Sociology 761 John Fox

Lecture Notes

Introduction to Survival Analysis

1. Introduction

- *Survival analysis* encompasses a wide variety of methods for analyzing the timing of events.
  - The prototypical event is *death*, which accounts for the name given to these methods.
  - But survival analysis is also appropriate for many other kinds of events, such as criminal recidivism, divorce, child-bearing, unemployment, and graduation from school.
- The wheels of survival analysis have been reinvented several times in different disciplines, where terminology varies from discipline to discipline:
  - *survival analysis* in biostatistics, which has the richest tradition in this area;
  - *failure-time analysis* in engineering;
  - *event-history analysis* in sociology.

- Sources for these lectures on survival analysis:
  - Paul Allison, *Survival Analysis Using the SAS System, Second Edition*, SAS Institute, 2010.
  - Paul Allison, *Event History and Surival Analyis, Second Edition*, Sage, 2014.
  - George Barclay, *Techniques of Population Analysis*, Wiley, 1958.
  - D. R. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, 1984.
  - David Hosmer, Jr., Stanley Lemeshow, and Susanne May, *Applied Survival Analysis, Second Edition*, Wiley, 2008.
  - Terry Therneau and Patricia Grambsch, *Modeling Survival Data*, Springer, 2000.

- *Outline*:
  - The nature of survival data.
  - Life tables.
  - The survival function, the hazard function, and their relatives.
  - Estimating the survival function.
  - The basic Cox proportional-hazards regression model
  - Topics in Cox regression:
    - Time-dependent covariates.
    - Model diagnostics.
    - Stratification.
    - Estimating the survival function.

2. The Nature of Survival Data: Censoring

- Survival-time data have two important special characteristics:
  - (a) Survival times are non-negative, and consequently are usually positively skewed.
    - This makes the naive analysis of untransformed survival times unpromising.
  - (b) Typically, some subjects (i.e., units of observation) have *censored* survival times
    - That is, the survival times of some subjects are not observed, for example, because the event of interest does not take place for these subjects before the termination of the study.
    - Failure to take censoring into account can produce serious bias in estimates of the distribution of survival time and related quantities.

- It is simplest to discuss censoring in the context of a (contrived) study:
  - Imagine a study of the survival of heart-lung transplant patients who are followed up after surgery for a period of 52 weeks.
  - The event of interest is death, so this is literally a study of *survival time*.
  - Not all subjects will die during the 52-week follow-up period, but all will die eventually.
- Figure 1 depicts the survival histories of six subjects in the study, and illustrates several kinds of censoring (as well as uncensored data):
  - My terminology here is not altogether standard, and does not cover all possible distinctions (but is, I hope, clarifying).
  - Subject 1 is enrolled in the study at the date of transplant and dies after 40 weeks; this observation is *uncensored*.
    - The solid line represents an *observed period at risk*, while the solid circle represents an *observed event*.

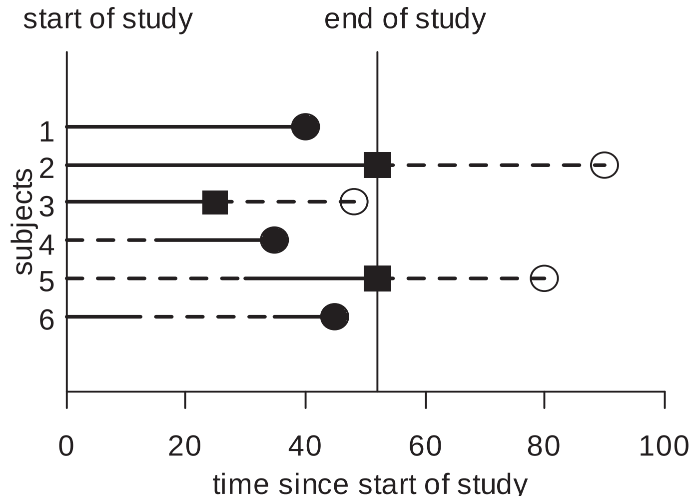

Figure 1. Data from an imagined study illustrating various kinds of subject histories: Subject 1, uncensored; 2, fixed-right censoring; 3, random-right censoring; 4 and 5, late entry; 6, multiple intervals of observation.

- Subject 2 is also enrolled at the date of transplant and is alive after 52 weeks; this is an example of *fixed-right censoring.*
  - The broken line represents an *unobserved period at risk*; the filled box represents the censoring time; and the open circle represents an *unobserved event*.
  - The censoring is fixed (as opposed to random) because it is determined by the procedure of the study, which dictates that observation ceases 52 weeks after transplant.
  - This subject dies after 90 weeks, but the death is unobserved and thus cannot be taken into account in the analysis of the data from the study.
  - Fixed-right censoring can also occur at different survival times for different subjects when a study terminates at a predetermined date.

- Subject 3 is enrolled in the study at the date of transplant, but is lost to observation after 30 weeks (because he ceases to come into hospital for checkups); this is an example of *random-right censoring*.
  - The censoring is random because it is determined by a mechanism out of the control of the researcher.
  - Although the subject dies within the 52-week follow-up period, this event is unobserved.
  - Right censoring — both fixed and random — is the most common kind.

- Subject 4 joins the study 15 weeks after her transplant and dies 20 weeks later, after 35 weeks; this is an example of *late entry* into the study.
  - Why can't we treat the observation as observed for the full 35-week period? After all, we *know* that subject 4 survived for 35 weeks after transplant.
  - The problem is that other potential subjects may well have died unobserved during the first 15 weeks after transplant, without enrolling in the study; treating the unobserved period as observed thus biases survival time upwards.
  - That is, had this subject died before the 15th week, she would not have had the opportunity to enroll in the study, and the death would have gone unobserved.
- Subject 5 joins the study 30 weeks after transplant and is observed until 52 weeks, at which point the observation is censored.
  - The subject's death after 80 weeks goes unobserved.

- Subject 6 enrolls in the study at the date of transplant and is observed alive up to the 10th week after transplant, at which point this subject is lost to observation until week 35; the subject is observed thereafter until death at the 45th week.
  - This is an example of *multiple intervals of observation*.
  - We only have an opportunity to observe a death when the subject is under observation.

▶ *Survival time*, which is the object of study in survival analysis, should be distinguished from *calendar time*.

- Survival time is measured relative to some relevant time-origin, such as the date of transplant in the preceding example.
- The appropriate time origin may not always be obvious.
- When there are alternative time origins, those not used to define survival time may be used as explanatory variables.
  - In the example, where survival time is measured from the date of transplant, *age* might be an appropriate explanatory variable.

- In most studies, different subjects will enter the study at different dates — that is, at different calendar times.
- Imagine, for example, that Figure 2 represents the survival times of three patients who are followed for at most 5 years after bypass surgery:
  - Panel (a) of the figure represent calendar time, panel (b) survival time.
  - Thus, subject 1 is enrolled in the study in 2000 and is alive in 2005 when follow-up ceases; this subject's death after 8 years in 2008 goes unobserved (and is an example of fixed-right censoring).
  - Subject 2 enrolls in the study in 2004 and is observed to die in 2007, surviving for 3 years.
  - Subject 3 enrolls in 2005 and is randomly censored in 2009, one year before the normal termination of follow-up; the subject's death after 7 years in 2012 is not observed.

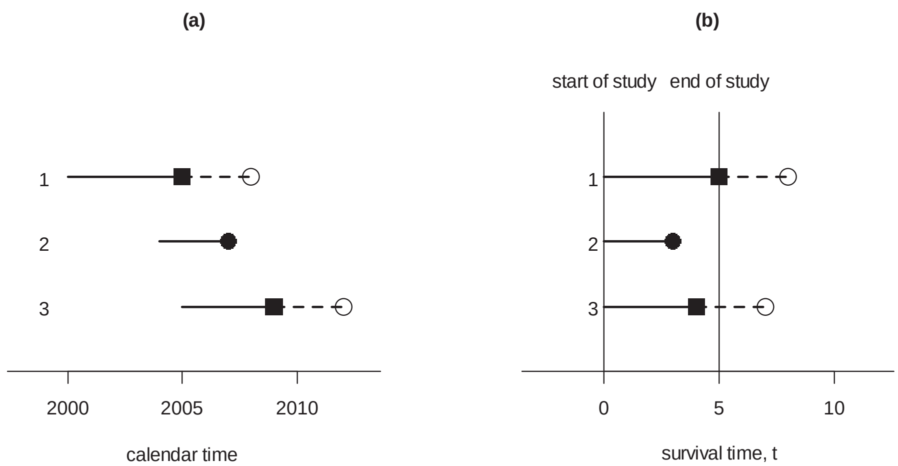

Figure 2. Calendar time (a) vs. survival time (b).

- Figure 3 shows calendar time and survival time for subjects in a study with a fixed *date* of termination: observation ceases in 2010.
  - Thus, the observation for subject 3, who is still alive in 2010, is fixed-right censored.
  - Subject 1, who drops out in 2009 before the termination date of the study, is randomly censored.

▶ Methods of survival analysis will treat as *at risk for an event* at survival time t those subjects who are under observation at that survival time.

- By considering only those subjects who are under observation, unbiased estimates of survival times, survival probabilities, etc., can be made, as long as those under observation are representative of *all* subjects.
- This implies that the censoring mechanism is unrelated to survival time, perhaps after accounting for the influence of explanatory variables.

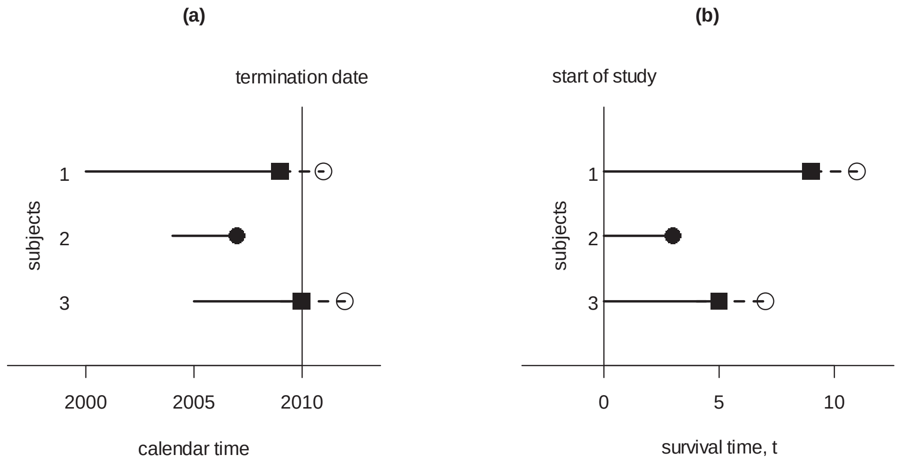

Figure 3. A study with a fixed date of termination.

- That is, the distribution of survival times of subjects who are censored at a particular time t is no different from that of subjects who are still under observation at this time.
- When this is the case, censoring is said to be *noninformative* (i.e., about survival time).
- With fixed censoring this is certainly the case.
- With random censoring, it is quite possible that survival time is not independent of the censoring mechanism.
  - For example, very sick subjects might tend to drop out of a study shortly prior to death and their deaths may consequently go unobserved, biasing estimated survival time upwards.
  - Another example: In a study of time to completion of graduate degrees, relatively weak students who would take a long time to finish are probably more likely to drop out than stronger students who tend to finish earlier, biasing estimated completion time downwards.

– Where random censoring is an inevitable feature of a study, it is important to include explanatory variables that are probably related to both censoring and survival time — e.g., seriousness of illness in the first instance, grade-point average in the second.

- Right-censored survival data, therefore, consist of two or three components:
  (a) The survival time of each subject, or the time at which the observation for the subject is censored.
  (b) Whether or not the subject's survival time is censored.
  (c) In most interesting analyses, the values of one or more explanatory variables (covariates) thought to influence survival time.
    – The values of (some) covariates may vary with time.
- Late entry and multiple periods of observation introduce complications, but can be handled by focusing on each interval of time during which a subject is under observation, and observing whether the event of interest occurs during that interval.

3. Life Tables

- At the dawn of modern statistics, in the 17th century, John Graunt and William Petty pioneered the study of mortality.
- The construction of life tables dates to the 18th century.
- A *life table* records the pattern of mortality with age for some population and provides a basis for calculating the expectation of life at various ages.
  - These calculations are of obvious actuarial relevance.
- Life tables are also a good place to start the study of modern methods of survival analysis:
  - Given data on mortality, the construction of a life table is largely straightforward.
  - Some of the ideas developed in studying life tables are helpful in understanding basic concepts in survival analysis.

  - Censoring is not a serious issue in constructing a life table.
- Here is an illustrative life table constructed for Canadian females using mortality data from Statistics Canada for the period 2009-2011 (the most recent data I could locate):

Age x	l_x	d_x	p_x	q_x	L_x	T_x	e_x
0	100000	449	.99551	.00449	99686	8359810	83.60
1	99551	21	.99979	.00021	99538	8260124	82.97
2	99530	16	.99984	.00016	99522	8160586	81.99
3	99514	13	.99987	.00013	99508	8061064	81.00
4	99501	10	.99990	.00010	99496	7961556	80.01
5	99491	9	.99991	.00009	99487	7862060	79.02
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
107	225	98	.56538	.43462	176	363	1.61
108	127	58	.54614	.45386	98	187	1.47
109	69	33	.52778	.47222	53	89	1.29

- The columns of the life table have the following interpretations:
  - x is age in years; in some instances (as explained below) it represents exact age at the xth birthday, in others it represents the one-year interval from the xth to the $(x + 1)$st birthday.
    - This life table is constructed for single years of age, but other intervals — such as five or ten years — are also common.
  - l_x (the lower-case letter "el") is the number of individuals surviving to their xth birthday.
    - The original number of individuals in the cohort, l_0 (here $100,000$), is called the *radix* of the life table.
    - Although a life table can be computed for a real *birth cohort* (individuals born in a particular year) by following the cohort until everyone is dead, it is more common, as here, to construct the table for a *synthetic cohort*.
    - A synthetic cohort is an imaginary group of people who die according to current age-specific rates of mortality.

    - Because mortality rates typically change over time, a synthetic cohort does not correspond to any real cohort.
  - d_x is the number of individuals dying between their xth and $(x + 1)$st birthdays.
  - p_x is proportion of individuals age x who survive to their $(x + 1)$st birthday — that is, the conditional probability of surviving to age $x + 1$ given that one has made it to age x.
  - q_x is the age-specific mortality rate — that is, the proportion of individuals age x who die during the following year.
    - q_x is the key column in the life table in that all other columns can be computed from it (and the radix), and it is the link between mortality data and the life table (as explained later).
    - q_x is the complement of p_x, that is, $q_x = 1 - p_x$.

- L_x is the number of person-years lived between birthdays x and $x + 1$.
  - At most ages, it is assumed that deaths are evenly distributed during the year, and thus $L_x = l_x - \frac{1}{2}d_x$.
  - In early childhood, mortality declines rapidly with age, and so during the first two years of life it is usually assumed that there is more mortality earlier in the year. (The details aren't important to us.)
- T_x is the number of person-years lived after the xth birthday.
  - T_x simply cumulates L_x from year x on.
  - A small censoring problem occurs at the end of the table if some individuals are still alive after the last year. One approach is to assume that those still alive live on average one more year.
    - In the example table, 8 people are alive at the end of their 109th year.
- e_x is the expectation of life remaining at birthday x — that is, the number of additional years lived on average by those making it to their xth birthday.

  - $e_x = T_x / l_x$.
  - e_0, the expectation of life at birth, is the single most commonly used number from the life table.

▶ This description of the columns of the life table also suggests how to compute a life table given age-specific mortality rates q_x:

- Compute the expected number of deaths $d_x = q_x l_x$.
  - d_x is rounded to the nearest integer before proceeding. This is why a large number is used for the radix.
- Then the number surviving to the next birthday is $l_{x+1} = l_x - d_x$.
- The proportion surviving is $p_x = 1 - q_x$.
- Formulas have already been given for L_x, T_x, and e_x.

- Figure 4 graphs the age-specific mortality rate q_x and number of survivors l_s as functions of age for the illustrative life table.
- As mentioned, the age-specific mortality rates q_x provide the link between the life table and real mortality data.
  - q_x must be *estimated* from real data.
  - The nature of mortality data varies with jurisdiction, but in most countries there is no population registry that lists the entire population at every moment in time.
  - Instead, it is typical to have estimates of population by age obtained from *censuses* and (possibly) sample surveys, and to have records of deaths (which, along with records of births, constitute so-called *vital statistics*).

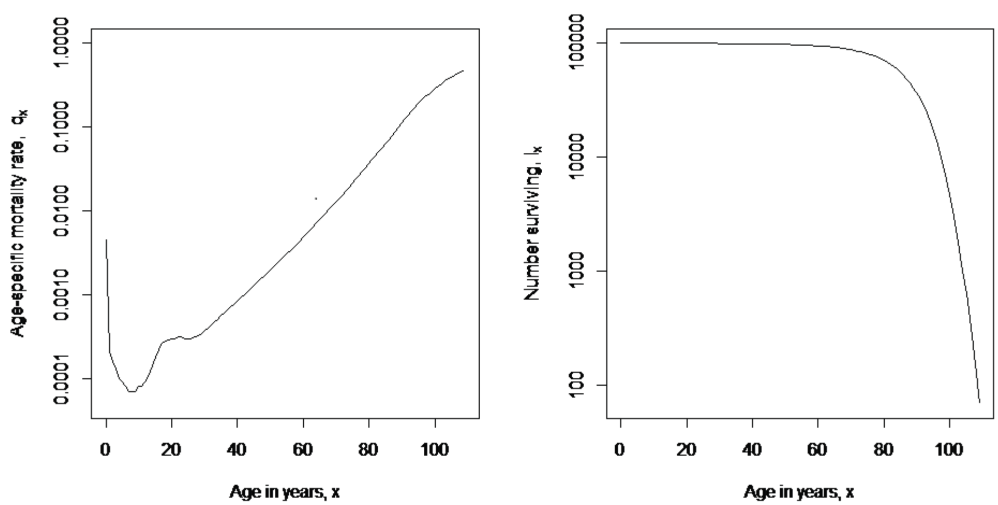

Figure 4. Age-specific mortality rates q_x and number of individuals surviving l_x as functions of age x, for Canadian females, 2009–2011. Both plots use logarithmic vertical axes.

- Population estimates refer to a particular point in time, usually the middle of the year, while deaths are usually recorded for a calendar year.
- There are several ways to proceed, and a few subtleties, but the following simple procedure is reasonable.
  - Let P_x represent the number of individuals of age x alive at the middle of the year in question.
  - Let D_x represent the number of individuals of age x who die during the year.
  - $M_x = D_x/P_x$ is the *age-specific death rate*. It differs from the age-specific *mortality* rate q_x in that some of the people who died during the year expired before the mid-year enumeration.

  - Assuming that deaths occur evenly during the year, an estimate of q_x is given by
$$q_x = \frac{D_x}{P_x + \frac{1}{2}D_x}$$
  - Again, an adjustment is usually made for the first year or two of life.

4. The Survival Function, the Hazard Function, and their Relatives

- The survival time T may be thought of as a random variable.
- There are several ways to represent the distribution of T: The most familiar is likely the *probability-density function*.
  - The simplest parametric model for survival data is the *exponential distribution*, with density function
$$p(t) = \lambda e^{-\lambda t}$$
  - The exponential distribution has a single *rate parameter* λ; the interpretation of this parameter is discussed below.
  - Figure 5 gives examples of several exponential distributions, with rate parameters $\lambda = 0.5$, 1, and 2.
  - It is apparent that the larger the rate parameter, the more the density is concentrated near 0.

  - The mean of an exponential distribution is the inverse of the rate parameters, $E(T) = 1/\lambda$.
- The *cumulative distribution function (CDF)*, $P(t) = \Pr(T \leq t)$, is also familiar.
  - For the exponential distribution
$$P(t) = \int_0^t p(x)dx = 1 - e^{-\lambda t}$$
  - The exponential CDF is illustrated in panel (b) of Figure 6 for rate parameter $\lambda = 0.5$.

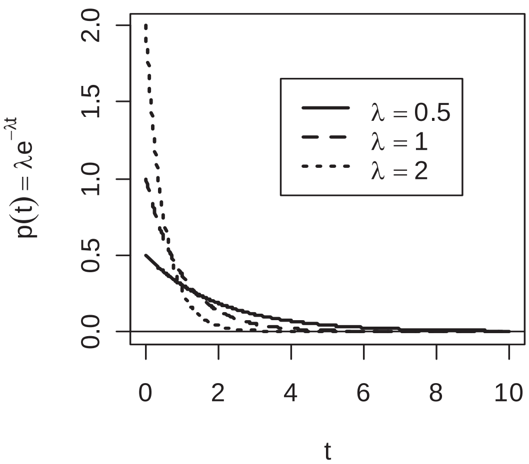

Figure 5. Exponential density functions for various values of the rate parameter λ.

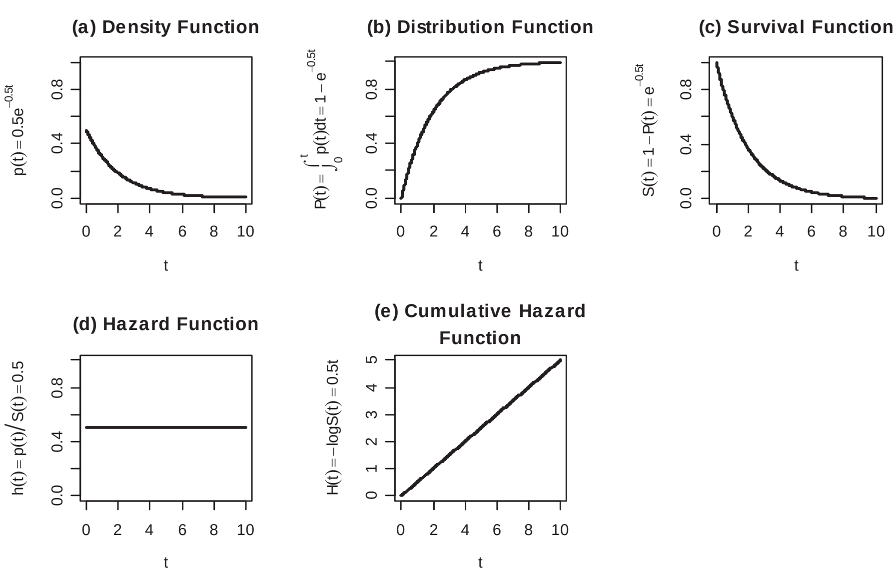

Figure 6. Several representations of an exponential distribution with rate parameter $\lambda = 0.5$.

- The *survival function*, giving the probability of surviving to time t, is the complement of the cumulative distribution function, $S(t) = \Pr(T > t) = 1 - P(t)$.
  - For the exponential distribution, therefore, the survival function is
$$S(t) = \int_t^\infty p(x)dx = e^{-\lambda t}$$
  - The exponential survival function is illustrated in Figure 6 (c) for rate parameter $\lambda = 0.5$.

4.1 The Hazard Rate and the Hazard Function

- Recall that in the life table, q_x represents the conditional probability of dying before age $x+1$ given that one has survived to age x.
  - If r is the radix of the life table, then the probability of living until age x is $l_x^* = l_x/r$ (i.e., the number alive at age x divided by the radix).
  - Likewise the *unconditional* probability of dying between birthdays x and $x+1$ is $d_x^* = d_x/r$ (i.e., the number of deaths in the one-year interval divided by the radix).
  - The *conditional* probability of dying in this interval is therefore
$$\begin{aligned} \Pr(\text{death between } x \text{ and } x+1 \mid \text{alive at age } x) &= \frac{\Pr(\text{death between } x \text{ and } x+1)}{\Pr(\text{alive at age } x)} \\ q_x &= \frac{d_x^*}{l_x^*} = \frac{d_x/r}{l_x/r} = \frac{d_x}{l_x} \end{aligned}$$

- Now, switching to our current notation, consider what happens to this conditional probability at time t as the time interval shrinks towards 0:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq T < t + \Delta t)|(T \geq t)]}{\Delta t} = \frac{p(t)}{S(t)}$$

- This continuous analog of the age-specific mortality rate is called the *hazard rate* (or just the *hazard*), and $h(t)$, the hazard rate as a function of survival time, is the *hazard function*.
- Note that the hazard is *not* a conditional probability (just as a probability-density is not a probability).
  - In particular, although the hazard cannot be negative, it can be larger than 1.

- The hazard is interpretable as the expected number of events per individual per unit of time.
  - Suppose that the hazard at a particular time t is $h(t) = 0.5$, and that the unit of time is one month.
  - This means that on average 0.5 events will occur per individual at risk per month (during a period in which the hazard remains constant at this value).
  - Imagine, for example, that the 'individuals' in question are light bulbs, of which there are 1000.
    - Imagine, further, that whenever a bulb burns out we replace it with a new one.
    - Imagine, finally, that the hazard of a bulb burning out is 0.5 and that this hazard remains constant over the life-span of bulbs.
    - Under these circumstances, we would expect 500 bulbs to burn out on average per month.

– Put another way, the expected life-span of a bulb is $1/0.5 = 2$ months.

- When, as in the preceding example, the hazard is constant, survival time is described by the exponential distribution, for which the hazard function is

$$h(t) = \frac{p(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

  – This is why λ is called the *rate* parameter of the exponential distribution.

  – The hazard function $h(t)$ for the exponential distribution with rate $\lambda = 0.5$ is graphed in Figure 6 (d).

- More generally, the hazard need not be constant.
- Because it expresses the instantaneous risk of an event, the hazard rate is the natural response variable for regression models for survival data.

▶ Return to the light-bulb example.

- The cumulative hazard function $H(t)$ represents the expected number of events that have occurred by time t; that is,

$$H(t) = \int_0^t h(x)dx = -\log_e S(t)$$

- For the exponential distribution, the cumulative hazard is proportional to time

$$H(t) = \lambda t$$

  – This is sensible because the hazard is constant.

- See Figure 6 (e) for the cumulative hazard function for the exponential distribution with rate $\lambda = 0.5$.

  – In this distribution, expected events accumulate at the rate of $1/2$ per unit time.

  – Thus, if our 1000 light bulbs burn for 12 months, replacing bulbs as needed when they fail, we expect to have to replace $1000 \times 0.5 \times 12 = 6000$ bulbs.

- Although the exponential distribution is particularly helpful in interpreting the meaning of the hazard rate, it is not a realistic model for most social (or biological) processes, where the hazard rate is not constant over time.
  - The hazard of death in human populations is relatively high in infancy, declines during childhood, stays relatively steady during early adulthood, and rises through middle and old age.
  - The hazard of completing a nominally four-year university undergraduate degree is essentially zero for at least a couple of years, rises to four years, and declines thereafter.
  - The hazard of a woman having her first child rises and then falls with time after menarche.

- There are other probability distributions that are used to model survival data and that have variable hazard rates.
  - Two common examples are the Gompertz distributions and the Weibull distributions, both of which can have declining, increasing, or constant hazards, depending upon the parameters of the distribution.
    - For a constant hazard, the Gompertz and Weibull distributions reduce to the exponential.
  - We won’t pursue this topic further, however, because we will not model the hazard function parametrically.

5. Estimating the Survival Function

- Allison (2010, 2014) discusses data from a study by Rossi, Berk, and Lenihan (1980) on recidivism of 432 prisoners during the first year after their release from Maryland state prisons.
  - Data on the released prisoners was collected weekly.
  - I will use this data set to illustrate several methods of survival analysis.
  - The variables in the data set are as follows (using the variable names employed by Allison):
    - `week`: The week of first arrest of each former prisoner; if the prisoner was not rearrested, this variable is censored and takes on the value `52`.
    - `arrest`: The censoring indicator, coded `1` if the former prisoner was arrested during the period of the study and `0` otherwise.

- `fin`: A dummy variable coded `1` if the former prisoner received financial aid after release from prison and `0` otherwise. The study was an experiment in which financial aid was randomly provided to half the prisoners.
- `age`: The former prisoner's age in years at the time of release.
- `race`: A dummy variable coded `1` for blacks and `0` for others.
- `wexp`: Work experience, a dummy variable coded `1` if the former prisoner had full-time work experience prior to going to prison and `0` otherwise.
- `mar`: Marital status, a dummy variable coded `1` if the former prisoner was married at the time of release and `0` otherwise.
- `paro`: A dummy variable coded `1` if the former prisoner was released on parole and `0` otherwise.
- `prio`: The number of prior incarcerations.

- `educ`: Level of education, coded as follows:
  - 2: 6th grade or less
  - 3: 7th to 9th grade
  - 4: 10th to 11th grade
  - 5: high-school graduate
  - 6: some postsecondary or more
- `emp1`–`emp52`: 52 dummy variables, each coded `1` if the former prisoner was employed during the corresponding week after release and `0` otherwise.

▶ For the moment, we will simply examine the data on `week` and `arrest` (that is, survival or censoring time and the censoring indicator).

- The following table shows the data for 10 of the former prisoners (selected not quite at random) and arranged in ascending order by week.

- Where censored and uncensored observations occurred in the same week, the uncensored observations appear first (this happens only in week 52 in this data set).
- The observation for prisoner 999, which is censored at week 30, is made up, since in this study censoring took place only at 52 weeks.

Prisoner	`week`	`arrest`
174	17	1
411	27	1
999	30	0
273	32	1
77	43	1
229	43	1
300	47	1
65	52	1
168	52	0
408	52	0

- More generally, suppose that we have n observations and that there are m unique event times arranged in ascending order, $t_{(1)}, t_{(2)}, ..., t_{(m)}$.
- The *Kaplan-Meier estimator*, the most common method of estimating the survival function $S(t) = \Pr(T > t)$, is computed as follows:
  - Between $t = 0$ and $t = t_{(1)}$ (i.e., the time of the first event), the estimate of the survival function is $\widehat{S}(t) = 1$.
  - Let n_i represent the number of individuals *at risk* for the event at time $t_{(i)}$.
    - The number at risk includes those for whom the event has not yet occurred, including individuals whose event times have not yet been censored.
  - Let d_i represent the number of events ("deaths") observed at time $t_{(i)}$.
    - If the measurement of time were truly continuous, then we would never observe "tied" event times and $t_{(i)}$ would always be 1.

- Real data, however, are rounded to a greater or lesser extent, resulting in *interval censoring* of survival time.
- For example, the recidivism data are collected at weekly intervals.
- Whether or not the rounding of time results in serious problems for data analysis depends upon degree.
- Some data are truly discrete and may require special methods: For example, the number of semesters required for graduation from university.

- The *conditional* probability of surviving past time $t_{(i)}$ given survival to that time is estimated by $(n_i - d_i)/n_i$.
- Thus, the *unconditional* probability of surviving past any time t is estimated by

$$\widehat{S}(t) = \prod_{t(i) \leq t} \frac{n_i - d_i}{n_i}$$

- This is the Kaplan-Meier estimate.

- For the small example:
  - The first arrest time is 17 weeks, and so $\widehat{S}(t) = 1$ for $t < t_{(1)} = 17$.
  - $d_1 = 1$ person was arrested at week 17 , when $n_1 = 10$ people were at risk, and so
$$\widehat{S}(t) = \frac{10-1}{10} = .9,$$
$$\text{for } t_{(1)} = 17 \leq t < t_{(2)} = 27$$
  - As noted, the second arrest occurred in week 27, when $d_2 = 1$ person was arrested and $n_2 = 9$ people were at risk for arrest:
$$\widehat{S}(t) = \frac{10-1}{10} \times \frac{9-1}{9} = .8,$$
$$\text{for } t_{(2)} = 27 \leq t < t_{(3)} = 32$$

- At $t_{(3)} = 32$, there was $d_3 = 1$ arrest and, because the observation for prisoner 999 was censored at week 30, there were $n_3 = 7$ people at risk:
$$\widehat{S}(t) = \frac{10-1}{10} \times \frac{9-1}{9} \times \frac{7-1}{7} = .6857,$$
$$\text{for } t_{(3)} = 32 \leq t < t_{(4)} = 43$$
- At $t_{(4)} = 43$, there were $d_4 = 2$ arrests and $n_4 = 6$ people at risk:
$$\widehat{S}(t) = \frac{10-1}{10} \times \frac{9-1}{9} \times \frac{7-1}{7} \times \frac{6-2}{6} = .4571,$$
$$\text{for } t_{(4)} = 43 \leq t < t_{(5)} = 47$$

- At $t_{(5)} = 47$, there was $d_5 = 1$ arrest and $n_5 = 4$ people at risk:

$$\widehat{S}(t) = \frac{10-1}{10} \times \frac{9-1}{9} \times \frac{7-1}{7} \times \frac{6-2}{6} \times \frac{4-1}{4} = .3429,$$

for $t_{(5)} = 47 \leq t < t_{(6)} = 52$

- At $t_{(6)} = 52$, there was $d_6 = 1$ arrest and $n_6 = 3$ people at risk (since censored observations are treated as observed *up to and including* the time of censoring):

$$\widehat{S}(t) = \frac{10-1}{10} \times \frac{9-1}{9} \times \frac{7-1}{7} \times \frac{6-2}{6} \times \frac{4-1}{4} \times \frac{3-1}{3} = .2286,$$

for $t = t_{(6)} = 52$

- Because the last two observations are censored, the estimator is undefined for $t > t_{(6)} = 52$.
  - Were the last observation an event, then the estimator would descend to 0 at $t_{(m)}$.
- The Kaplan-Meier estimate is graphed in Figure 7.

▶ It is, of course, useful to have information about the sampling variability of the estimated survival curve.

- An estimate of the variance of $\widehat{S}(t)$ is given by *Greenwood's formula*:

$$\widehat{V}[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{t(i) \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- The square-root of $\widehat{V}[\widehat{S}(t)]$ is the standard error of the Kaplan-Meier estimate, and $\widehat{S}(t) \pm 1.96\sqrt{\widehat{V}[\widehat{S}(t)]}$ gives a point-wise 95-percent confidence envelope around the estimated survival function.

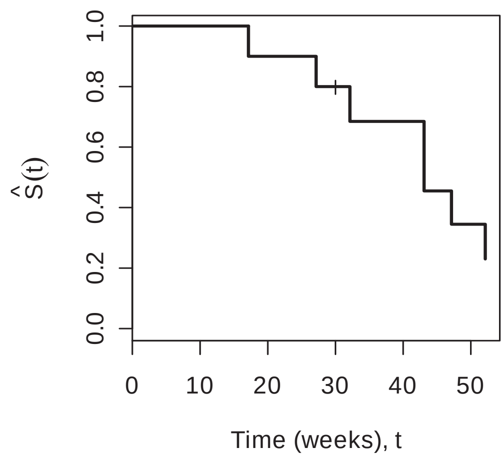

Figure 7. Kaplan-Meier estimate of the survival function for 10 observations drawn from the recidivism data. The censored observation at 30 weeks is marked with a "+."

- Figure 8 shows the Kaplan-Meier estimate of time to first arrest for the full recidivism data set.

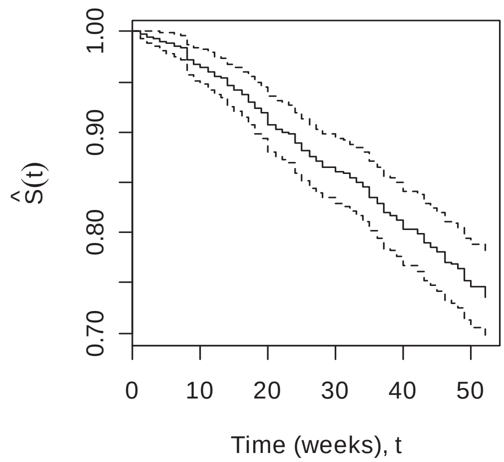

Figure 8. Kaplan-Meier estimate of the survival function for Rossi et al.'s recidivism data. The broken lines give a 95-percent point-wise confidence envelope around the estimated survival curve.

5.1 Estimating Quantiles of the Survival-Time Distribution

- Having estimated a survival function, it is often of interest to estimate quantiles of the survival distribution, such as the median time of survival.
- If there are any censored observations at the end of the study, as is often the case, it is not possible to estimate the expected (i.e., the mean) survival time.
- The estimated pth quantile of survival time is
$$\widehat{t}_p = \min(t\colon \widehat{S}(t) \leq p)$$
  - For example, the estimated median survival time is
$$\widehat{t}_{.5} = \min(t\colon \widehat{S}(t) \leq .5)$$
  - This is equivalent to drawing a horizontal line from the vertical axis to the survival curve at $\widehat{S}(t) = .5$; the left-most point of intersection with the curve determines the median, as illustrated in Figure 9 for the small sample of 10 observations drawn from Rossi et al.'s data.

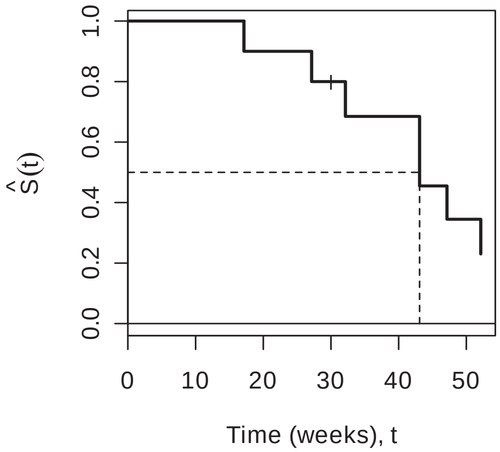

Figure 9. Determining the median survival time for the sample of 10 observations drawn from Rossi et al.'s recidivism data. The median survival time is $\widehat{t}_{.5} = 43$.

- It is not possible to estimate the median survival time for the full data set, since the estimated survival function doesn't dip to $.5$ during the 52-week period of study.

5.2 Comparing Survival Functions

- There are several tests to compare survival functions between two or among several groups.
- Most tests can be computed from contingency tables for those at risk at each event time.
  - Suppose that there are two groups, and let $t_{(i)}$ represent the ith ordered event time in the two groups combined.
  - Form the following contingency table:

	Group 1	Group 2	Total
Event	d_{1i}	d_{2i}	d_i
No event	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
At Risk	n_{1i}	n_{2i}	n_i

where
    - d_{ji} is the number of people experiencing the event at time $t_{(i)}$ in group j;

    - n_{ji} is the number of people at risk in group j at time $t_{(i)}$;
    - d_i is the total number experiencing the event in both groups;
    - n_i is the total number at risk.
  - Unless there are tied event times, one of d_{1i} and d_{2i} will be 1 and the other will be 0.
  - Under the hypothesis that the population survival functions are the same in the two groups, the estimated expected number of individuals experiencing the event at $t_{(i)}$ in group j is
$$\widehat{e}_{ji} = \frac{d_i n_{ji}}{n_i}$$
  - The variance of $\widehat{e}_{ji}$ may be estimated as
$$\widehat{V}(\widehat{e}_{ji}) = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$
  - There is one such table for every observed event time, $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$.

- A variety of test statistics can be computed using the expected and observed counts; probably the simplest and most common is the *Mantel-Haenszel* or *log-rank test*,

$$Q = \frac{\left(\sum_{i=1}^{m} d_{1i} - \sum_{i=1}^{m} \widehat{e}_{1i} \right)^2}{\sum_{i=1}^{m} \widehat{V}(\widehat{e}_{1i})}$$

- This test statistic is distributed as χ_1^2 under the null hypothesis that the survival functions for the two groups are the same.
- The test generalizes readily to more than two groups.

▶ Consider, for example, Figure 10 which shows Kaplan-Meier estimates separately for released prisoners who received and did not receive financial aid.

- At all times in the study, the estimated probability of not (yet) re-offending is greater in the financial aid group than in the no-aid group.
- The log-rank test statistic is $Q = 3.84$, which is associated with a p-value of almost exactly $.05$, providing marginally significant evidence for a difference between the two groups.

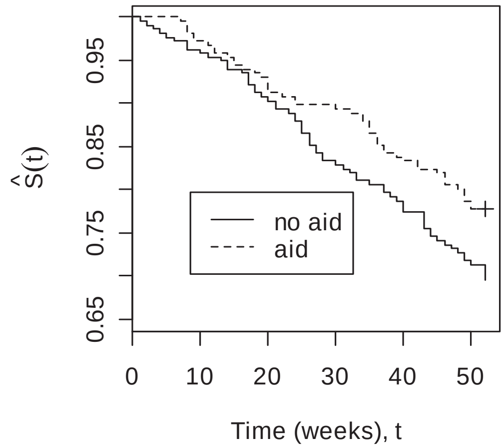

Figure 10. Kaplan-Meier estimates for released prisoners receiving financial aid and for those receiving no aid.

6. Cox Proportional-Hazards Regression

- Most interesting survival-analysis research examines the relationship between survival — typically in the form of the hazard function — and one or more explanatory variables (or *covariates*).
- Most common are linear-like models for the log hazard.
  - For example, a parametric regression model based on the exponential distribution:
$$\log_e h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$
or, equivalently,
$$\begin{aligned} h_i(t) &= \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \\ &= e^{\alpha} \times e^{\beta_1 x_{i1}} \times e^{\beta_2 x_{i2}} \times \cdots \times e^{\beta_k x_{ik}} \end{aligned}$$
where
    – i indexes subjects;
    – $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the values of the covariates for the ith subject.

- This is therefore a linear model for the log-hazard or a multiplicative model for the hazard itself.
- The model is *parametric* because, once the regression parameters $\alpha, \beta_1, \ldots, \beta_k$ are specified, the hazard function $h_i(t)$ is fully characterized by the model.
- The regression constant α represents a kind of *baseline hazard*, since $\log_e h_i(t) = \alpha$, or equivalently, $h_i(t) = e^{\alpha}$, when all of the x's are 0.
- Other parametric hazard regression models are based on other distributions commonly used in modeling survival data, such as the Gompertz and Weibull distributions.
- Parametric hazard models can be estimated with the `survreg` function in the `survival` package.

▶ Fully parametric hazard regression models have largely been superseded by the Cox model (introduced by David Cox in 1972), which leaves the baseline hazard function $\alpha(t) = \log_e h_0(t)$ unspecified:

$$\log_e h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

or equivalently,

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

- The Cox model is termed *semi-parametric* because while the baseline hazard can take any form, the covariates enter the model through the *linear predictor*

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

  – Notice that there is no constant term (intercept) in the linear predictor: The constant is absorbed in the baseline hazard.

- The Cox regression model is a *proportional-hazards model*:
  - Consider two observations, i and i', that differ in their x-values, with respective linear predictors
    $$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$
    and
    $$\eta_{i'} = \beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_k x_{i'k}$$
  - The hazard ratio for these two observations is
    $$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)e^{\eta_i}}{h_0(t)e^{\eta_{i'}}} = \frac{e^{\eta_i}}{e^{\eta_{i'}}} = e^{\eta_i - \eta_{i'}}$$
  - This ratio is constant over time.
- In this initial formulation, I am assuming that the values of the covariates x_{ij} are constant over time.
- As we will see later, the Cox model can easily accommodate *time-dependent covariates* as well.

6.1 Partial-Likelihood Estimation of the Cox Model

▶ In the same remarkable paper in which he introduced the semi-parametric proportional-hazards regression model, Cox also invented a method, which he termed *partial likelihood*, to estimate the model.

- Partial-likelihood estimates are not as efficient as maximum-likelihood estimates for a correctly specified parametric hazard regression model.
- But not having to assume a possibly incorrect form for the baseline hazard more than makes up for small inefficiencies in estimation.
- Having estimated a Cox model, it is possible to recover a nonparametric estimate of the baseline hazard function.

- As is generally the case, to estimate a hazard regression model by maximum likelihood we have to write down the probability (or probability density) of the data as a function of the parameters of the model.
  - To keep things simple, I'll assume that the only form of censoring is right-censoring and that there are no tied event times in the data.
    - Neither restriction is an intrinsic limitation of the Cox model.
  - Let $p(t|\mathbf{x}, \boldsymbol{\beta})$ represent the probability density for an event at time t given the values of the covariates $\mathbf{x}$ and regression parameters $\boldsymbol{\beta}$.
    - Note that $\mathbf{x}$ and $\boldsymbol{\beta}$ are vectors.
  - A subject i for whom an event is observed at time t_i contributes $p(t_i|\mathbf{x}_i, \boldsymbol{\beta})$ to the likelihood.
  - For a subject i who is censored at time t_i, all we know is that the subject survived to that time, and therefore the observation contributes $S(t_i|\mathbf{x}_i, \boldsymbol{\beta})$ to the likelihood.
    - Recall that the survival function $S(t)$ gives the probability of surviving to time t.

- It is convenient to introduce the *censoring indicator variable* c_i, which is set to 1 if the event time for the ith subject is observed and to 0 if it is censored.
- Then the likelihood function for the data can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [p(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{c_i} [S(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{1-c_i}$$

- From our previous work, we know that $h(t) = p(t)/S(t)$, and so $p(t) = h(t)S(t)$.
- Using this fact:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} [h(t_i|\mathbf{x}_i, \boldsymbol{\beta})S(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{c_i} [S(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{1-c_i} \\ &= \prod_{i=1}^{n} [h(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{c_i} S(t_i|\mathbf{x}_i, \boldsymbol{\beta}) \end{aligned}$$

- According to the Cox model,

$$\begin{aligned} h(t_i|\mathbf{x}_i, \boldsymbol{\beta}) &= h_0(t_i) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \\ &= h_0(t_i) \exp(\mathbf{x}_i' \boldsymbol{\beta}) \end{aligned}$$

- Similarly, we can show that

$$S(t_i|\mathbf{x}_i, \boldsymbol{\beta}) = S_0(t_i)^{\exp(\mathbf{x}_i' \boldsymbol{\beta})}$$

where $S_0(t)$ is the *baseline survival function*.

- Substituting these values from the Cox model into the formula for the likelihood produces

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[h_0(t_i) \exp(\mathbf{x}_i' \boldsymbol{\beta})\right]^{c_i} S_0(t_i)^{\exp(\mathbf{x}_i' \boldsymbol{\beta})}$$

- Full maximum-likelihood estimates would find the values of the parameters $\boldsymbol{\beta}$ that, along with the baseline hazard and survival functions, maximize this likelihood, but the problem is not tractable.

▶ Cox's proposal was instead to maximize what he termed the *partial-likelihood function*

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{i' \epsilon R(t_i)} \exp(\mathbf{x}_{i'}' \boldsymbol{\beta})} \right]^{c_i}$$

- The *risk set* $R(t_i)$ includes those subjects at risk for the event at time t_i, when the event was observed to occur for subject i (or at which time subject i was censored) — that is, subjects for whom the event has not yet occurred or who have yet to be censored.
- Censoring times are effectively excluded from the likelihood because for these observations the exponent $c_i = 0$.

- Thus we can re-express the partial likelihood as

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{i' \epsilon R(t_i)} \exp(\mathbf{x}_{i'}'\boldsymbol{\beta})}$$

where i now indexes the m observed event times, $t_1, t_2, \ldots, t_m$.

- The ratio

$$\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{i' \epsilon R(t_i)} \exp(\mathbf{x}_{i'}'\boldsymbol{\beta})}$$

has an intuitive interpretation:

  – According to the Cox model, the hazard for subject i, for whom the event was actually observed to occur at time t_i, is proportional to $\exp(\mathbf{x}_i'\boldsymbol{\beta})$.
  – The ratio, therefore, expresses the hazard for subject i relative to the cumulative hazard for all subjects at risk at the time that the event occurred to subject i.
  – We want values of the parameters that will predict that the hazard was high for subjects at the times that events actually were observed to occur to them.

- Although they are not fully efficient, maximum partial-likelihood estimates share the other general properties of maximum-likelihood estimates.
  - We can obtain estimated asymptotic sampling variances from the inverse of the information matrix.
  - We can perform likelihood ratio, Wald, and score tests for the regression coefficients.
- When there are many ties in the data, the computation of maximum partial-likelihood estimates, though still feasible, becomes time-consuming.
  - For this reason, approximations to the partial likelihood function are often used.
  - Two commonly employed approximations are due to Breslow and to Efron.

  - Breslow's method is more popular, but Efron's approximation is generally the more accurate of the two (and is the default for the `coxph` function in the `survival` package in R).

6.2 An Illustration Using Rossi et al.'s Recidivism Data

▶ Recall Rossi et al.'s data on recidivism of 432 prisoners during the first year after their release from Maryland state prisons.
- Survival time is the number of weeks to first arrest for each former prisoner.
- Because former inmates were followed for one year after release, those who were not rearrested during this period were censored at 52 weeks.

- The Cox regression reported below uses the following time-constant covariates:
  – `fin`: A dummy variable coded `1` if the former prisoner received financial aid after release from prison and `0` otherwise.
  – `age`: The former prisoner's age in years at the time of release.
  – `race`: A dummy variable coded `1` for blacks and `0` for others.
  – `wexp`: Work experience, a dummy variable coded `1` if the former prisoner had full-time work experience prior to going to prison and `0` otherwise.
  – `mar`: Marital status, a dummy variable coded `1` if the former prisoner was married at the time of release and `0` otherwise.
  – `paro`: A dummy variable coded `1` if the former prisoner was released on parole and `0` otherwise.
  – `prio`: The number of prior incarcerations.

- The results of the Cox regression for time to first arrest are as follows:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
`fin`	-0.379	0.684	0.191	-1.983	$.047$
`age`	-0.057	0.944	0.022	-2.611	$.009$
`race`	0.314	1.369	0.308	1.019	$.310$
`wexp`	-0.150	0.861	0.212	-0.706	$.480$
`mar`	-0.434	0.648	0.382	-1.136	$.260$
`paro`	-0.085	0.919	0.196	-0.434	$.660$
`prio`	0.091	1.096	0.029	3.195	$.001$

where:

- b_j is the maximum partial-likelihood estimate of β_j in the Cox model.
- e^{b_j}, the exponentiated coefficient, gives the effect of x_j in the multiplicative form of the model — more about this shortly.
- SE(b_j) is the standard error of b_j, that is the square-root of the corresponding diagonal entry of the estimated asymptotic coefficient-covariance matrix.

- $z_j = b_j/\text{SE}(b_j)$ is the Wald statistic for testing the null hypothesis H_0: $\beta_j = 0$; under this null hypothesis, z_j follows an asymptotic standard-normal distribution.
- p_j is the two-sided p-value for the null hypothesis H_0: $\beta_j = 0$.
  - Thus, the coefficients for `age` and `prio` are highly statistically significant, while that for `fin` is marginally so.

- The estimated coefficients b_j of the Cox model give the linear, additive effects of the covariates on the log-hazard scale.
  - Although the *signs* of the coefficients are interpretable (e.g., other covariates held constant, getting financial aid decreases the hazard of rearrest, while an additional prior incarceration increases the hazard), the *magnitudes* of the coefficients are not so easily interpreted.

- It is more straightforward to interpret the exponentiated coefficients, which appear in the multiplicative form of the model,

$$\widehat{h}_i(t) = \widehat{h}_0(t) \times e^{b_1 x_{i1}} \times e^{b_2 x_{i2}} \times \cdots \times e^{b_k x_{ik}}$$

  - Thus, increasing x_j by 1, holding the other x's constant, multiplies the estimated hazard by e^{bj}.
  - For example, for the dummy-regressor `fin`, $e^{b_1} = e^{-0.379} = 0.684$, and so we estimate that providing financial aid *reduces* the hazard of rearrest — other covariates held constant — by a factor of 0.684 — that is, by $100(1 - 0.684) = 31.6$ percent.
  - Similarly, an additional prior conviction *increases* the estimated hazard of rearrest by a factor of $e^{b_7} = e^{0.091} = 1.096$ or $100(1.096 - 1) = 9.6$ percent.

7. Topics in Cox Regression

7.1 Time-Dependent Covariates

- It is often the case that the values of some explanatory variables in a survival analysis change over time.
  - For example, in Rossi et al.'s recidivism data, information on ex-inmates' employment status was collected on a weekly basis.
  - In contrast, other covariates in this data set, such as race and the provision of financial aid, are time-constant.
  - The Cox-regression model with time-dependent covariates takes the form

$$\begin{aligned} \log_e h_i(t) &= \alpha(t) + \beta_1 x_{i1}(t) + \beta_2 x_{i2}(t) + \cdots + \beta_k x_{ik}(t) \\ &= \alpha(t) + \mathbf{x}_i'(t)\boldsymbol{\beta} \end{aligned}$$

    – Of course, not *all* of the covariates have to vary with time: If covariate x_j is time-constant, then $x_{ij}(t) = x_{ij}$.

- Although the inclusion of time-dependent covariates can introduce non-trivial data-management issues (that is, the construction of a suitable data set can be tricky), the Cox-regression model can easily handle such covariates.
  - Both the data-management and conceptual treatment of time-dependent covariates are facilitated by the so-called "counting-process" representation of survival data.
  - We focus on each time interval for which data are available, recording the start time of the interval, the end time, whether or not the event of interest occurred during the interval, and the values of all covariates during the interval.
    - This approach naturally accommodates censoring, multiple periods of observation, late entry into the study, and time-varying data.
  - In the recidivism data, where we have weekly information for each ex-inmate, we create a separate data record for each week during which the subject was under observation.

- For example, the complete initial data record for the first subject is as follows:

```
> Rossi[1,]
  week arrest fin age race wexp mar paro prio educ emp1
1   20      1   0  27    1    0   0    1    3    3    0
  emp2 emp3 emp4 emp5 emp6 emp7 emp8 emp9 emp10 emp11
1    0    0    1    1    0    0    0    0     0     0
  emp12 emp13 emp14 emp15 emp16 emp17 emp18 emp19 emp20
1     0     0     0     0     0     0     0     0     0
  emp21 emp22 emp23 emp24 emp25 emp26 emp27 emp28 emp29
1    NA    NA    NA    NA    NA    NA    NA    NA    NA
  emp30 emp31 emp32 emp33 emp34 emp35 emp36 emp37 emp38
1    NA    NA    NA    NA    NA    NA    NA    NA    NA
  emp39 emp40 emp41 emp42 emp43 emp44 emp45 emp46 emp47
1    NA    NA    NA    NA    NA    NA    NA    NA    NA
  emp48 emp49 emp50 emp51 emp52
1    NA    NA    NA    NA    NA
```


– Subject 1 was rearrested in week 20, and consequently the employment dummy variables are available only for weeks 1 through 20 (and missing thereafter).
  · The subject had a job at weeks 4 and 5, but was otherwise unemployed when he was under observation.
  · *Note*: Actually, this subject was unemployed at *all* 20 weeks prior to his arrest — I altered the data for the purpose of this example.

– We therefore have to create 20 records for subject 1:

```
      start stop arrest fin age race . . . educ emp
1.1       0    1      0   0  27    1 . . .    3   0
1.2       1    2      0   0  27    1 . . .    3   0
1.3       2    3      0   0  27    1 . . .    3   0
1.4       3    4      0   0  27    1 . . .    3   1
1.5       4    5      0   0  27    1 . . .    3   1
1.6       5    6      0   0  27    1 . . .    3   0
. . .
1.18     17   18      0   0  27    1 . . .    3   0
1.19     18   19      0   0  27    1 . . .    3   0
1.20     19   20      1   0  27    1 . . .    3   0
```

– The subject $\times$ time-period data set has many more records than the original data set: 19,809 vs. 432.

- Except for missing data after subjects were rearrested, the data-collection schedule in the recidivism study is regular.
  - That is, all subjects are observed on a weekly basis, starting in week 1 of the study and ending (if the subject is not arrested during the period of the study) in week 52.
  - The counting-process approach, however, does not require regular observation periods.

▶ Once the subject × time-period data set is constructed, it is a simple matter to fit the Cox model to the data.

- In maximizing the partial likelihood, we simply need to know the risk set at each event time, and the contemporaneous values of the covariates for the subjects in the risk set.
- This information is available in the subject × time-period data.

▶ For Rossi et al.'s recidivism data, we get the following results when the time-dependent covariate employment status is included in the model:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
`fin`	-0.357	0.700	0.191	-1.866	$.062$
`age`	-0.046	0.955	0.022	-2.132	$.033$
`race`	0.339	1.403	0.310	1.094	$.270$
`wexp`	-0.026	0.975	0.211	-0.121	$.900$
`mar`	-0.294	0.745	0.383	-0.767	$.440$
`paro`	-0.064	0.938	0.195	-0.330	$.740$
`prio`	0.085	1.089	0.029	2.940	$.003$
`emp`	-1.328	0.265	0.251	-5.30	$\ll .001$

- The time-dependent employment covariate has a very large apparent effect: $e^{-1.328} = 0.265$.
  – That is, other factors held constant, the hazard of rearrest is 73.5 percent lower during a week in which an ex-inmate is employed
- As Allison points out, however, the direction of causality here is ambiguous, since a subject cannot work when he is in jail.

7.1.1 Lagged Covariates

▶ One way to address this kind of problem is to used a *lagged covariate*.
- Rather than using the contemporaneous value of the employment dummy variable, we can instead use the value from the previous week.
  – When we lag employment one week, we lose the observation for each subject for the first week.

– For example, person 1's subject × time-period data records are as follows:

```
     start stop arrest fin age race . . . educ emp
1.2      1    2      0   0  27    1 . . .    3   0
1.3      2    3      0   0  27    1 . . .    3   0
1.4      3    4      0   0  27    1 . . .    3   0
1.5      4    5      0   0  27    1 . . .    3   1
1.6      5    6      0   0  27    1 . . .    3   1
. . .
1.18    17   18      0   0  27    1 . . .    3   0
1.19    18   19      0   0  27    1 . . .    3   0
1.20    19   20      1   0  27    1 . . .    3   0
```

– Recall that subject 1 was (supposedly) employed at weeks 4 and 5.

- With employment lagged one week, a Cox regression for the recidivism data produces the following results:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
fin	-0.351	0.704	0.192	-1.831	$.067$
age	-0.050	0.951	0.022	-2.774	$.023$
race	0.321	1.379	0.309	1.040	$.300$
wexp	-0.048	0.953	0.213	-0.223	$.820$
mar	-0.345	0.708	0.383	-0.900	$.370$
paro	-0.471	0.954	0.196	-0.240	$.810$
prio	0.092	1.096	0.029	3.195	$.001$
emp (lagged)	-0.787	0.455	0.218	-3.608	$< .001$

  - The coefficient for the time-varying covariate employment is still large and statistically significant, but much smaller than before: $e^{-0.787} = 0.455$, and so the hazard of arrest is 54.5 percent lower following a week in which an ex-inmate is employed.

7.2 Cox-Regression Diagnostics

▶ As for a linear or generalized-linear model, it is important to determine whether a fitted Cox-regression model adequately represents the data.

▶ I will consider diagnostics for three kinds of problems, along with possible solutions:
- violation of the assumption of proportional hazards;
- influential data;
- nonlinearity in the relationship between the log-hazard and the covariates.

7.2.1 Checking for Non-Proportional Hazards

▶ A departure from proportional hazards occurs when regression coefficients are dependent on time — that is, when time *interacts* with one or more covariates.

▶ Tests and graphical diagnostics for interactions between covariates and time may be based on the *scaled Schoenfeld residuals* from the Cox model.
- The formula and rationale for the scaled Schoenfeld residuals are complicated, and so I won't give them here (but see Hosmer and Lemeshow, 1999, or Therneau and Grambsch, 2000).
- The scaled Schoenfeld residuals comprise a matrix, with one row for each record in the data set to which the model was fit and one column for each covariate.

- Plotting scaled Schoenfeld residuals against time, or a suitable transformation of time, reveals unmodelled interactions between covariates and time.
  - One choice is to use the Kaplan-Meier estimate of the survival function to transform time.
  - A systematic tendency of the scaled Schoenfeld residuals to rise or fall more or less linearly with (transformed) time suggests entering a linear-by-linear interaction (i.e., the simple product) between the covariate and time into the model.
- A test for non-proportional hazards can be based on the estimated correlation between the scaled Schoenfeld residuals and (transformed) time.
  - This test can be performed on a per-covariate basis and also cumulated across covariates.

▶ For a simple illustration, I'll return to the original Cox regression that I performed for Rossi's recidivism data.

- It is conceivable that a variable with a nonsignificant coefficient in the initial model nevertheless interacts significantly with time, so I'll start with the model as originally specified.
- Tests for non-proportional hazards in this model are as follows:

Covariate	$\widehat{\rho}$	X^2	df	p
`fin`	$.006$	0.005	1	$.943$
`age`	$-.265$	11.279	1	$< .001$
`race`	$-.112$	1.417	1	$.234$
`wexp`	$.230$	7.140	1	$.007$
`mar`	$.073$	0.686	1	$.407$
`paro`	$-.036$	0.155	1	$.694$
`prio`	$-.014$	0.023	1	$.879$
Global Test		17.659	7	$.014$

– $\widehat{\rho}$ is the estimated correlation between the scaled Schoenfeld residuals and transformed time.
– Under the null hypothesis of proportional hazards, each X^2 test statistic is distributed as χ^2 with the indicated degrees of freedom.
– Thus, the tests for age (which had a statistically significant coefficient in the initial Cox regression) and wexp (which did not) are statistically significant, as is the global test for non-proportional hazards.

- Figure 11 shows plots of scaled Schoenfeld residuals vs. the covariates age and wexp.
  – The line on each plot is a smoothing spline (a method of nonparametric regression); the broken lines give a point-wise 95-percent confidence envelope around this fit.
  – The tendency for the effect of age to fall with time, and for that of wexp to rise with time, is clear in these plots.

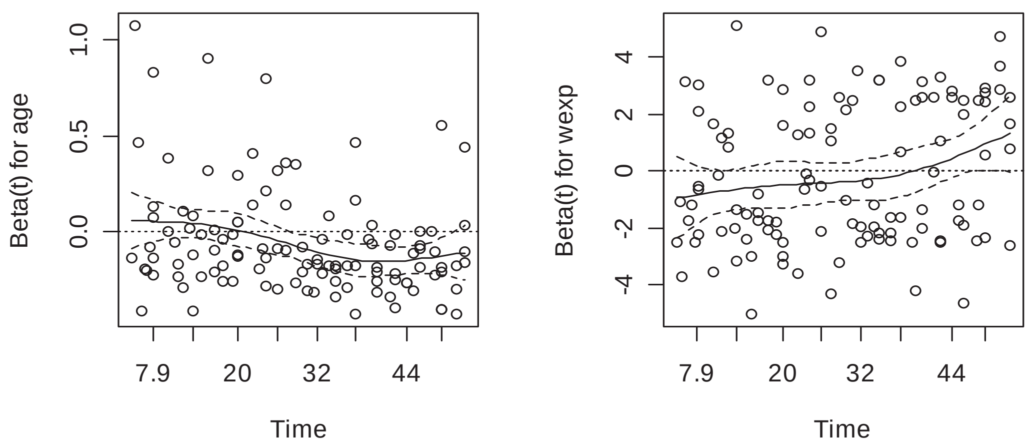

Figure 11. Plots of scaled Schoenfeld residuals against transformed time for the covariates age and wexp.

- I fit a respecified Cox regression model to the data, including `fin`, `age`, `wexp`, and `prio`, along with the products of time and `age` and of time and `wexp`:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
`fin`	-0.3575	0.699	0.1902	-1.88	$.060$
`age`	0.0778	1.081	0.0400	1.95	$.052$
`wexp`	-1.4644	0.231	0.4777	3.07	$.002$
`prio`	0.0876	1.092	0.0282	3.10	$.002$
`age` $\times$ time	-0.0053	0.995	0.0015	-3.47	$< .001$
`wexp` $\times$ time	0.0439	1.045	0.0145	3.02	$.003$

  – The products of time and `age` and of time and `wexp` are time-dependent covariates, and so the model must be fit to the subject $\times$ time-period form of the data set.

  – Although interactions with time appear in the model, time itself does not appear as a covariate: The “main effect” of time is in the baseline hazard function.
  – The effect of `age` on the hazard of re-offending is initially positive, but this effect declines with time and eventually becomes negative (by 15 weeks).
  – The effect of `wexp` is initially strongly negative, but eventually becomes positive (by 34 weeks).
  – The respecified model shows no evidence of non-proportional hazards: For example, the global test gives $X^2 = 1.12$, $df = 6$, $p = .98$.

7.2.2 Fitting by Strata

- An alternative to incorporating interactions with time is to divide the data into *strata* based on the values of one or more covariates.
  - Each stratum may have a different baseline hazard function, but the regression coefficients in the Cox model are assumed to be constant across strata.
- An advantage of this approach is that we do not have to assume a particular form of interaction between the stratifying covariates and time.
- There are a couple of disadvantages, however:
  - The stratifying covariates disappear from the linear predictor into the baseline hazard functions.
    - Stratification is therefore most attractive when we are not really interested in the effects of the stratifying covariates, but wish simply to control for them.

  - When the stratifying covariates take on many different (combinations of) values, stratification — which divides the data into groups — is not practical.
    - We can, however, recode a stratifying variable into a small number of relatively homogeneous categories.
- For the example, I divided age into three categories: those 20 years old or less; those 21 to 25 years old; and those older than 25.
  - Cross-classifying categorized age and work experience produces the following contingency table:

	Age		
Work Experience	20 or less	21 – 25	26 or more
No	87	73	25
Yes	40	102	105

- Fitting the stratified Cox model to the data:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
`fin`	-0.387	0.679	0.192	-2.02	$.043$
`prio`	0.080	1.084	0.028	2.83	$.005$

- For this model, as well, there is no evidence of non-proportional hazards: The global test statistic is $X^2 = 0.15$ with $2\ df$, for which $p = .93$.

7.2.3 Detecting Influential Observations

▶ As in linear and generalized linear models, we don't want the results in Cox regression to depend unduly on one or a small number of observations.

▶ Approximations to changes in the Cox regression coefficients attendant on deleting individual observations (*dfbeta*), and these changes standardized by coefficient standard errors (*dfbetas*), can be obtained for the Cox model.

▶ Figure 12 shows index plots of *dfbeta* and *dfbetas* for the two covariates, `fin` and `prio`, in the stratified Cox model that I fit to the recidivism data.
- All of the *dfbeta* are small relative to the sizes of the corresponding regression coefficients, and the *dfbetas* are small as well.

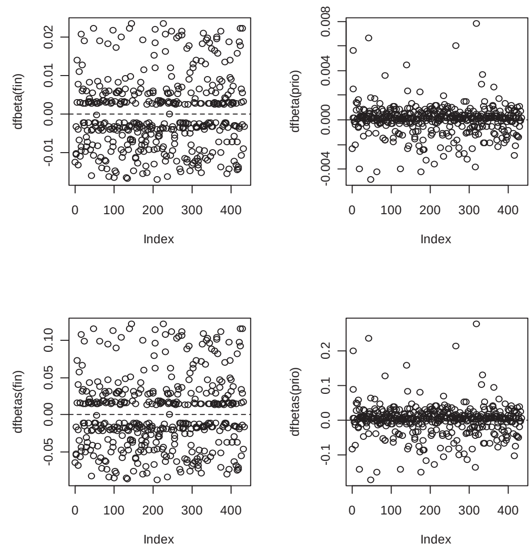

Figure 12. *dfbeta* (top row) and *dfbetas* (bottom) row for `fin` (left) and `prio` (right) in the stratified Cox model fit to Rossi et al.'s recidivism data.

7.2.4 Detecting Nonlinearity

- Another kind of Cox-model residuals, called *martingale residuals*, are useful for detecting nonlinearity in Cox regression.
  - As was the case for scaled Schoenfeld residuals, the details of the martingale residuals are beyond the level of this lecture.
  - Plotting residuals against covariates, in a manner analogous to plotting residuals against covariates from a linear model, can reveal nonlinearity in the partial relationship between the log hazard and the covariates.
- Let m_i represent the martingale residual for observation i.
  - Plotting $m_i + b_j x_{ij}$ against x_{ij} is analogous to a *component+residual* (or *partial-residual*) plot for a covariate in a linear model.
  - Here, b_j is the estimated coefficient for the jth covariate, and x_{ij} is the value of the jth covariate for observation i.

- As is the case for component+residual plots in a linear or generalized-linear model, it aids interpretation to add a nonparametric-regression smooth and a least-squares line to these plots.
- Plotting martingale residuals and partial residuals against `prio` in the last Cox regression produces the results shown in Figure 13.
  - The plots appear quite straight, suggesting that nonlinearity is not a problem here.
  - As is typical of residuals plots for survival data, the patterned nature of these plots makes smoothing important to their visual interpretation.
- There is no issue of nonlinearity in the partial relationship between the log-hazard and `fin` (the other covariate in the model), since `fin` is a dummy variable.

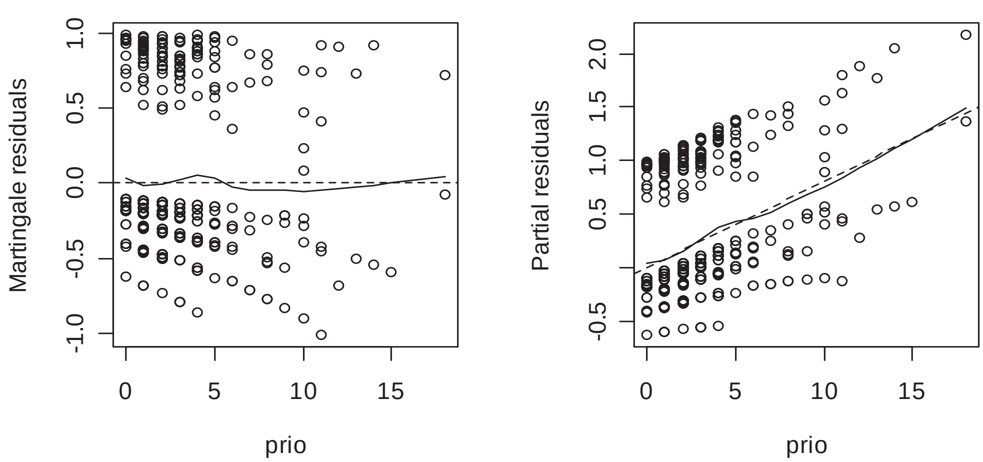

Figure 13. Plots of Martingale residuals (left) and partial residuals (right) against `prio`.

7.3 Estimating Survival in Cox Regression Models

- Because of the unspecified baseline hazard function, the estimated coefficients of the Cox model do not fully characterize the distribution of survival time as a function of the covariates.
- By an extension of the Kaplan-Meier method, it is possible, however, to estimate the survival function for a real or hypothetical subject with any combination of covariate values.
- In a stratified model, this approach produces an estimated survival curve for each stratum.

- To generate a simple example, suppose that a Cox model is fit to the recidivism data employing the time-constant covariates `fin`, `age`, and `prio`, producing the following results:

Covariate	b_j	e^{b_j}	SE(b_j)	z_j	p_j
`fin`	-0.347	0.707	0.190	-1.82	$.068$
`age`	-0.067	0.021	0.021	-3.22	$.001$
`prio`	0.097	0.027	0.027	3.56	$< .001$

  - Figure 14 shows the estimated survival functions for those receiving and not receiving financial aid (i.e., for `fin` $= 1$ and 0, respectively) at average age ($\overline{\texttt{age}} = 24.6$) and average prior number of arrests ($\overline{\texttt{prio}} = 2.98$).
  - Similarly, an estimate of the baseline survival function can be recovered by setting all of the covariates to 0.

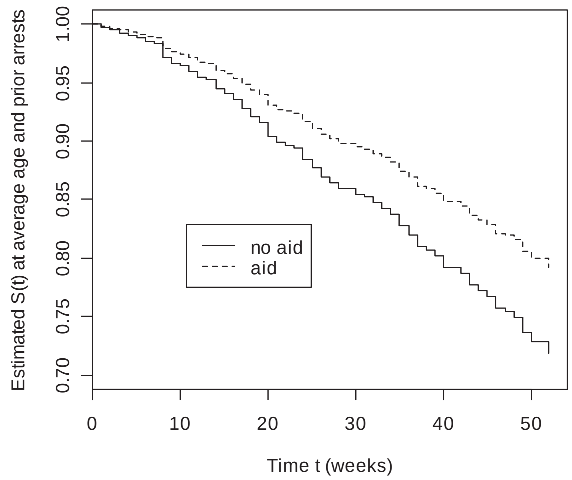

Figure 14. Estimated survival functions from the Cox regression model with `fin`, `age`, and `prio` as predictors — setting `age` and `prio` to their average values, and letting `fin` take on the values 0 and 1.