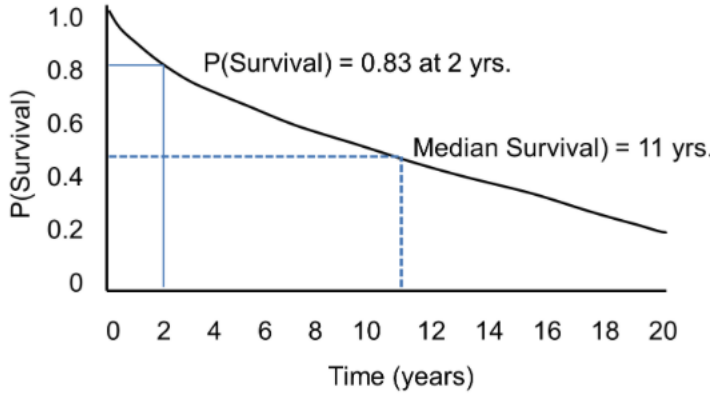
Chapter 16. Survival Analysis

16.1. Time-to-Event

- In many clinical/medical researches, the 'time to event' *T* is a variable of primary interest.
  - *T*: None-negative random variable
  - Event: Death, failure, equipment breakdown, development of some disease, etc.
  - Clinical endpoint, survival time, of failure time
- Generally, not symmetrically distributed.
  - Only few subjects survive longer compared to the majority.
- Survival time is right censored.
  - At the end of the study, some subjects may not have reached the endpoint of interest.
  - Assumption: Time-to-event is independent of the censoring mechanism.

- Example
  - Time until cardiovascular death after some treatment intervention
  - Time until tumor recurrence
  - Remission duration of certain disease in clinical trials
  - Incubation time of certain disease (e.g. AIDS, Hepatitis C)



  - The 10-year survival rate of patients with stage 3 colon cancer after the diagnosis
  - Which gender is more likely to survive 3 years after a surgery?

16.2. Incomplete Data: Censoring

- Censoring happens when a value occurs outside the range of a measuring instrument.
- Reasons of censoring: Withdrawal, lost to follow-up, event-free at last follow-up, death due to another cause, etc.

Type	Description
Right censoring	The individual is still alive or has not experienced the event of interest at the end of the study.
Left censoring	The individual has already experienced the event of interest prior to the start of the study. We know that the event occurred, but are not sure when exactly it happened. e.g. First time smoking, Alzheimer disease (onset hard to determine)
Interval censoring	The event occurs within some interval. Due to discrete observation times, actual event time is unknown.
Type I censoring	An experiment has a set number of subjects and the study ends at a predetermined *time*. Some subjects remain right-censored.
Type II censoring	An experiment has a set of number of subjects and the study ends when a predetermined *number* of subjects experience the event of interest. Some subjects remain right-censored.

16.3. Incomplete Data: Truncation

- Truncation ≠ Censoring
- Truncation occurs when the incomplete nature of the observation is due to a systematic selection process inherent to the study design (sampling bias).
- Only those individuals whose time of event lies within a certain interval $[Y_L, Y_R]$ are included.

Type	Description
Right truncation	Only individuals who have experienced the event by a specified time are included in the sample. e.g. Patients with AIDS from transfusion (Only patients who were infected with AIDS virus after March 1, 2005 and developed AIDS by June 30, 2014 are included.)
Left truncation	Only individuals who survive a sufficient time are included in the sample. e.g. Death time of elderly residents of a retirement community (Only the elderly people of a certain age can be admitted into the community. People died before this age cannot be included.)

16.4. Important Functions

- Let *T* be the survival time (time-to-event) with pdf $f(t),\ t \in [0, \infty)$.
- Cumulative distribution function $F(t)$

$$F(t) = P(T \leq t) = \int_0^t f(s)\, ds$$

- Survival function $S(t)$

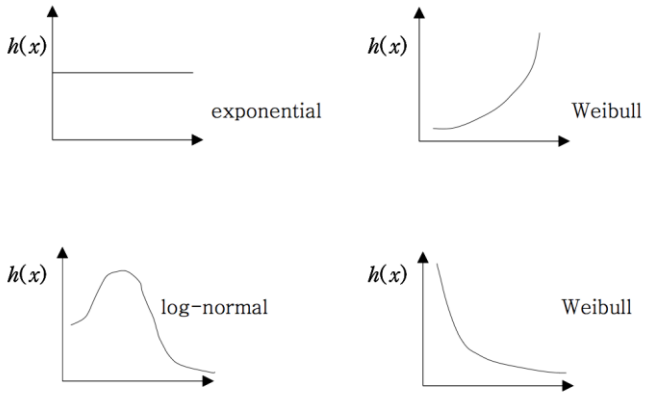
$$S(t) = P(T > t) = \int_t^\infty f(s)\, ds = 1 - P(T \leq t) = 1 - F(t)$$

  - $S(0) = 1$
  - $S(\infty) = \lim_{t \to \infty} S(t) = 0$
  - $S(t)$ is non-increasing in t and right-continuous.

- Hazard function $\lambda(t)$

$$\lambda(t) = \lim_{h \to 0+} \frac{P(t \leq T < t + h \mid T \geq t)}{h}$$

  - *Instantaneous* failure rate at *t* given survival up to *t*
  - Conditional failure rate / Intensity function / Force of mortality / Instantaneous hazard

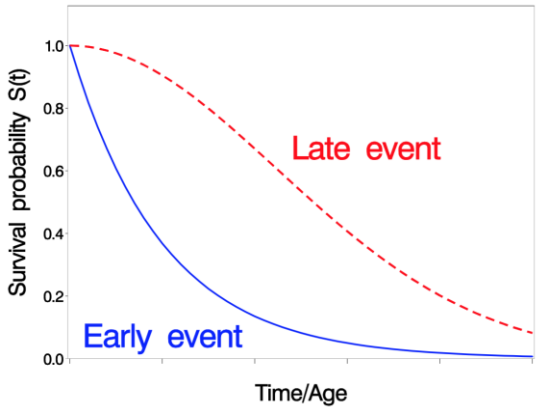


- Cumulative hazard function $\Lambda(t)$

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

- Relationship between functions
  - $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$
  - $\lambda(t) = \frac{d\Lambda(t)}{dt} = -\frac{d\log S(t)}{dt} = \frac{f(t)}{S(t)}$
  - $\Lambda(t) = -\log S(t)$
  - $\mathrm{S(t)} = e^{-\Lambda(t)} = \exp\left(-\int_0^t \lambda(s)\, ds\right)$
  - $\Lambda(\infty) = \infty\ (\because S(\infty) = 0)$

- Example: Compare survival or hazard functions of different groups. (e.g. treatment / gender)



Survival analysis considers *censoring* and *time-dependent* covariates.

- ✓ Compare mean time to events (t-test or linear regression) → Ignore censoring
- ✓ Compare proportion of events (Relative risk, OR, or logistic regression) → Ignore time

16.5. Survival Analysis: Notation

- Mainly focus on right censored data
  - *X*: True survival time
  - *C*: Censoring time
  - $\Delta = I(X \leq C)$: Censoring indicator (0 if censored)
  - $Y = \min(X, C)$: What we actually observe
  - Censoring time is independent of the event of interest. (i.e. *X* is independent of *C*)
- Survival analysis

Approach	Description
Parametric	e.g. Exponential / Weibull distribution
Nonparametric	No specification about the distribution of survival time Kaplan-Meier (product-limit) estimator

16.6. Kaplan-Meier (K-M) Estimator

- Suppose for *n* subjects,
  - Observed at distinct, ordered time points $t_1 < t_2 < \cdots < t_k,\ k \leq n$
  - d_i: The number of failures at time t_i
  - n_i: The number of subjects at risk (i.e. no event and not censored) just prior to t_i

    (Size of the risk set)
- If there exist right censored individuals,
  - Estimated hazard function

$$\hat{\lambda}(t_i) = \frac{d_i}{n_i}, \qquad i = 1, 2, \cdots k$$

  - Kaplan-Meier estimator of survival function

$$\hat{S}(t) = \prod_{i:\, t_i \leq t} \left(1 - \hat{\lambda}(t_i)\right) = \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

- Note that the K-M estimator is undefined after the largest observed failure time.

  DO NOT extrapolate!

Example

Data A small study is looking at time to relapse after a cancer treatment.
Data from 10 patients is shown below; censored observations are marked by (+):

10, 20+, 35, 40+, 50+, 55, 70+, 71+, 80, 90+

K-M Estimator

Time (t)	Died (d_i)	At risk (n_i)	$\hat{\lambda}(t_i) = d_i/n_i$	$1 - \hat{\lambda}(t_i)$	$\hat{S}(t)$
10	1	10	1/10	9/10	0.9
20	0	9	0	1	0.9 X 1 = 0.9
35	1	8	1/8	7/8	0.9 X 7/8 = 0.79
40	0	7	0	1	0.79 X 1 = 0.79
50	0	6	0	1	0.79 X 1 = 0.79
55	1	5	1/5	4/5	0.79 X 4/5 = 0.63
70	0	4	0	1	0.63 X 1 = 0.63
71	0	3	0	1	0.63 X 1 = 0.63
80	1	2	1/2	1/2	0.63 X 1/2 = 0.32
90	0	1	0	1	0.32 X 1 = 0.32

- Median survival time
  - The median survival time ($t_{50\%}$) is the time beyond which 50% of the individuals in the population of interest are expected to survive. (i.e. $S(t_{50\%}) = 0.5$)
  - The estimated median survival time ($\hat{t}_{50\%}$) is defined as the smallest observed time for which the estimated survival function is less than 0.5.
  - Sometimes, the estimated survival function is greater than 0.5 for all values of *t*, then there is no median survival time.

- PROC LIFETEST
  - Estimate survival functions / K-M table
  - SAS output includes K-M table.
  - Generate graphs and confidence limits.
  - If dataset contains only complete and right-censored observations, PROC LIFETEST requires two components:

    a) Time (of event or censoring)

    b) Censoring indicator (1: event / 0: censored)

General Syntax

```
proc lifetest data=dataset plots=(survival);
     time time-variable*censoring-indicator(0);
run;
```

  - Time variable always comes first.
  - The censoring indicator needs to be numeric.
  - SAS needs to know which observations are censored.

Example

Raw Data

Obs	Day	Treatment	Status
1	4	1	1
2	5	1	1
3	9	1	1
4	10	1	1
5	11	1	1

SAS Code

```
* Kaplan-Meier (K-M) Estimator;
proc lifetest data=leukemia plots=(survival(cl));
     time day*status(0);
run;
```

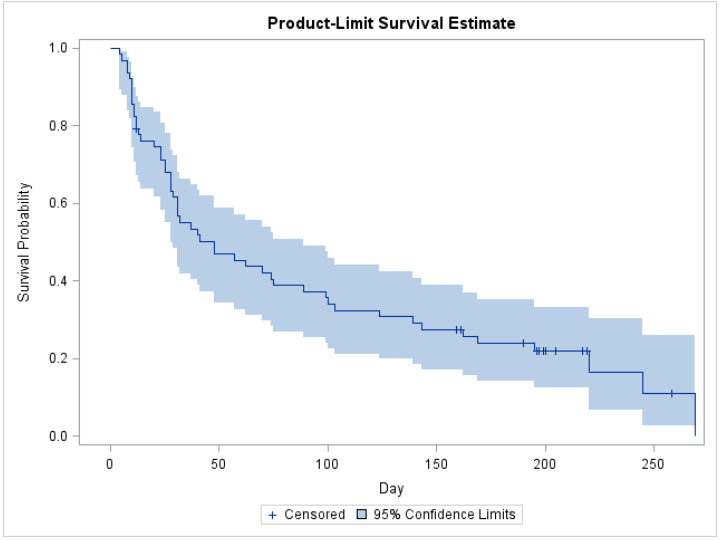
Output

Product-Limit Survival Estimates

Days		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	63
4.000		0.9841	0.0159	0.0157	1	62
5.000		0.9683	0.0317	0.0221	2	61
8.000		.	.	.	3	60
8.000		0.9365	0.0635	0.0307	4	59

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	169.000	LOGLOG	99.000	269.000
50	48.000	LOGLOG	28.000	89.000
25	20.000	LOGLOG	10.000	28.000



16.7. Log-Rank Test

- Compare the survival functions between (2+) groups.

  (e.g. treatment, some demographic characteristics)

- No assumptions about the distribution of survival functions are required.
- Suppose there are *J* groups. The null hypothesis here will be

$$H_0: S_1(t) = \cdots = S_J(t) \text{ for all } t.$$

- Optimal power for detecting differences when hazards are proportional.

General Syntax

```
proc lifetest data=dataset;
     time time-variable*censoring-indicator(0);
     strata strata-variable / test=(all) adjust=mc-method diff=control("ref");
run;
```

- STRATA: Specify the grouping variable.
- TEST=(all): Provide all the available nonparametric tests.
- ADJUST=: Select the multiple comparison adjustment method.[10]
- DIFF=: Declare the reference group.
- If hazards cross, the log-rank test may *not* be suitable.
- Weighting allows the test to depend on the event time and the censoring distribution. (e.g. Gehan, Peto-Peto, Fleming-Harrington)
- Stratified analysis
  - Compare survival functions among one category across other categories.
  - e.g. Clinic (multicenter clinical trial), age group, gender

[10] Check Chapter 12.9. for possible multiple comparison options.

Example

Raw Data

Obs	Day	Treatment	Status
1	4	1	1
2	5	1	1
3	9	1	1
4	10	1	1
5	11	1	1

SAS Code

```
* Log-Rank Test;
proc lifetest data=leukemia plots=(survival);
     time day*status(0);
     strata treatment / test=(all);
run;
```

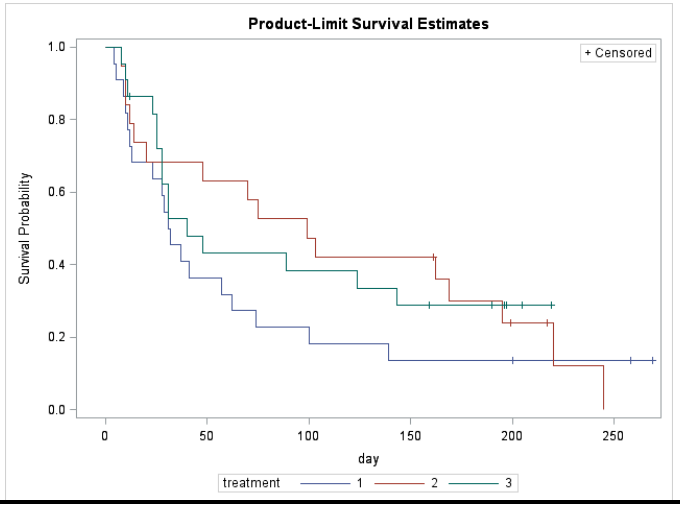
Output

Rank Statistics

treatment	Log-Rank	Wilcoxon	Tarone	Peto	ModifiedPeto	Fleming
1	4.0501	213.00	31.26	3.2454	3.2074	3.2399
2	-1.9887	-137.00	-19.57	-1.9661	-1.9631	-2.0043
3	-2.0614	-76.00	-11.69	-1.2793	-1.2443	-1.2356

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1.6425	2	0.4399
Wilcoxon	2.6843	2	0.2613
Tarone	2.5914	2	0.2737
Peto	2.5625	2	0.2777
Modified Peto	2.6242	2	0.2693
Fleming(1)	2.3804	2	0.3042



16.8. Proportional Hazards (PH) Model

- Also known as Cox's regression
- Link the survival functions to multiple covariates (explanatory variables).
- Quantify the *effect* of a certain predictor on survival function.
- Allow to predict survival functions based on a set of covariates.
- Semi-parametric model
  - Make a parametric assumption on the effect of predictors on hazard function.
  - No assumption regarding the nature of the hazard function itself
- With non-time-varying covariates $Z = (Z_1, Z_2, \cdots, Z_k)$, the PH model specifies that

$$\lambda(t|Z) = \lambda_0(t)\, e^{\sum_{i=1}^{k} \beta_i Z_i}$$

  - $\lambda_0(t)$: Arbitrary baseline hazard rate (nonparametric).
  - $\beta = (\beta_1, \beta_2, \cdots, \beta_k)$: Regression coefficients
  - More complicated case: Time-varying covariates $Z_i(t), i = 1, 2, \cdots, k$.

- Hazard ratio

$$\frac{\lambda_{z_1=1}(t|Z_2,\cdots,Z_k)}{\lambda_{z_1=0}(t|Z_2,\cdots,Z_k)} = \frac{\lambda_0(t)\, e^{\beta_1+\sum_{i=2}^{k}\beta_i Z_i}}{\lambda_0(t)\, e^{\sum_{i=2}^{k}\beta_i Z_i}} = e^{\beta_1}$$

  - HR > 1 ($\beta_1 > 0$) : Greater hazard / Worse survival
  - HR < 1 ($\beta_1 < 0$) : Less hazard / Better survival / Protective

- PROC PHREG

General Syntax

```
proc phreg data=dataset;
     model time-variable*censoring-indicator(0) = list-of-independent-variables;
run;
```

- Analysis steps

1. Start by checking the K-M estimates.
2. Fit the Cox proportional hazard (PH) model and get the hazard ratio (HR).
3. Test the proportionality assumption.

   a) The hazard ratio is constant over time, but proportional (multiplicative factor).
   b) The risk does not depend on time. That is, the risk is constant over time.
   ⇒ For each covariate,
   i) Plot survival functions / cumulative hazard functions / log(cumulative hazard).
   ii) Include an interaction term with time (usually log(time)).
   Proportionality condition is met if the interaction terms are not significant.

3.* If the proportionality assumption is not satisfied,

   a) Stratify the model by the non-proportional covariates.
   b) Run Cox models on time intervals rather than on entire time domain.
   c) Include a covariate interaction with time as a predictor.

4. Check the functional form of continuous variables.
   (e.g. Linear, quadratic, categorized form)
5. Look at the residuals. (Random pattern of residuals evenly distributed around zero)

Example

Raw Data

Obs	id	age	beck	hercoc	ivhx	ndrugtx	race	treat	site	los	time	censor
1	1	39	9	4	3	1	0	1	0	123	188	1
2	2	33	34	4	2	8	0	1	0	25	26	1
3	3	33	10	2	3	3	0	1	0	7	207	1
4	4	32	20	4	3	1	0	0	0	66	144	1
5	5	24	5	2	1	5	1	1	0	173	551	0

SAS Code

```
* Cox PHM - multiple predictors;
proc phreg data=uis;
      class treat(ref="0") race(ref="1");
      model time*censor(0)
             = treat age race/rl;
run;
```

Output

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.8131	3	0.0012
Score	15.5781	3	0.0014
Wald	15.5092	3	0.0014

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	1	1	-0.21981	0.08984	5.9861	0.0144	0.803	0.673	0.957	treat 1
age		1	-0.01210	0.00720	2.8219	0.0930	0.988	0.974	1.002	
race	0	1	0.25989	0.10653	5.9515	0.0147	1.297	1.052	1.598	race 0