An Introduction to Event History Analysis

Oxford Spring School
June 18-20, 2007

Day One: Exploring Survival Data

Survival Analysis

Survival analysis is also known as "event history analysis" (sociology), "duration models" (political science, economics), "hazard models" / "hazard rate models" (biostatistics, epidemiology), and/or "failure-time models" (engineering, reliability analysis).

- Commonality: Models for *time-to-event data*.
- Roots in biostats/epidemiology...
- Also: Engineering, etc., plus sociology, economics.
- Widely used in all the social sciences now...

Examples of problems/applications in political science include:

- In American politics: Confirmation durations, political careers, position-taking, bill cosponsorship, campaign contributions, state constitutions, policy innovation/adoption, Congressional overrides of Supreme Court decisions, etc.
- In comparative politics: Cabinet/government durations, length of civil wars, coalition durability, etc.
- In international relations: War duration, peace duration, alliance longevity, length of trade agreements, etc.

Outside of political science, these models have been used to study (e.g.):

- Strike durations, work careers (including promotions, firings, etc.), and (in economics),
- Criminal careers, marriage and child-bearing behavior, and even induction into the Major League Baseball Hall of Fame (in sociology), and
- A range of health- and development-related issues (in anthropology).

More generally, survival models are useful for any issue in which:

- *The phenomenon of interest is a duration*, and/or
- *The response is the occurrence of a discrete event in time.*

General Issues with Modeling Survival Data

Characteristics of Time-To-Event Data

- *Discrete* events (i.e., not continuous),
- Take place over *time*,
- May not (even never) experience the event (i.e., possibility of *censoring*).

Thus, modeling duration data presents several rather sticky issues...

- Like count data, duration data are strictly *nonnegative*,
- Also, the data are *conditional*: to survive to some time t, one must necessarily have survived up to $t-1$ as well,
- Additionally, we regularly encounter observations which haven't failed yet (i.e., *censored* data).

All of these things present challenges for the data analyst.

An Example: Retirements from the U.S. House of Representatives

To consider an example, suppose we were interested in exploring the determinants of the time until a member retires from Congress. Possible influences on that decision might include:

- The member's age,
- The member's tenure,
- Partisan control of the chamber,
- etc.

How might we model these influences on the member's decision to leave office?

One option: OLS regression of length of tenure.

- This has some serious substantive/practical drawbacks...
    - It would allow us to estimate the effect of Congressperson-level variables that didn't change over time (like political party, region, etc.), but
    - It would not let us assess the impact of variables that changed over time (w/o aggregation bias).
    - It could also predict negative tenures. Moreover,
    - How would we handle sitting members?...

- This is also a bad idea statistically; in particular, it will yield biased estimates of the things we're interested in...
    - For example, if you include sitting members, you're estimating something other than what you're really interested in (a mixture of the actual "survival" time and the "censoring" mechanism – more on this later).
    - On the other hand, if you exclude them, you're systematically leaving out data (which are more likely to be on members with longer tenures).

Another option: OLS/Poisson regression of the count of members retiring in a given year. This option also has some limitations.

- While it will allow us to estimate the effect of time-varying factors,
- It will not let us estimate the effect of member-level influences.
- Also raises the problem of aggregation bias, if members are not homogenous in the reasons they leave.

A third option would be to estimate a logit model of the decision to retire (=1) or not (=0) on member/year data.

- This is definitely an improvement – it addresses nearly all the problems we raised above.
- But, what if there is time dependence?

The best option is some form of *survival model*.

Survival Data Basics

Let's start by defining some terms. Assume we have N units $i = \{1, 2, ...N\}$, each of which will experience some event in which we're interested. Call:

$$\begin{array}{rcl} Y_i & = & \text{the duration until the event occurs,} \\ Z_i & = & \text{the duration until the observation is "censored" (more on this later),} \\ T_i & = & \min\{Y_i, Z_i\}, \\ C_i & = & 0 \text{ if observation } i \text{ is censored, 1 if it is not.} \end{array}$$

A "normal" MLE model might model $\Pr(T_i = t)$, that is, the probability that the duration observed is equal to some particular value t. We might refer to this *density* as $f(t)$:

$$f(t) = \Pr(T_i = t) \tag{1}$$

that is, the instantaneous probability of the event occurring.

There are (at least) two potential problems with this approach:

1. The probability in (1) is *conditional* (that is, you can't have the event at t unless you haven't had the event at $1, 2, ...t-1$), and
2. Some observations may be *censored* (that is, we only know that they haven't had the event by t).

Consider the example of human mortality. The density $f(t)$ is U-shaped and skewed, but

- $\Pr(T_i = 80)$ is meaningless of the subject dies at age 78.
- So, we want to know the *cumulative probability* of death...

This is the cumulative distribution function (the CDF, often generically called $F(t)$):

$$\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t)\, dt \tag{2}$$

That is, $F(t)$ is the probability of death on or before t.

Since everyone (eventually) dies, we can get the probability of *survival to* t as

$$\Pr(T_i \geq t) \equiv S(t) = 1 - F(t). \tag{3}$$

Now, think about what we really want to know...

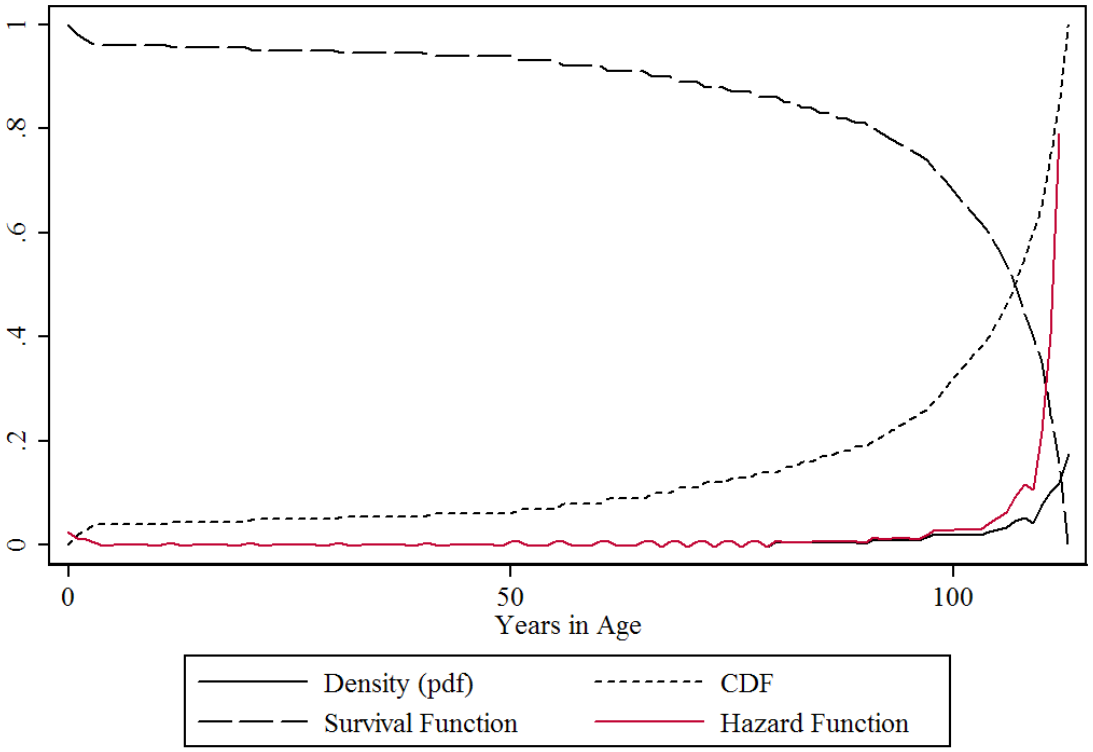
- We want to know the probability of having the event at t, given that we haven't had it prior to that point.
- That is, we want to know $\Pr(T_i = t | T_i \geq t)$.
- This is known as the *hazard*, and is typically denoted $h(t)$.

By the rule for conditional probability, the hazard is equal to

$$h(t) = \frac{f(t)}{S(t)}. \tag{4}$$

Think of the hazard as a kind of a probability. It is important that it accounts for the time-path of the data. Moreover, it is the reverse direction from the duration – higher hazards correspond to shorter durations, and vice-versa.

Density, CDF, Survival, and Hazard Functions for Human Mortality



Some Useful Equivalencies

These definitions of the density, survival, and hazard functions allow us to express some equivalencies which will become important in later work. In particular, note that, because $S(t) = 1 - \int_0^t f(t)\,dt$, it is also the case that

$$f(t) = \frac{-\partial S(t)}{\partial t}. \tag{5}$$

This means that we can express the hazard rate as

$$\begin{aligned} h(t) &= \frac{\frac{-\partial S(t)}{\partial t}}{S(t)} \\ &= \frac{-\partial ln S(t)}{\partial t} \end{aligned} \tag{6}$$

In words: the hazard equals the negative partial derivative of the log–survival function with respect to t.

Now, given the way we've defined the density and survival functions, it must be the case that $S(0) = 1$ and $S(\infty) = 0$; these are true because (a) all observations start out the observation

period "alive," and (b) all observations will eventually have the event of interest.[1]

It is also useful to consider what happens if we integrate (6). Define

$$H(t) = \int_0^t h(t)\, dt \tag{7}$$

as the *integrated* (or *cumulative*) *hazard*. Note that, because of (6), we can write

$$\begin{aligned} H(t) &= \int_0^t \frac{-\partial ln S(t)}{\partial t}\, dt \\ &= -ln[S(t)] \end{aligned} \tag{8}$$

and, conversely,

$$S(t) = \exp^{-H(t)} \tag{9}$$

In words, the integrated hazard equals the negative log of the survival function, and the survival function is equal to the exponent of the negative integrated hazard.

The larger point is that any particular function $f(t)$, $F(t)$, $S(t)$, $h(t)$, and/or $H(t)$ determines the values of the rest of them. Thus, we can consider models of the density, hazard, or survival of some set of observations equivalently.

Censoring

Censoring is removal from the data for reasons other than the event of interest. A few key things to remember when thinking about censoring:

- Censoring is *defined by the researcher*.
    - Censoring usually means the "end" of observation.
    - One person's censoring may be another person's event of interest.
    - e.g., Congressional careers (retirement vs. defeat).
- Censoring is also typically assumed to be (conditionally) *independent of both the event of interest and the covariates*.
    - Indeed, it needs to be, in order to get the easy conditional probability rendition of the hazard given above.
    - Example of when it isn't: Congressional departure by reelection defeat and death...

[1] If you're skeptical about this latter assumption, give me a little while... we'll come back to it.

- Finally, note that censoring *doesn't mean that the observation provides no information for us...*
  - An observation censored at t still tells us that it has a survival time at least to t.
  - So, we can use this information as well.
  - We'll get to *how* we incorporate that information in just a minute.

Univariate and Bivariate Survival Analysis

Estimating $S(t)$

Typically, in the univariate context, $S(t)$ is the focus. Suppose for the moment that we have N observations, events are *absorbing*, and there are no ties. Then we can define:

$$\begin{aligned} n_t &= \text{the number of observations "at risk" for the event at } t, \text{ and} \\ d_t &= \text{the number of observations which experience the event at time } t. \end{aligned}$$

For any particular time t_k, we can get an estimate of the survival function $S(t)$ as the product of the conditional proportions of all survivors to that point. That is,

$$\widehat{S(t_k)} = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

This is known as the "Kaplan-Meier" estimate of the survivor function. Note a few things about this estimate:

- Event occurrences alter the estimate itself, but
- Censored observations just shrink the "risk sets" at their respective time points. Also,
- Note that this is *not* the same as the "proportion surviving to t."

If we're interested in inference, or just want to know the uncertainty surrounding our estimates, we need some measure of the variability of these estimates. The most common is the "Greenwood" variance estimator:

$$\text{Var}[\widehat{S(t_k)}] = \left[\widehat{S(t_k)}\right]^2 \sum_{t \leq t_k} \frac{d_t}{n_t(n_t - d_t)}$$

This is a product of the (squared) survival estimate and an "inflation factor." Note that:

- The variance is increasing in $S(t)$, and
- The variance is also increasing in d_t, but
- is decreasing in n_t (the size of the data).

There are a number of other alternatives for estimating the survival function (see the Hosmer and Lemeshow book, and/or Cox and Oakes), but they are all largely equivalent.

Estimating $H(t)$

Estimating the integrated hazard follows a similar logic. Consider the integrated/cumulative hazard $H(t)$. Since $H(t)$ is the cumulative conditional likelihood of the event of interest, a simple estimate is:

$$\hat{H}(t_k) = \sum_{t \leq t_k} \frac{d_t}{n_t}$$

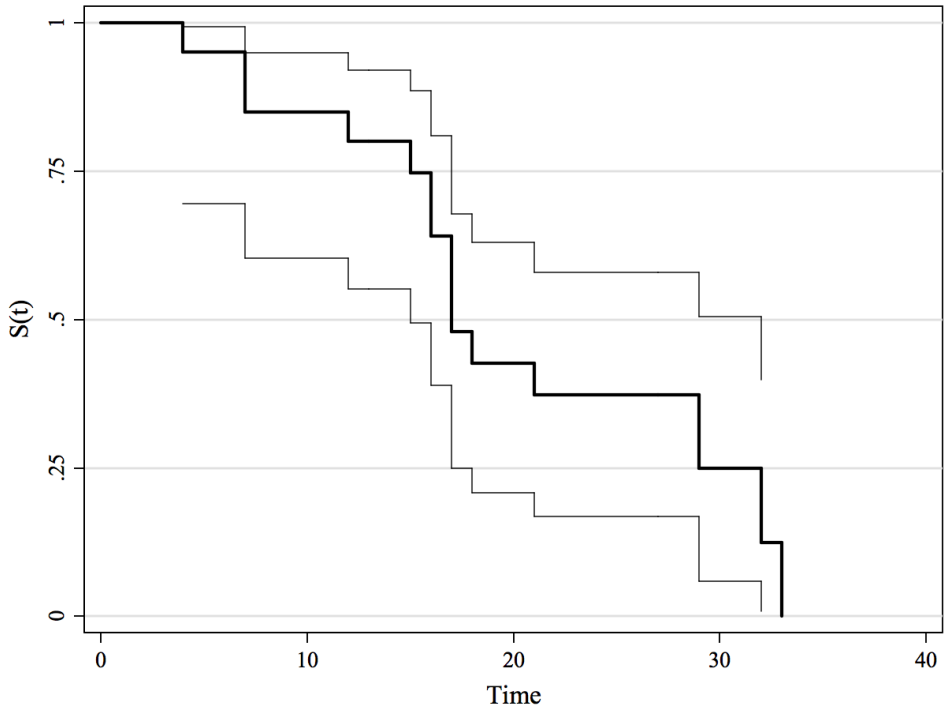
This is known as the "Nelson-Aalen" estimate of the cumulative hazard. It is a cumulative sum of the hazards – it "accumulates" over time, and so increases...

Q: What do we do with $S(t)$ and $H(t)$?

Answer #1: We *plot* them...

Consider first some made-up data ($N = 20$):

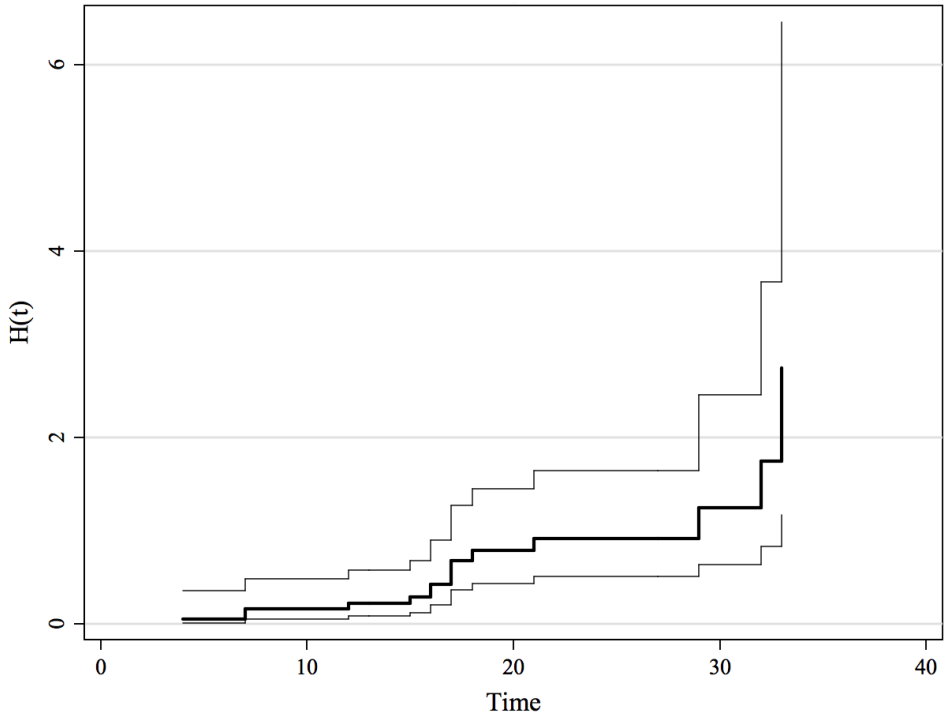
Figure 1: Kaplan-Meier Survival Function (with 95% Greenwood c.i.s)



Plotting the survival and cumulative hazards allows us to "see" what they look like...

- Hazards can be odd-shaped – more on the importance of this later...

Figure 2: Nelson-Aalen Cumulative Hazard Function (with 95% c.i.s)



- But don't read too much into this – these are unconditional hazards, and don't account for the effect(s) of covariates (which is typically what we're interested in...).

Plotting them also allows us to *compare* survival and/or hazard functions for different groups. Consider the survival and integrated hazard functions for two different groups in the data, defined by $X = 0$ and $X = 1$:

Figure 3: Kaplan-Meier Survival Function (with 95% Greenwood c.i.s), by X

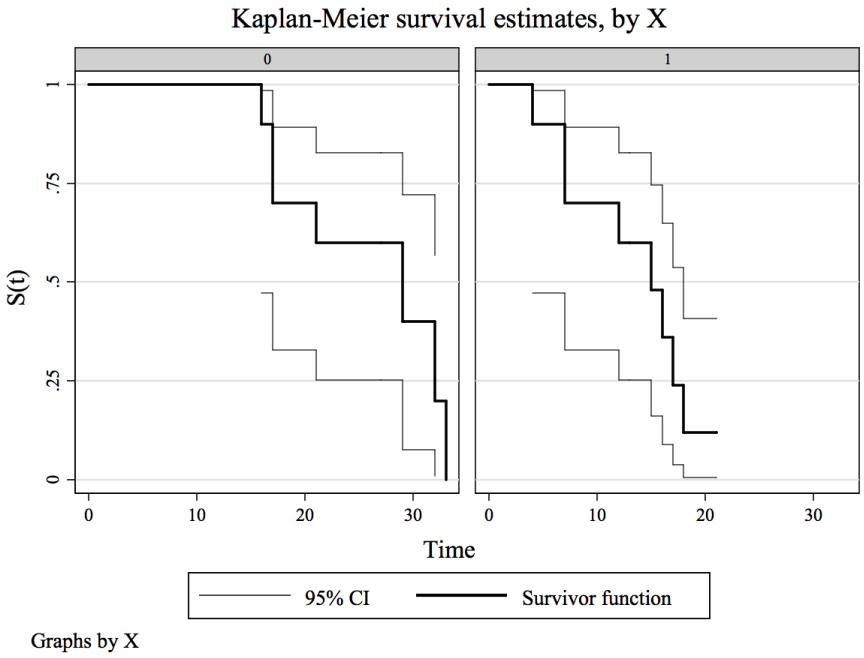
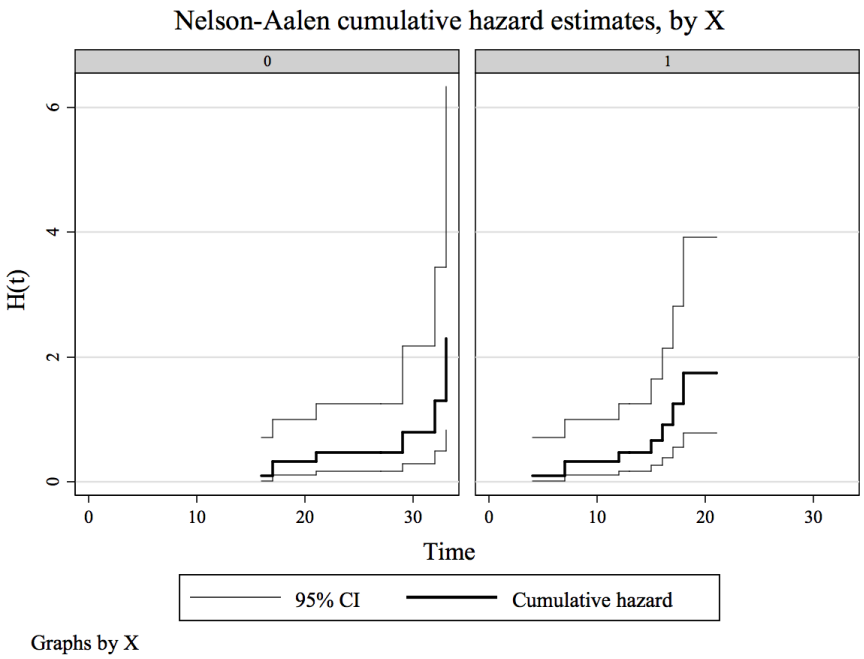


Figure 4: Nelson-Aalen Cumulative Hazard Function (with 95% c.i.s), by X



Here, we can see that the group defined by $X = 0$ has substantially longer average survival times than those for which $X = 1$. This suggests that the variable X has a positive influence on the hazard (i.e., that the presence of X leads to shorter survival times).

Answer #2: We can *test* for the equality of two functions...

Suppose we have two groups, defined as $X = 1$ for the "treatment" and $X = 0$ for the "placebo," and we want to know if the treatment increases survival times. We can think of this as analogous to a 2×2 table at each time point in the data:

	Treatment	Placebo	Total
Event	d_{1t}	d_{0t}	d_t
No Event	$n_{1t} - d_{1t}$	$n_{0t} - d_{0t}$	$n_t - d_t$
Total	n_{1t}	n_{0t}	n_t

Normally, we'd do a chi-square test here, using the observed and expected number of events per cell. In a survival context, the same idea applies, except that we conduct a similar test for each time period t. This is the *log-rank test*:

$$Q = \frac{\left[\sum(d_{1t} - \hat{e}_{1t})\right]^2}{\left[\frac{n_{1t}n_{0t}d_t(n_t - d_t)}{n_t^2(n_t - 1)}\right]}$$

where:

$$\hat{e}_{1t} = \frac{n_{1t}d_t}{n_t}$$

Now, think about this for a minute...

- $\hat{e}_{1t}$ is just the "expected" number of events at that time period...
- The numerator of Q is the sum of the (squared) observed minus expected events.
- The denominator reflects the possible ways in which the number of events occurring might have happened.

Q is $\sim \chi_1^2$, and tests the null hypothesis that $S_0(t) = S_1(t)$. The test assumes:

- Large N (that is, it is asymptotically χ_1^2), and
- Independent censoring.

A Diversion: Survival Models and Counting Processes

Beginning in the late 1970s, statisticians began to consider survival models as a form of counting process. The intuition is that survival models can be thought of as a very slow counting (e.g., Poisson) process...

- Many treatments of this are hard (Fleming and Harrington 1991, Andersen et al. 1993), but
- the intuition is useful, and point up commonalities b/w survival and event count models that will be very useful later.

Consider an observation i we're observing over time, waiting for an event to happen. Assume for now that:

- The event in question is *absorbing*,
- There is a duration variable called Y_i,
- The observation can also "fail" for other reasons: call the "censoring" duration defined by these other (unrelated) failures Z_i
- We observe a time variable $T_i = \min(Y_i, Z_i)$, and that
- We also have an indicator C_i of "how" the observation failed:
  - $C_i = 0$ if $T_i = Z_i$,
  - $C_i = 1$ if $T_i = Y_i$.
- Finally, assume that no two observations have events occurring at exactly the same time...

There are three key variables which are functions of time in the counting process approach:

1. the *counting process* variable,
2. the *risk indicator*, and
3. the *intensity process*.

The *counting process* is an indicator of whether, at a particular time t, the observation in question has experienced the event of interest (1) or not (0):

$$N_i(t) = I(T_i \leq t, C_i = 1) \tag{9}$$

Think of this as just "counting" the number of events which have occurred to i by time t. Consider the case of a model of retirements from the U.S. Supreme Court. If a justice lives for 20 years after being appointed to the Court ($Z_i = 20$), and retires after 15 years ($Y_i = 15$), the observations would have:

- $N_i(t) = 0$ for $t \in [1, 14]$, and
- $N_i(t) = 1$ for $t \in [15, 20]$.

The *risk indicator* is simply a dummy variable indicating whether observation i was "at risk" for the event of interest at time t:

$$R_i(t) = I(T_i > t) \tag{9}$$

So, for the aforementioned justice, we would have:

- $R_i(t) = 1$ for $t \in [0, 15]$, and
- $R_i(t) = 0$ for $t \in [16, 20]$.

since after $T = 15$ the justice has already retired, and so is no longer at risk for retiring.

Finally, the *intensity process* is a function of the risk indicator and the hazard:

$$\lambda_i(t) = R_i(t)h(t) \tag{9}$$

where $h(t)$ is the hazard function defined above (that is, $\Pr(t \leq T_i < t + dt, C_i = 1 | T_i \geq t)$.

- As we noted above, the hazard function at t equals the probability density at t divided by the survival function at t.
- The intensity process is thus a simple function of the hazard and the risk indicator:
    - The intensity is zero whenever the observation is no longer "at risk" for the event,
    - When the observation is "at risk" the intensity is a sort of "expected number of events."
        - Sort of like a binomial expected value, in that it is a function of the form $n \times p$, where
        - n is the possible number of events (either one or zero) at t, and
        - p is a term for the (here, conditional) probability (risk) of an event at t.

If we follow each observation from the beginning of the study period until either the occurrence of an event or censoring, we can think of the cumulative "expected number of events" as simply the integral, over time, of the intensity process:

$$\Lambda_i(t) = \int_0^t \lambda_i(t)\, dt \tag{9}$$

This is akin the "total number of expected events" up to time t.

So, the counting process $N_i(t)$ is the observed number of events, and the cumulative intensity $\Lambda_i(t)$ is the expected number of events. This suggests that we might think about their relationship as akin to a familiar regression-like model:

$$N_i(t) = \Lambda_i(t) + M_i(t) \tag{9}$$

where we have an observed "count" of events, which we can decompose into systematic and stochastic (random) parts. The difference between the two is thus something like a residual, and is called the *counting process martingale*:

$$M_i(t) = N_i(t) - \Lambda_i(t) \tag{9}$$

They're called that because, under proper model specification, Aalen (1978) showed that these residual-like things are martingales. A martingale process X_t is a stochastic process over time, in which, at any point in time t, the expected value of X_t at some future point, conditional on all past realizations of X_t, is equal to X_t. That is,

$$E(X_{t+s}|X_0, X_1, ...X_i, ...X_t) = X_t \, \forall s > 0$$

Martingale processes are used in pricing theory in economics, among other places. Martingales are very useful, in that they have a good deal of developed theory which we can use to prove the distributional traits of survival models. In particular, these residuals can be useful in that they allow for model checking, in a manner analogous to that for OLS.

We'll get to the issue of why all this ought to be of interest to you in the coming days...

Practical Issues

Data Structure and Organization

Data for duration models may be either *non-time-varying* or *time-varying*.

Non-Time-Varying Data

These are data that do not have covariates $\mathbf{X}$ that vary over time. Because of this, it is possible to record all the relevant information in them using one line of data for each unit of analysis:

id	durat	censor	timein	timeout	X1
1	4	0	30	34	0.12
2	2	1	12	14	0.19
3	5	1	5	10	0.09
...	...	...	...	...	...
N	10	1	21	31	0.22

Here,

- `id` is just the unit identification number,
- `durat` is the duration variable; that is, the survival (or censoring) time,
- `censor` is a (misnamed) variable that is coded 1 if the event occurred, and 0 if the observation was censored,
- `timein` indicates the time at which the observation "entered" the data, while
- `timeout` is the time on which the observation "exited" the data (either by the event occurring or through censoring), and
- `X1` is a (non-time-varying) covariate.

Time-Varying Data

Often it is the case that we have variables that vary over time, and we expect these variables to influence the hazard in question. For example, in a model of federal court retirements, we might expect the relative ideological distance between the sitting president and the judge in question to influence his/her decision to retire; this variable varies as different presidents come into and leave office.

To set up such data, we require one line of data, per unit of analysis, per time period. This means that, in addition to a unit identification variable, we also have a variable which indicates the time period, as well as our duration and censoring variables:

id	durat	censor	timein	timeout	X1	X2
1	1	0	30	31	0.12	331
1	2	0	31	32	0.12	412
1	3	0	32	33	0.12	405
1	4	0	33	34	0.12	416
2	1	0	12	13	0.19	226
2	2	1	13	14	0.19	296
3	1	0	5	6	0.09	253
3	2	0	6	7	0.09	311
3	3	0	7	8	0.09	327
3	4	0	8	9	0.09	344
3	5	1	9	10	0.09	301
...	...	...	...	...	...	...

These data are, in theory, the same as those above, but with a few slight changes:

- `id` is again the unit identification number,
- `durat` is the duration variable; that is, the survival (or censoring) time,

- `censor` is a (again, misnamed) variable that is coded 1 if the event occurred, and 0 if the observation was censored,
- `timein` indicates the time at which the observation "entered" the data, while
- `timeout` is the time on which the observation "exited" the data (either by the event occurring or through censoring). Also,
- `X1` is a (non-time-varying) covariate, while
- `X2` is a covariate that varies over time.

A variant on time-varying data are what are termed "counting process" data. The distinction is important when we get to the issue of "setting up" the data in our software; otherwise, we can just think of counting process data as a form of time-varying data.

Stata Preliminaries: `stset`

In Stata, all survival analysis commands require that we first indicate that the data are "survival-time" data. We do this by "`stset`-ing" the data.

- The basic command is

  ```
  . stset durat
  ```

  if the data are non-time-varying and there is no censoring at all; that is, if every observation experiences the event of interest ($C_i = 1 \, \forall \, i$).

- If some observations are censored, then we use

  ```
  . stset durat, failure(censor)
  ```

  where `censor` is the aforementioned censoring indicator.

- If not all the observations in the data entered at the same time, then it is (sometimes) important to record this fact; we do so by indicating, via two variables, when the observation entered and exited the data:

  ```
  . stset durat, failure(censor) en(timein) ex(timeout)
  ```

  Note that, provided that the data are set up as we described them above, we can use this command irrespective of whether or not the data are time-varying.

- Finally, the `id()` option is used to indicate that you have time-varying data(and possibly that observations experience more than one event):

```
. stset durat, failure(censor) en(timein) ex(timeout) id(id)
```

`stset` is a useful command...

- `stset` will create some new variables (`_t`, `_d`, etc.); don't worry about them.
- `stset` will also tell you if your data are "buggy" – e.g., have durations that "end before they start," and other such problems.

Basic Summaries, Plots and Tests

Stata has a host of simple descriptive, univariate, and bivariate commands specific to survival data and analysis...

- The basic data description command is `stdes`, which provides a host of information about your survival-time data once it has been `stset`.
- `stsum` summarizes survival-time data, including providing the overall incidence rate (that is, total number of events divided by total time at risk) and percentiles for the survival times. It can be used with the `by` option to get summary statistics for subgroups within the data.
- Stata will estimate basic univariate and bivariate survival quantities, using the `sts` command. This is a very flexible command; see the handout for some examples.

Analyzing Survival Data in R

In S-Plus and R, we have to create a "survival object," that tells the program that the data consist of information on durations and censoring status. In the simplest case, to do this, we use the `Surv` command:

`Surv(`*Duration, Censoring Indicator*`)`

So, for example:

```
exampledata<-read.dta(exampledata.dta)
survobject<-Surv(exampledata$duration, exampledata$censor)
```

A survival object of this sort is a $N \times 2$ matrix, containing the analysis duration T_i and the censoring indicator C_i. It is used when we have non-time-varying data.

For time-varying data, we need to adopt a little different means of creating our survival object. For such data, we can designate the start and stop times in the `Surv` command:

```
TVdata<-read.dta(TVdata.dta)
TVobject<-Surv(TVdata$starttime, TVdata$endtime, TVdata$censor)
```

Again, see the handout for examples of how to do this.

R is an excellent package for doing survival analysis. While we'll focus on Stata more in this class, we'll also use R from time to time, and include some illustrations as well.

Appendix: Stata Commands Used

Figure 1:

```
. sts graph, gwood recast(line) lcolor(black) lpattern(solid) lwidth(medthick)
ciopts(recast(rline) lcolor(black) lwidth(vthin)) ytitle(S(t)) xtitle(Time)
title(, size(zero)) caption(, size(zero)) legend(off) graphregion(margin(small))
```

Figure 2:

```
. sts graph, na cna recast(line) lcolor(black) lpattern(solid) lwidth(medthick)
ciopts(recast(rline) lcolor(black) lwidth(vthin)) ytitle(H(t)) xtitle(Time)
title(, size(zero)) caption(, size(zero)) legend(off) graphregion(margin(small))
```

Figure 3:

```
. sts graph, by(X) separate gwood recast(line) lcolor(black) lpattern(solid)
lwidth(medthick) ciopts(recast(rline) lcolor(black) lwidth(vthin)) ytitle(S(t))
xtitle(Time) title(Kaplan-Meier Survival Estimates, size(zero) ring(0))
subtitle(, size(small)) legend(off) graphregion(margin(small))
```

Figure 4:

```
. sts graph, by(X) na cna recast(line) lcolor(black) lpattern(solid)
lwidth(medthick) ciopts(recast(rline) lcolor(black) lwidth(vthin)) ytitle(H(t))
xtitle(Time) title(, size(zero) ring(0)) subtitle(, size(small)) legend(off)
graphregion(margin(small))
```