Survival Analysis

```
library(survival)
library(survminer)
library(ggplot2)
library(KMsurv)
```

Goals

- To introduce the format and structure of Survival Analysis with different kinds of Survival Models.
- To describe diagnostics for Survival Analysis.
- To introduce Cox Regression.

Censoring

Introduction

Survival analysis encompasses a wide variety of methods for analyzing the timing of events. Interest lies in succinctly describing whether and/or when events occur.

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, time to relapse to smoking or other substance use, and time to end of remission in leukemia.

Survival analysis has 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.

Studies that are amenable to a survival analysis share several common methodological features:

- A well-defined target event under study

Event occurrence represents an individual's transition from one well-defined state to another. These states must be defined precisely. For example, a recently treated ex-alcoholic is (a) abstinent or (b) starts drinking, a leukemia patient is (a) in remission or (b) no longer in remission, or a heart transplant patient is (a) alive or (b) dead.

- A well-defined starting point for the study at which time no one in the study has experienced the target event

The "beginning of time" is a moment when everyone in the population occupies one, and only one, of the possible states. For example, the day after surgery, a surviving heart transplant patient is alive.

- A meaningful metric for recording time

Time should be measured in well-defined units: days, months, years

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 problematic.

(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.

Censoring

The two major reasons for censoring are:

- Some individuals will never experience the event at all
- Some individuals will experience the event, but not during data collection

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 hypothetical study:

Suppose heart-lung transplant patients 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.

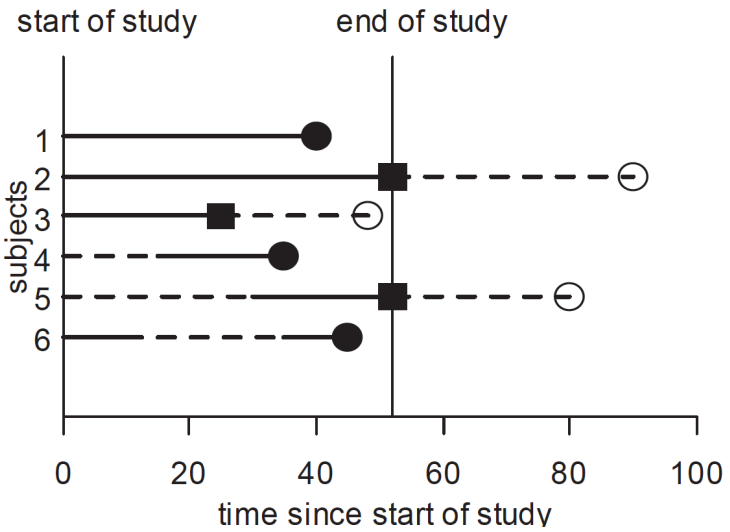


Figure 1: Survival histories of six subjects illustrating several kinds of censoring (as well as uncensored data).

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.

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.

Subject 2 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 and occurs because the event of interest is not observed during the study.

Left censoring occurs because the beginning of time is not known for the individual. For example, **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.

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.

Methods of survival analysis will treat as "at risk" for an event 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. That is, the distribution of survival times of subjects who are censored at a particular time is no different from that of subjects who are still under observation at this time.

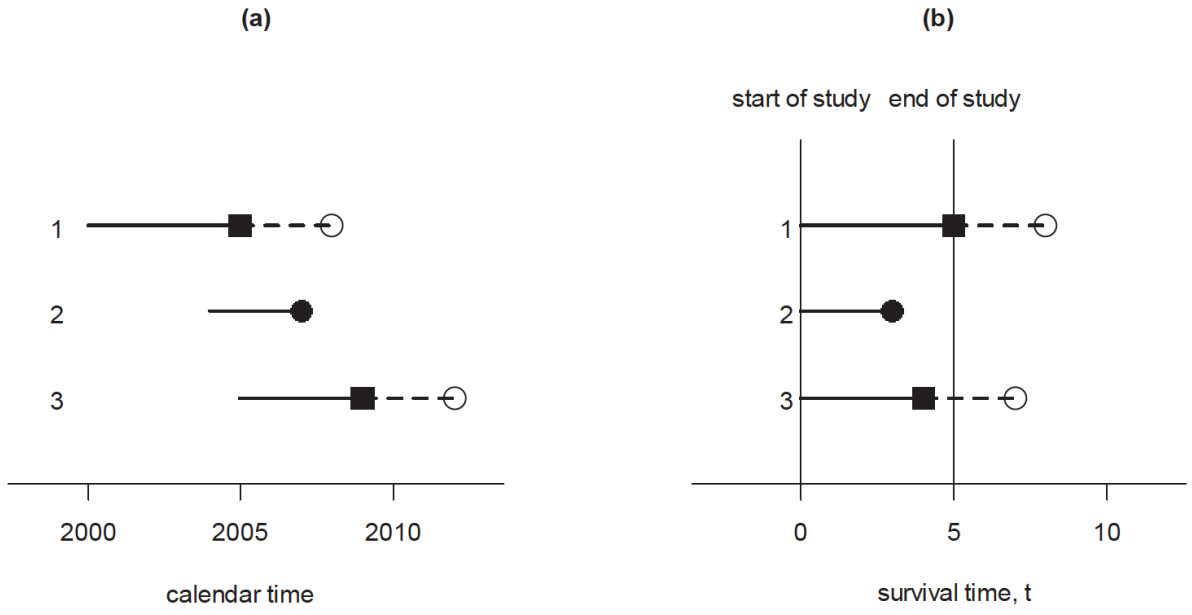


Figure 2: Survival times of three patients who are followed for at most 5 years after bypass surgery.

Noninformative censoring occurs for reasons independent of the event occurrence or the risk of event occurrence. Fixed censoring is noninformative.

Informative censoring occurs when non-observation of the individual is due to occurrence or imminent occurrence of the event of interest.

With random censoring, it is quite possible that survival time is not independent of the censoring mechanism.

- For example, if someone is about to relapse into alcoholism and stops answering the experimenter's calls during the period of data gathering as a result, the censoring is informative.
- Or 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.

When 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.

Survival Probability and Life Tables

Survival Probability

I will introduce the survival probability in the context of discrete time first before extending the concept to continuous time.

Discrete time occurs when events are recorded to the nearest week, month, or year. In other words, we do not have an exact date on which the event occurred. This type of data is sometimes called *interval-censored* because we only know that an event occurred at some point during an interval of time.

Let T be the survival time, a discrete random variable, that takes the values $t_1 < t_2 < \ldots$ with probabilities denoted

$$f(t_j) = Pr(T = t_j)$$

The *survival probability* at time t_j is the probability that the survival time T is at least t_j:

$$S(t_j) = Pr(T \geq t_j) = \sum_{k=j}^{\infty} f(t_j)$$

If there is not censoring, the survival probability could easily be estimated by the observed proportion of subjects surviving time point t.

But when there is censoring, we need the number of *at-risk* subjects at time t, which is the number alive under observation and not censored. So to estimate the survival probability, we can divide the span of the study into a series of intervals and consider the conditional probability of surviving given being at-risk at the beginning of each interval.

The simplest estimate is known as the life table or actuarial estimate.

Life Tables

The life table estimate is useful when the data are grouped; that is, discrete time intervals. We do not know exactly when during each time interval an event occurs.

Three steps to a life table analysis:

1. Divide the times into fixed disjoint intervals.
2. Estimate the conditional proability of survival across each interval
3. Estimate $S(t_j)$ at the interval endpoint.

Intervals:

$$I_j[a_{j-1}, a_j), j = 1, 2, ..., k+1 \text{ such that } a_0 = 0 \text{ and } a_{k+1} = \infty$$

For jth interval:

N_j is the number at risk right before the beginning of I_j

D_j is the number of events that occur in I_j

W_j is the number censored in I_j

Two questions arise in the calculation in Step 2:

How do those at risk in the $(j-1)$th interval propagate into the jth interval?

How to estimate the survival probability across I_j, given being alive at the start of I_j?

Example: The following table gives the survival times for 913 patients with malignant melanoma who were treated at the M.D. Anderson Tumor Clinic between 1944 and 1960. Survival time is defined as the time (in years) from treatment to the death of the patient.

j	I_j interval(yrs)	N_j at risk	D_j events	W_j censored	N'_j Eff # at risk	$\tilde{q}$	$\tilde{p} = 1 - \tilde{q}$	$\tilde{S}(t)$
	0	913						1
1	[0, 1)	913	312	96	865.0	0.361	0.639	0.639
2	[1, 2)	505	96	74	468.0	0.205	0.795	0.508
3	[2, 3)	335	45	62	?	0.148	0.852	0.433
4	[3, 4)	228	29	30	213.0	?	0.864	0.374
5	[4, 5)	169	7	40	149.0	0.047	0.953	?
6	[5, 6)	122	9	37	103.5	0.087	0.913	0.325
7	[6, 7)	76	3	17	67.5	0.044	0.956	0.311
8	[7, 8)	56	1	12	50.0	0.020	0.980	0.305
9	[8, 9)	43	3	8	39.0	0.077	0.923	0.281
10	[9, ∞)	32	32	-	32.0	1.000	0.000	0.000

Basically, we do not know when the censored individuals left during the interval, so we assume that they left uniformally throughout the interval.

$$N'_j = N_j - \frac{W_j}{2}$$

Thus, $N'_3 = 335 - 62/2 = 304$.

Estimate $S(t_j)$:

The event cannot have occurred by duration 0 so $S(0) = 1$.

$\widetilde{q}_j$ is the year-specific mortality rate — that is, the proportion of individuals in year j who die during that year. $\widetilde{q}_j$ is the complement of $\widetilde{p}_j$; that is, $\widetilde{q}_j = 1 - \widetilde{p}_j$ where $\widetilde{p}_j$ is the proportion of individuals in year j who survive to the $(j+1)$st year — that is, the conditional probability of surviving to year $j+1$ given that one has made it to year j.

So $\widetilde{q}_4 = 1 - \widetilde{p}_4 = 1 - .864 = 0.136$

Alternatively, $\widetilde{q}_j$ is obtained by D_j/N'_j. So $\widetilde{q}_4 = 29/213 = 0.136$.

The survival probabilities are:

$$S(t_1) = \widetilde{p}_1$$

$$S(t_j) = S(t_{j-1}) \times \widetilde{p}_j$$

Thus, $S(t_5) = 0.374 \times 0.953 = 0.356$.

If we know the exact time at which an event occurred or censoring occurred, then we do not need to assume that individuals were censored uniformly throughout the interval and we can obtain better estimates using Kaplan-Meier.

Kaplan-Meier Estimates

The Kaplan-Meier (K-M) estimator is useful when the event and censored times are observed.

Suppose that we have N observations and that there are m unique event times arranged in ascending order, $t_1 < t_2 < \ldots < t_m$.

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_j represent the number of individuals at risk for the event at time j. 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_j represent the number of events observed at time j.

The conditional probability of surviving past time j given survival to that time is estimated by $(N_j - D_j)/N_j$.

The unconditional probability of surviving past any time j is estimated by

$$\widehat{S}(t) = \prod_{t_j \leq t} \frac{N_j - D_j}{N_j}$$

For the K-M estimate, censored observations are treated as observed up to and including the time of censoring.

Example: The remission times of 42 patients with acute leukemia were reported in a clinical trial undertaken to assess the ability of 6-mercaptopurine (6-MP) to maintain remission (i.e., to remain disease free). Each patient was randomized to receive 6-MP or a placebo. Patients were observed until they had a recurrence of the disease, they were censored, or until the study was terminated after one year. The variables in the dataset are:

`time` - the remission times in weeks;

`status` - 1 if patient had a recurrence of leukemia; 0 if censored

`group` - 1 if patient is in the 6-MP group; 0 if in the placebo group

```
leuk <- read.table("leukemia.dat", header = TRUE)
```

Using the `survival` package, there is a function called `Surv` that creates a survival object, which contains the follow-up time and the event indicator together. This object is used as the outcome in the `survfit()` function to obtain the K-M estimates.

```
with(leuk, Surv(time, status))
```

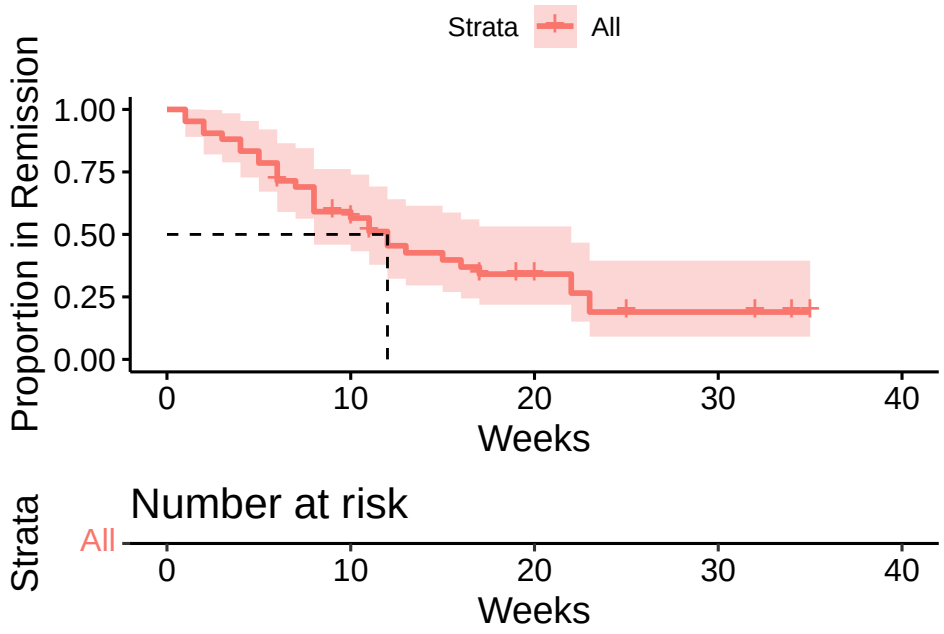
```
##  [1]  6   6   6   7  10  13  16  22  23   6+  9+ 10+ 11+ 17+ 19+ 20+ 25+ 32+ 32+
## [20] 34+ 35+  1   1   2   2   3   4   4   5   5   8   8   8   8  11  11  12  12
## [39] 15  17  22  23
```

```
KM <- survfit(Surv(time, status) ~ 1, data=leuk)
summary(KM)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = leuk)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     42       2    0.952  0.0329       0.8901        1.000
##     2     40       2    0.905  0.0453       0.8202        0.998
##     3     38       1    0.881  0.0500       0.7883        0.985
##     4     37       2    0.833  0.0575       0.7279        0.954
##     5     35       2    0.786  0.0633       0.6709        0.920
##     6     33       3    0.714  0.0697       0.5899        0.865
##     7     29       1    0.690  0.0715       0.5628        0.845
##     8     28       4    0.591  0.0764       0.4588        0.762
##    10     23       1    0.565  0.0773       0.4325        0.739
##    11     21       2    0.512  0.0788       0.3783        0.692
##    12     18       2    0.455  0.0796       0.3227        0.641
##    13     16       1    0.426  0.0795       0.2958        0.615
##    15     15       1    0.398  0.0791       0.2694        0.588
##    16     14       1    0.369  0.0784       0.2437        0.560
```

```
##     17      13       1    0.341  0.0774       0.2186       0.532
##     22       9       2    0.265  0.0765       0.1507       0.467
##     23       7       2    0.189  0.0710       0.0909       0.395
```

```
ggsurvplot(KM, surv.median.line = "hv", risk.table = TRUE, xlab = "Weeks", ylab= "Proportion in Remissi
```



The estimated survival curve jumps only at observed failure times, and the information from the censored observations contributes to the sizes of the steps.

As the number of subjects increases, the number of failure (event) times increases, so the sizes of the steps decreases, which yields a smoother curve.

Median Survival Time

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.

Median survival time, τ is estimated as $S(\tau) = 0.5$. In practice, we do not usually hit the median survival at exactly one of the failure times. In this case, the estimated median survival is the smallest time τ such that $S(\tau) \leq 0.5$.

```
KM
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = leuk)
##
##       n events median 0.95LCL 0.95UCL
## [1,] 42     30     12       8      22
```

The estimated median survival time is 12 weeks. Thus, at 12 weeks, half (i.e., 50%) of the sample had a recurrance of the disease and half are still in remission.

The pth quantile of survival time is $\tau_p = \min(\tau : \widehat{S}(\tau) \leq p)$ so for the median, $\tau_{.5} = \min(\tau : \widehat{S}(\tau) \leq 0.5)$.

For example:

```
quantile(KM)
```

```
## $quantile
## 25 50 75
##  6 12 23
##
## $lower
## 25 50 75
##  4  8 16
##
## $upper
## 25 50 75
## 10 22 NA
```

The 25th percentile of the distribution of survival time is 6 weeks. This means that by 6 weeks, 75% of the sample were still in remission, or alternatively, 25% had had a recurrance of the disease.

Greenwood's formula

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:

$$\widehat{V}[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{t_j \leq t} \frac{D_j}{N_j(N_j - D_j)}$$

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

Comparing two survival curves

There are two treatment groups: 6-MP and placebo. We can look at the estimated survival curves to compare effectiveness of treatment vs. control.

```
KM2 <- survfit(Surv(time, status) ~ group, data=leuk)
KM2
```

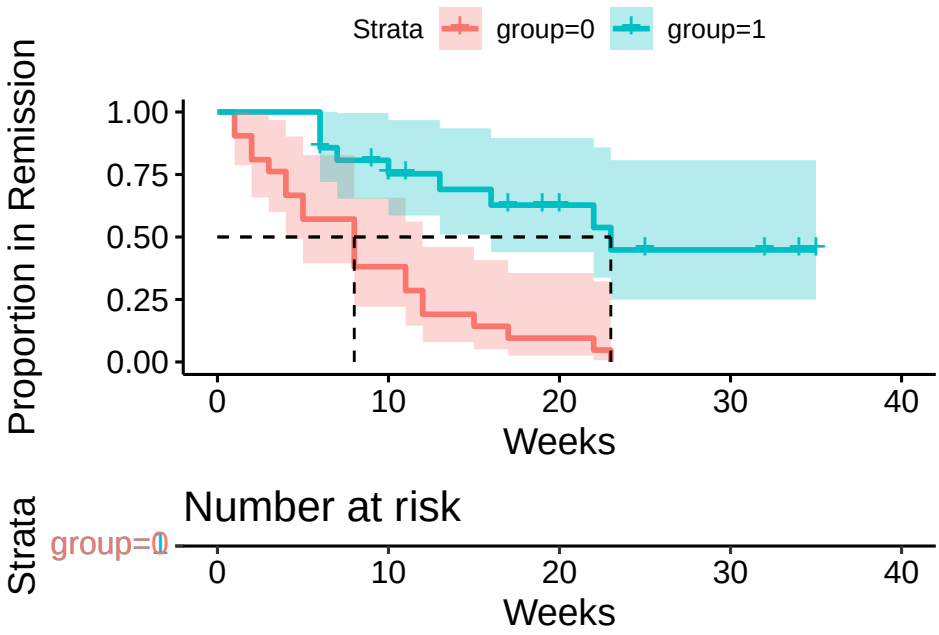
```
## Call: survfit(formula = Surv(time, status) ~ group, data = leuk)
##
##          n events median 0.95LCL 0.95UCL
## group=0 21     21      8       4      12
## group=1 21      9     23      16      NA
```

```
summary(KM2)
```

```
## Call: survfit(formula = Surv(time, status) ~ group, data = leuk)
##
##                 group=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.78754        1.000
##     2     19       2   0.8095  0.0857      0.65785        0.996
##     3     17       1   0.7619  0.0929      0.59988        0.968
##     4     16       2   0.6667  0.1029      0.49268        0.902
##     5     14       2   0.5714  0.1080      0.39455        0.828
##     8     12       4   0.3810  0.1060      0.22085        0.657
##    11      8       2   0.2857  0.0986      0.14529        0.562
```

```
##    12       6       2   0.1905  0.0857       0.07887         0.460
##    15       4       1   0.1429  0.0764       0.05011         0.407
##    17       3       1   0.0952  0.0641       0.02549         0.356
##    22       2       1   0.0476  0.0465       0.00703         0.322
##    23       1       1   0.0000     NaN            NA            NA
##
##                 group=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.720        1.000
##     7     17       1    0.807  0.0869        0.653        0.996
##    10     15       1    0.753  0.0963        0.586        0.968
##    13     12       1    0.690  0.1068        0.510        0.935
##    16     11       1    0.627  0.1141        0.439        0.896
##    22      7       1    0.538  0.1282        0.337        0.858
##    23      6       1    0.448  0.1346        0.249        0.807
```

```
ggsurvplot(KM2, surv.median.line = "hv", risk.table = TRUE, conf.int = TRUE, xlab = "Weeks", ylab= "Pro
```



Comparing Survival Curves: Log-rank Test

To test whether those in the 6-MP group survive longer, we can use the log-rank test. It compares the observed number of events in each group to what would be expected if the null hypothesis, that there is no difference between the groups, were true. The log-rank statistic is approximately distributed as a χ^2 test statistic.

```
survdiff(Surv(time, status) ~ group, data=leuk)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ group, data = leuk)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## group=0 21       21     10.7      9.77      16.8
## group=1 21        9     19.3      5.46      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

The log-rank test extends directly to more than two groups. For example, if you have a categorical variable with three categories represented by two dummy variables, `D1` and `D2`, then you could perform the log-rank test using code similar to the following:

```
survdiff(Surv(time,status) ~ D1 + D2, data=leuk)
```

However, if the log-rank χ^2 is statistically significant, it only tells you that at least one of the groups is significantly different from the reference category.

Non-parametric tests, such as the log-rank test, are particularly feasible when comparing survival functions across the levels of a factor. They are fairly robust, efficient, and usually simple/intuitive. However, as the number of factors of interest increases, non-parametric tests become difficult to conduct and interpret. Regression models, instead, are more flexible for exploring the relationship between survival and predictors.

In survival analysis, we are studying the distribution of time to event T, which can be described or characterized by not only the survival function but also the hazard function, which describes a different aspect of survival times.

Discrete Time Survivial Models

For discrete time, the *hazard* at time t_j is the conditional probability of event at that time given that one has not experienced the event up to that point:

$$\lambda(t_j) = Pr(T = t_j | T \geq t_j) = \frac{f(t_j)}{S(t_j)}$$

The survival probability at time t_j can be written in terms of the hazard at all prior times $t_1, \ldots, t_{j-1}$,

$$S(t_j) = (1 - \lambda(t_1))(1 - \lambda(t_2)) \ldots (1 - \lambda(t_{j-1})) = \prod_{i=1}^{j} 1 - \lambda(t_j)$$

so in order to survive to time t_j one must first survive t_1, then one must survive t_2 given that one survived t_1, and so on, finally surviving t_{j-1} given survival up to that point.

Finally, the *cumulative hazard* is

$$\Lambda(t_j) = \sum_{j:t_j \leq t} \lambda(t_j)$$

Because the hazard is a conditional probability, it is bounded by 0 and 1. Thus, we have the same situation as we did in modeling the probability for a binary response variable. Similar to the transformation to the log-odds for logistic regression, we can work with the conditional odds of the event at each time t_j given that the event has not occurred up to that point, using the following model:

$$\frac{\lambda(t_j|\mathbf{x}_i)}{1 - \lambda(t_j|\mathbf{x}_i)} = \frac{\lambda_0(t_j)}{1 - \lambda_0(t_j)} \exp(\mathbf{x}_i'\beta)$$

where $\lambda(t_j|\mathbf{x}_i)$ is the hazard at time t_j for an individual with covariate values, $\mathbf{x}_i$, $\lambda_0(t_j)$ is the baseline hazard at time t_j, and $\exp(\mathbf{x}_i'\beta)$ is the relative risk associated with covariate values $\mathbf{x}_i$.

Taking logs, we obtain a model for the logit of the hazard (i.e., conditional probability) of event occurrance at t_j given the event has not occurred up to that time,

$$\text{logit } \lambda(t_j|\mathbf{x}_i) = \alpha_j + \mathbf{x}_i'\beta$$

where $\alpha_j = \text{logit } \lambda_0(t_j)$ is the logit of the baseline hazard and $\mathbf{x}_i'\beta$ is the effect of the covariates on the logit of the hazard. The model essentially treats time as a discrete factor by introducing one parameter α_j for each possible time of event occurrance, t_j. Interpretation of the parameters, β, associated with the other covariates follows along the same lines as in logistic regression.

As before, we can get back to the hazard using the inverse transformation:

$$\lambda(t_j) = \frac{1}{1 + \exp[-(\alpha_j + \mathbf{x}_i'\beta)]}$$

In fact, we can fit the model using the `glm` function with `family=binomial` just as we fit logistic regression models a few weeks ago.

This model is an alternative version of the proportional hazards model for discrete time and assumes that the *conditional* odds of surviving t_j are proportional to a baseline odds.

Final points about the logit hazard model

By modeling the discrete time periods with dummy variables, the baseline function is guaranteed to essentially mimic the shape observed in the life table.

But we may want to make some assumptions about its shape so that we could estimate it using fewer parameters and therefore have a more parsimonous model. Moreover, the completely general specification, besides being rather unparsimonious, might capitalize excessively on chance variations around a parametric functional form, especially when sample size is small.

When the study involves many time periods, the number of dummy predictors will be excessively large, thereby reducing statistical power.

As with the typical logistic regression that we discussed several weeks ago, deviance residuals can be computed and examined.

Continuous predictor variables can be included. Interaction terms can be included.

Different nested models can be compared using the deviance, like we have discussed previously.

Complementary Log-Log Model

A few weeks ago when we discussed generalized linear models, I briefly mentioned the complementary log-log link.

Comparisons between the logit and complementary log-log functions

At low probability values, the logit and complementary log-log functions are virtually identical.

Although odds of 1 (probability of 0.5) correspond to a convenient and easily remembered value of 0 on the logit scale, the corresponding value on the complementary log-log scale is a not-so-memorable. It is -0.3665.

The logit link builds in a proportional odds assumption in the discrete-time model, while the complementary log-log function builds in a proportional hazards assumption.

We can model the survival probability as

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

where $S(t_j|\mathbf{x}_i)$ is the probability that an individual with covariate values $\mathbf{x}_i$ will survive up to time point t_j, and $S_0(t_j)$ is the baseline survival.

Because

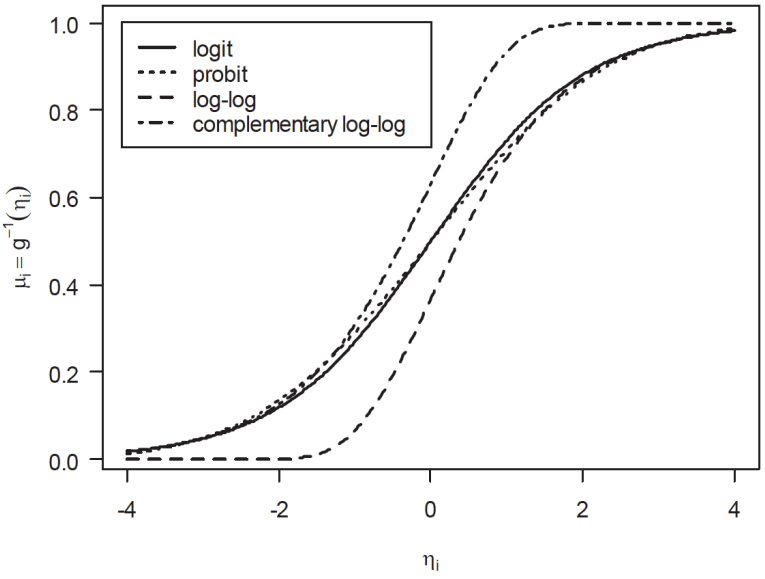


Figure 3: The logit function is symmetric around 0.5, while the complementary log-log function is not symmetric.

$$S(t_j) = (1 - \lambda(t_1)) \times (1 - \lambda(t_2)) \times \ldots \times (1 - \lambda(t_{j-1}))$$

we can obtain a similar relationship for the complement of the hazard,

$$1 - \lambda(t_j|\mathbf{x}_i) = [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\beta\}}$$

and solving for the hazard for individual i at time point t_j, we obtain

$$\lambda(t_j|\mathbf{x}_i) = 1 - [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\beta\}}$$

The transformation that makes the right hand side a linear function of the parameters is the complementary log-log link. Applying this transformation we obtain the model

$$\log(-\log(1 - \lambda(t_j|\mathbf{x}_i))) = \alpha_j + \mathbf{x}_i'\beta$$

where $\alpha_j = \log(-\log(1 - \lambda_0(t_j)))$ is the complementary log-log transformation of the baseline hazard.

Piecewise Exponential Model

Different kinds of proportional hazard models may be obtained by making different assumptions about the baseline survival function, or equivalently, the baseline hazard function.

If the baseline risk is constant over time, so $\lambda_0(t) = \lambda_0$, we obtain what is known as the exponential regression model:

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0 \exp\{\mathbf{x}_i'\beta\}$$

But the hazard need not be constant, and realistically, probably is not. For example, 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 piecewise exponential model, relaxes this assumption by assuming that the baseline hazard is constant within each interval. First, partition duration into J intervals with cutpoints $0 < \tau_0 < \tau_1 < \ldots < \tau_J < \infty$. Define the jth interval as $[\tau_{j-1}, \tau_j)$ extending from the $(j-1)$st boundary to the jth and including the former but not the latter.

We can then model the baseline hazard $\lambda_0(t)$ using J parameters, $\lambda_1, \ldots, \lambda_J$

$$\lambda_0(t) = \lambda_j \text{ for } t \text{ in } [\tau_{j-1}, \tau)$$

each representing the risk for the reference group in one particular interval. Because the risk is assumed to be piecewise constant, the corresponding survival function is often called a piecewise exponential.

The judicious choice of the cutpoints should allow us to approximate reasonably well almost any baseline hazard, using closely-spaced boundaries where the hazard varies rapidly and wider intervals where the hazard changes more slowly.

The piecewise exponential model is given as

$$\lambda_{ij} = \lambda_j \exp\{\mathbf{x}_i'\beta\}$$

where λ_{ij} is the hazard for individual i in interval j, λ_j is the baseline hazard for interval j, and $\exp\{\mathbf{x}_i'\beta\}$ is the relative risk for an individual with covariate values $\mathbf{x}_i$ compared to the baseline, at any given time.

Taking the logs, we obtain the additive log-linear model

$$\log \lambda_{ij} = \log t_{ij} + \log \lambda_j + \mathbf{x}_i'\beta$$

This is a standard log-linear model where the duration interval categories are treated as a factor. This model can be fit using Poisson regression that we discussed a few weeks ago with one difference. In this model there is an offset, $\log t_{ij}$, the log of the exposure time, which we actually know before fitting the model.

Why the offset? Recall the hazard is interpretable as the expected number of events per individual per unit of time. Thus, $\lambda_{ij} = \mu_{ij}/t_{ij}$ and $\log(\mu_{ij}/t_{ij}) = \log \mu_{ij} - \log t_{ij}$ and

$$\log \mu_{ij} - \log t_{ij} = \log \lambda_j + \mathbf{x}_i'\beta \log \mu_{ij} = \log t_{ij} + \log \lambda_j + \mathbf{x}_i'\beta$$

Cox Regression

Continuous Time Survival Models

In continuous survival analysis models, time is treated as inherently continuous. In the practical sense, of course, all time measurement is discrete. The issue is one of degree. In continuous survival models, the number of time points at which occurrence of the event of interest is assessed is large enough that *ties* are virtually impossible. This means that one and only one person failed at distinct time t_j

In discrete survival models, events are "collected" at various well defined intervals and counted, or the measurement times are small in number relative to the number of subjects and so ties are inevitable.

Survival and Hazard Functions

Let T be a non-negative random variable representing the time until the occurrence of an event. Assume that T is a *continuous* random variable with probability density function (p.d.f.), $f(t)$, and cumulative distribution function (c.d.f.), $F(t) = Pr(T < t)$, giving the probability that the event has occurred by duration t.

The complement of the c.d.f. is the *survival function*,

$$S(t) = Pr(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x)dx$$

the probability of being alive just before duration t, or more generally, the probability that the event of interest has not occurred by duration t. That is, the survival function gives the probability of surviving a specific amount of time.

An alternative characterization of the distribution of T is given by the *hazard function*, or the instantaneous rate of occurrence of the event, defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t; t + dt)$ given that it has not occurred before, and the denominator is the width of the interval, dt. Dividing one by the other, we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes to zero, we obtain an instantaneous rate of occurrence.

The interpretation of the hazard function is *not* as a probability but as a rate per unit time. The hazard assesses the risk—at a particular moment—that an individual who has not yet done so will experience the event. Unlike probabilities, rates can exceed 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 $\lambda(t) = 0.5$ and that the unit of time is one month. Then, on average, 0.5 events will occur per individual at risk per month (assuming the hazard remains constant during this period).

Because it expresses the instantaneous risk of an event, the hazard rate is the natural response variable for regression models for survival data.

The hazard function is sometimes written as:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

That is, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event.

Since $f(t)$ is the derivative of $S(t)$,

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

Since the event cannot have occurred by duration 0, $S(0) = 1$. By integrating from 0 to t, we can solve for the probability of surviving up to duration t as a function of the hazard at all durations up to t:

$$S(t) = \exp\left[-\int_0^t \lambda(x)dx\right]$$

The integral in braces in this equation is called the cumulative hazard (or cumulative risk) and is denoted:

$$\Lambda(t) = \int_0^t \lambda(x)dx = -\log_e S(t)$$

the sum of the risks you face going from duration 0 to t.

Model Fitting

Most interesting survival analysis examines the relationship between survival — typically in the form of the hazard function — and one or more explanatory variables (or covariates).

The hazard is typically modeled using the log transformation because it is a rate rather than a probability. Therefore, rather than being bounded by 0 and 1, it is only bounded below, at 0. Thus, we can use a linear model for the log-hazard or a multiplicative model for the hazard itself.

There are essentially three approaches to fitting survival models:

- The first and perhaps most straightforward is the parametric approach, where we assume a specific functional form for the baseline hazard $\lambda_0(t)$. Examples are models based on the exponential, Weibull, gamma, and generalized F distributions.
- The second approach is a semi-parametric approach that focuses on estimation of the regression coefficients β leaving the baseline hazard $\lambda_0(t)$ completely unspecified. This approach relies on a partial likelihood function proposed by Cox (1972) in his original paper.
- A third approach is a flexible, semi-parametric approach, where we make mild assumptions about the baseline hazard $\lambda_0(t)$. Specifically, we may subdivide time into reasonably small intervals and assume that the baseline hazard is constant in each interval, leading to a piece-wise exponential model, which can be fit using Poisson regression. This approach is sufficiently flexible to provide a useful tool with wide applicability.

We will not discuss the first approach. If we have time, we may discuss the third approach but we will focus primarily on the second approach.

Cox Proportional Hazard Model

The Cox proportional hazards model assumes a baseline hazard function, $\lambda_0(t)$, (i.e., the hazard for the reference group) that may vary over time and is not required to be estimated.

The model is given as:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{ik}x_{ik}\}$$

or equivalently,

$$\log_e \lambda_i(t) = \log_e \lambda_0(t) + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{ik}x_{ik}$$

Thus, the linear predictor is:

$$\eta_i = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{ik}x_{ik}$$

The intercept is absorbed in the baseline hazard.

The major assumption of the Cox model is that the hazard ratio, e^{β_j}, for a predictor, X_j, is constant and does not depend on time (i.e., the hazards in the two groups are proportional over time). Each predictor has a multiplicative effect on the hazard.

For example, using the leukemia data, our model would be:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_{i1}\ \text{group}_{i1})$$

If we take the ratio of the two hazard functions, $\lambda_i(t)$ when group=1 and $\lambda_0(t)$ when group=0, we get

$$\lambda_i(t)/\lambda_0(t) = \exp(\beta_1)$$

implying that the hazards are proportional. This assumption implies that, the hazard functions for the groups should be proportional and cannot cross.

This generalizes to more than one predictor and to quantitative predictors. If two individuals differ in their predictor values, the hazard ratio for these two individuals is independent of time. Thus, if an individual has a risk of death at some initial time point that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.

The Cox Proportional Hazard model is fit using the method of partial likelihood. We will not go into mathematical details of the partial likelihood but there are several implications.

1. The shape of the baseline hazard is irrelevant - it is nowhere to be seen in the log likelihood calculation.
2. The precise event times are irrelevant as well.
3. Ties substantially complicate calculations, in which case an exact method is available, but often it is impractical, and one of several approximations can be tried. Breslow's approximation 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).