are independent—basically, we act as if subjects censored at time t were randomly selected to be censored from all subjects still at risk at time t
- Under independence of T and C, it turns out there's actually no mathematical difference between fixed and random censoring: the likelihood function is the same
- (Random censoring also includes cases of *delayed/random entry*, where not all subjects start at the same time)
  - The recidivism study follows each man for one year after release, so the "origin" is the same for everyone—time following release, regardless of the actual calendar date of release
  - Instead, the investigators could've just done the study for one year, so the men would enter the study as they were released

Non-informative censoring

- We also assume that random censoring is **non-informative**, meaning that censoring is caused by something independent of the event
  - Subject is censored because the study ended
  - Subject is censored because they moved away
- Informative censoring would be something like withdrawal due to side effects of a treatment, worsening conditions, etc.
- These independence and non-informative assumptions are nice because they allow us to avoid having to model the censoring mechanism (i.e., the reason for censoring), which we generally don't care about

Left- & interval-censoring

- **Left-censoring** is when the event occurs before `time = t`, but we don't know exactly when, so we only observe $T < t$
  - A person initially doesn't have a disease, but tests positive at time t_1. If the time scale is the time to developing the disease (rather than time to diagnosis), we don't know exactly when they got the disease, only that it happened sometime before that first appointment: $T < t_1$
  - If we want to know the age at which college students begin smoking (so age is the time scale), any student who started smoking before college is left-censored
- **Interval censoring** is when the event occurs between two times, but we don't know exactly when
  - A person tests negative during a doctor's appointment at time t_1 and positive during the appointment at t_2. Time to developing disease is $t_1 < T < t_2$

Truncation

- With censoring, we might not know T for some subjects, but we know that it's either greater than or less than t
- With **truncation**, we don't even know that the subject exists **at all**
  - Survivor/selection bias
  - Car insurance/accidents
- Essentially, it's impossible to observe *any* subjects for which $T < t$