Machine Learning for Business

Module 8: Survival analysis
Day 4, 13.00 – 16.00

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

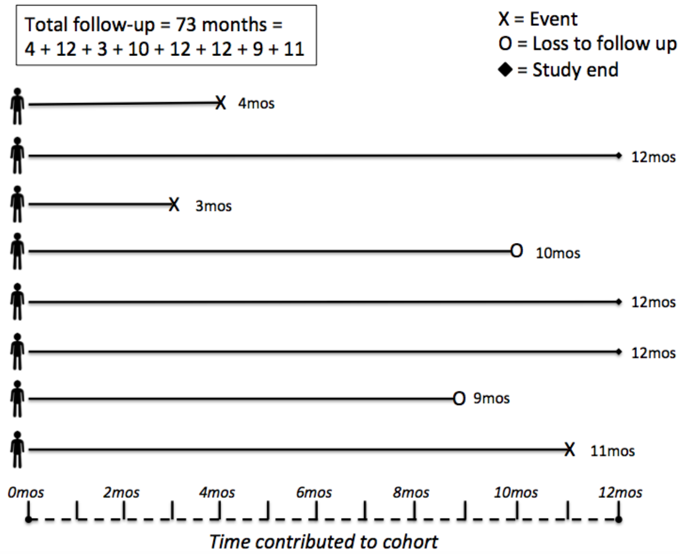
BIG DATA EXPERIENCE CENTER

Logistic regression vs time

- In logistic regression, we were interested in studying how risk factors were associated with presence or absence of disease.
- Sometimes, we are interested in how a risk factor or treatment affects time to disease or some other event.
- In these cases, logistic regression is not appropriate.

Survival analysis

- Survival analysis is used to analyze data in which the time until the event is of interest.
- The response is often referred to as a failure time, survival time, or event time.



Example:

- Time until tumor recurrence
- Time until a machine part fails
- Time until mobile phone recharge
- Time until the next credit card usage

The survival time response

- Usually continuous
- May be incompletely determined for some subjects
  - For some subjects we may know that their survival time was at least equal to some time *t*.
  - Whereas, for other subjects, we will know their exact time of event.
- Incompletely observed responses are **censored**
- Is always≥0.

Analysis issue

- If there is no censoring, standard linear regression procedures could be used.
- However, these may be inadequate because
  - Time to event is restricted to be positive and has a skewed distribution.
  - The probability of surviving past a certain point in time may be of more interest than the expected time of event.
  - The hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

Censoring

- Censoring is present when we have some information about a subject's event time, but we don't know the exact event time.
- For the analysis methods we will discuss to be valid, censoring mechanism must be independent of the survival mechanism.

Reasons censoring might occur

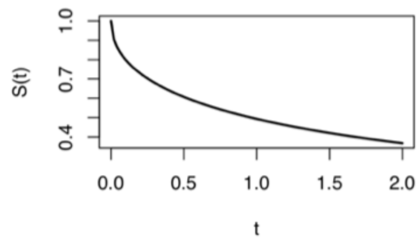
- A subject does not experience the event before the study ends
- A person is lost to follow-up during the study period
- A person withdraws from the study

These are all examples of **right-censoring.**

Terminology and notation

- T denotes the response variable, $T \geq 0$.
- The survival function is

$$S(t) = Pr(T > t) = 1 - F(t).$$



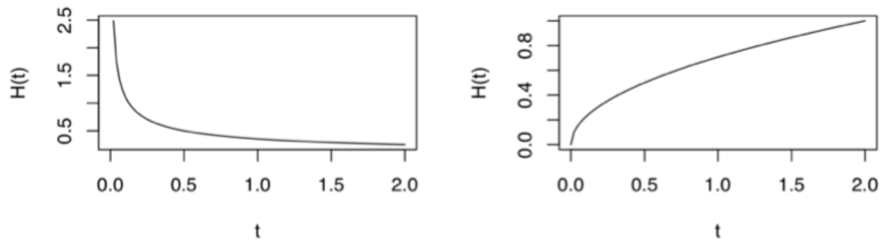
Survival function

- The survival function gives the probability that a subject will survive past time *t*.
- As t ranges from 0 to ∞ , the survival function has the following properties
  - It is non-increasing
  - At time t= 0, S(t) = 1. In other words, the probability of surviving past time 0 is 1.
  - At time t=∞, S(t)=S(∞)=0. As time goes to infinity, the survival curve goes to 0.
- In theory, the survival function is smooth.
- In practice, we observe events on a discrete time scale (days, weeks, etc.).

- The hazard function, $h(t)$, is the instantaneous rate at which events occur, given no previous events.

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

- The cumulative hazard describes the accumulated risk up to time t, $H(t) = \int_0^t h(u)du$.



If we know any one of the functions $S(t)$, $H(t)$, or $h(t)$, we can derive the other two functions.

$$h(t) = -\frac{\partial \log(S(t))}{\partial t}$$

$$H(t) = -\log(S(t))$$
$$S(t) = \exp(-H(t))$$

Survival data

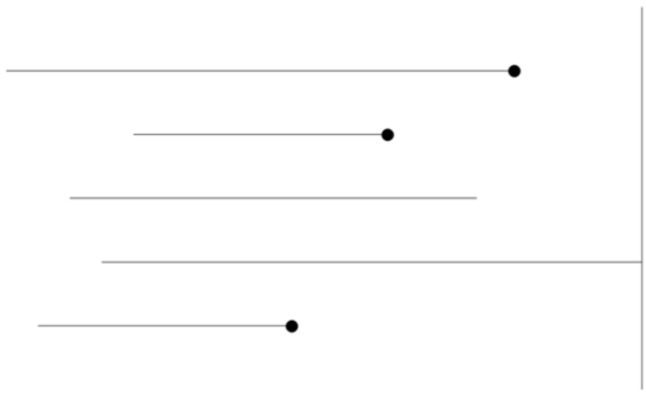
How do we record and represent survival data with censoring?

- T_i denotes the response for the ith subject.
- Let C_i denote the censoring time for the ith subject
- Let δ_i denote the event indicator

$$\delta_i = \begin{cases} 1 \text{ if the event was observed } (T_i \leq C_i) \\ 0 \text{ if the response was censored } (T_i > C_i). \end{cases}$$

- The observed response is $Y_i = \min(T_i, C_i)$.

Example



T_i	C_i	Y_i	δ_i
80	100	80	1
40	80	40	1
74+	74	74	0
85+	85	85	0
40	95	40	1

Termination of study

Estimating $S(t)$ and $H(t)$

If we are assuming that every subject follows the same survival function (no covariates or other individual differences), we can easily estimate $S(t)$.

- We can use nonparametric estimators like the Kaplan-Meier estimator
- We can estimate the survival distribution by making parametric assumptions
  - exponential
  - Weibull
  - Gamma
  - log-normal

Non-parametric estimation of S

- When no event times are censored, a non-parametric estimator of $S(T)$ is $1 - F_n(t)$, where $F_n(t)$ is the empirical cumulative distribution function.
- When some observations are censored, we can estimate $S(t)$ using the Kaplan-Meier product-limit estimator.

t	No. subjects at risk	Deaths	Censored	Cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	0	1	$20/20 \times 0.769 = 0.769$
377	19	0	1	$19/19 \times 0.769 = 0.769$
421	18	0	1	$18/18 \times 0.769 = 0.769$
431	17	1	0	$16/17 \times 0.769 = 0.688$
⋮				⋮
⋮				⋮

R: Kaplan-Meier estimator
Loading data

```
library(survival)
data(ovarian)
head(ovarian)
```

```
  futime fustat     age resid.ds rx ecog.ps
1     59      1 72.3315        2  1       1
2    115      1 74.4932        2  1       1
3    156      1 66.4658        2  1       2
4    421      0 53.3644        2  2       1
5    431      1 50.3397        2  1       1
6    448      0 56.4301        1  1       2
```

Survival object

?Surv

```
Surv(time, time2, event,
    type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'
    origin=0)
is.Surv(x)
```

Arguments

`time`	for right censored data, this is the follow up time. For interval data, the first argument is the starting time for the interval.
`event`	The status indicator, normally 0=alive, 1=dead. Other choices are `TRUE`/`FALSE` (`TRUE` = death) or 1/2 (2=death). For interval censored data, the status indicator is 0=right censored, 1=event at `time`, 2=left censored, 3=interval censored. Although unusual, the event indicator can be omitted, in which case all subjects are assumed to have an event.

R: Kaplan-Meier estimator
Loading data

```
S1 <- Surv(ovarian$futime,ovarian$fustat)
S1
```

```
 [1]   59   115   156   421+  431   448+  464   475   477+
[10]  563   638   744+  769+  770+  803+  855+ 1040+ 1106+
[19] 1129+ 1206+ 1227+  268   329   353   365   377+
```

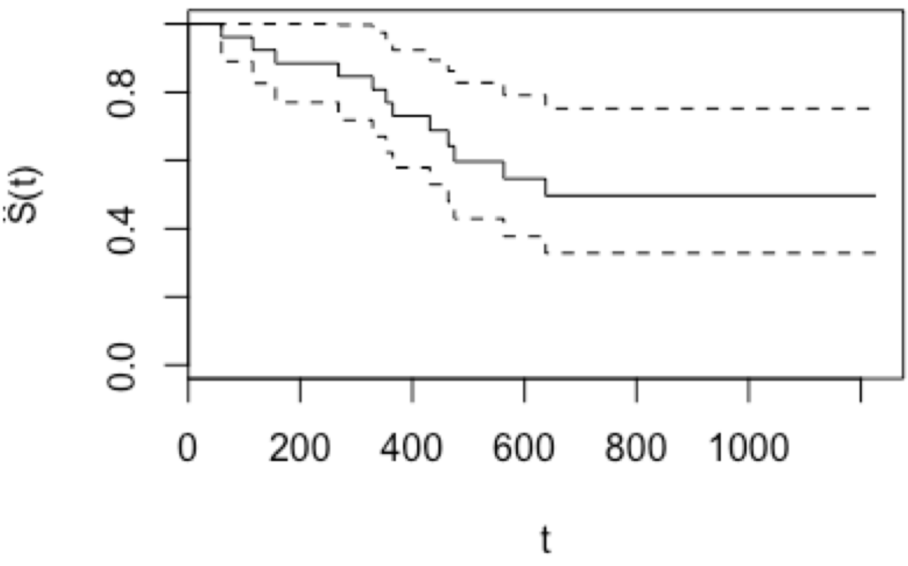
Fitting survival data

```
fit1 <- survfit(S1 ~ 1)
summary(fit1)
```

```
Call: survfit(formula = S1 ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

```
plot(fit1,xlab="t",
        ylab=expression(hat(S)*"(t)"))
```



Parametric survival functions

- The Kaplan-Meier estimator is a very useful tool for estimating survival functions.
- Sometimes, we may want to make more assumptions that allow us to model the data in more detail.

Benefit of using parametric survival functions

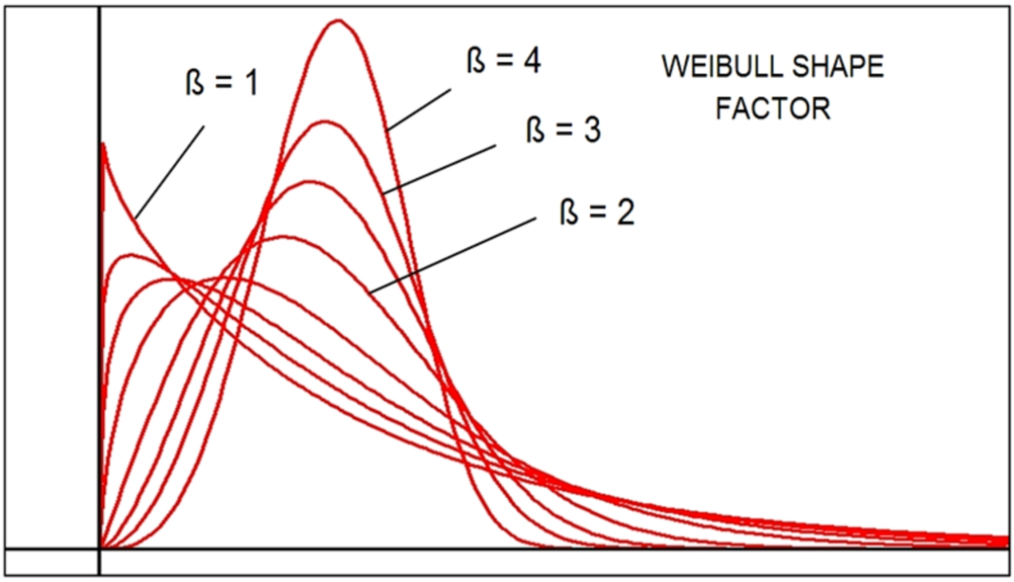
By specifying a parametric form for S(t), we can

- easily compute selected quantiles of the distribution
- estimate the expected failure time
- derive a concise equation and smooth function for estimating S(t),H(t) and h(t)
- estimate S(t) more precisely than KM assuming the parametric form is correct!

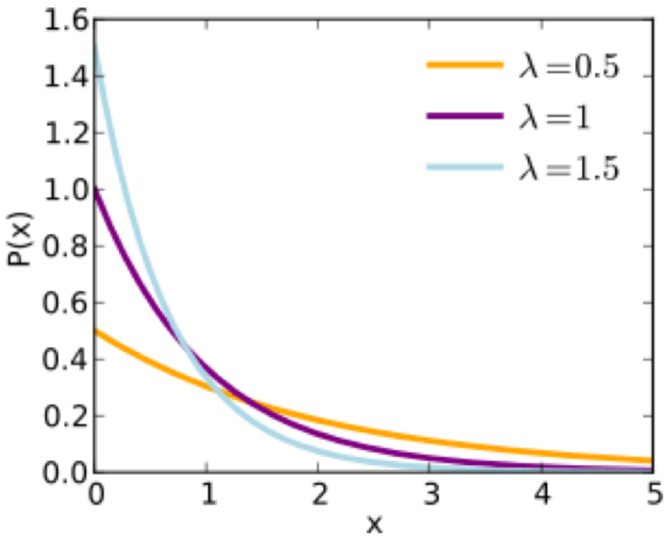
Suitable distribution for survival analysis

- Weibull
- Exponential
- log-normal (log(T) has a normal distribution)
- log-logistic

Weibull distribution



Exponential distribution



Estimation for parametric $S(t)$

We will use maximum likelihood estimation to estimate the unknown parameters of the parametric distributions.

- If Y_i is uncensored, the ith subject contributes $f(Y_i)$ to the likelihood
- If Y_i is censored, the ith subject contributes $Pr(y > Y_i)$ to the likelihood.

The joint likelihood for all n subjects is

$$L = \prod_{i:\delta_i=1}^{n} f(Y_i) \prod_{i:\delta_i=0}^{n} S(Y_i).$$

The log-likelihood can be written as

$$\log L = \sum_{i:\delta_i=1}^{n} \log(h(Y_i)) - \sum_{i=1}^{n} H(Y_i).$$

Example

- Let's look at the ovarian data set in the survival library in R.
- Suppose we assume the time-to-event follows an exponential distribution, where

$$h(t) = \lambda$$

and

$$S(t) = \exp(-\lambda t).$$

R: parametric survival function

```
s2 <- survreg(Surv(futime, fustat)~1,
                   ovarian, dist='exponential')
summary(s2)
```

```
Call:
survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "exponential")
            Value Std. Error    z        p
(Intercept)  7.17      0.289 24.8 3.72e-136

Scale fixed at 1

Exponential distribution
Loglik(model)= -98   Loglik(intercept only)= -98
Number of Newton-Raphson Iterations: 4
n= 26
```

Interpreting parameter

In the R output,

$$\begin{aligned}\lambda &= \exp(-(\text{Intercept})) \\ &= \exp(-7.17)\end{aligned}$$

Therefore,

$$S(t) = \exp(-\exp(-7.17)t).$$

Survival analysis with covariate

- If we assumes that "rx", may affect the survival function, we may put it in as a factor

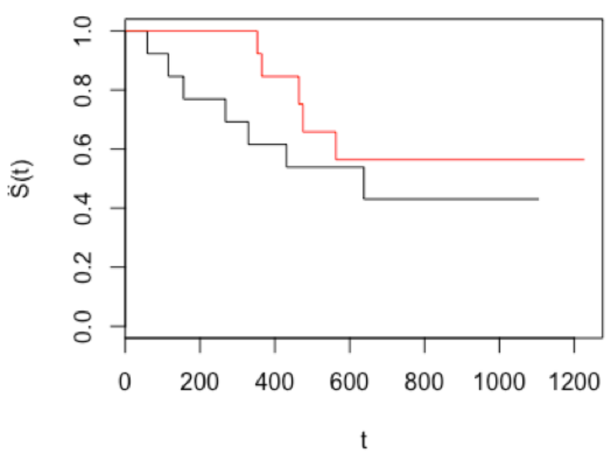
```
fit3 <- survfit(Surv(futime,fustat) ~ rx, data = ovarian)
summary(fit3)
```

```
Call: survfit(formula = Surv(futime, fustat) ~ factor(rx), data = ovarian)

                factor(rx)=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     13       1    0.923  0.0739        0.789        1.000
  115     12       1    0.846  0.1001        0.671        1.000
  156     11       1    0.769  0.1169        0.571        1.000
  268     10       1    0.692  0.1280        0.482        0.995
  329      9       1    0.615  0.1349        0.400        0.946
  431      8       1    0.538  0.1383        0.326        0.891
  638      5       1    0.431  0.1467        0.221        0.840

                factor(rx)=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  353     13       1    0.923  0.0739        0.789        1.000
  365     12       1    0.846  0.1001        0.671        1.000
  464      9       1    0.752  0.1256        0.542        1.000
  475      8       1    0.658  0.1407        0.433        1.000
  563      7       1    0.564  0.1488        0.336        0.946
```

```
plot(fit3,xlab="t",
ylab=expression(hat(S)*"(t)"),col=1:2)
```



Thank you

Question?