Introduction to survival analysis

Jorge N. Tendeiro

Department of Psychometrics and Statistics
Faculty of Behavioral and Social Sciences
University of Groningen

✉ j.n.tendeiro@rug.nl
🌐 www.jorgetendeiro.com
 jorgetendeiro/Seminar-2020-Survival-Analysis

Plan for today

Gentle introduction to survival analysis.

Source:
Harrell, F. E., Jr. (2015). *Regression Modeling strategies*, 2nd edition.
Springer

Chapters:
17, 18, and 20.

Survival analysis (SA)

Data:
For which the *time until the event* is of interest.

▶ This goes beyond *logistic regression*, which focuses on the *occurrence* of the event.

Outcome variable:

▶ T = Time until the event.
▶ Often referred to as *failure time*, *survival time*, or *event time*.

Examples

Survival time: Time until. . .

▶ death, desease, relapse.

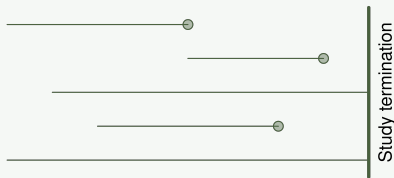*Failure time:* Time until. . .

▶ product malfunction.

Event time: Time until. . .

▶ graduation, marriage, divorce.

Advantages of SA over typical regression models

▶ SA allows modeling units that did not fail up to data collection (*censored on the right* data).

▶ Regression could be considered to model the expected survival time. *But*:
   - ✓ Survival time is often not normally distributed.
   - ✓ $P(\text{survival} > t)$ is often more interesting than $\mathbb{E}(\text{survival time})$.
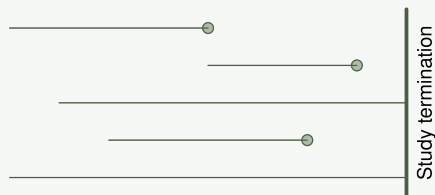
Censoring

▶ For some subjects, the event did not occur up to the end of data collection.
▶ These data are right-censored.

Define random variables for the ith subject:

▶ T_i = time to event
▶ C_i = censoring time
▶ e_i = event indicator = $\left\{ \begin{array}{ll} 1 & \text{if event is observed } (T_i \leq C_i) \\ 0 & \text{if event is not observed } (T_i > C_i) \end{array} \right.$
▶ $Y_i = \min(T_i, C_i)$ = what occurred first (failure or censoring)

Variables $\{Y_i, e_i\}$ include all the necessary information.

Typical data set

T_i	C_i	Y_i	e_i
5	10	5	1
4	12	4	1
13+	13	13	0
5	10	5	1
15+	15	15	0

Observe the flexibility of SA data:

▶ Subjects may join the study at different moments.
▶ Censoring times may differ among subjects.

$\{Y_i, e_i\}$ does include all the necessary information.

But, *assumption*: Censoring is non-informative, i.e., it is independent of the risk of the event.

Three main functions

Recall that the outcome variable is T = time until event.

▶ Survival function:

$$S(t) = P(T > t) = 1 - F(t),$$

where $F = P(T \leq t)$ is distribution function of T.

▶ Cumulative hazard function:

$$\Lambda(t) = -\log(S(t))$$

▶ Hazard function:

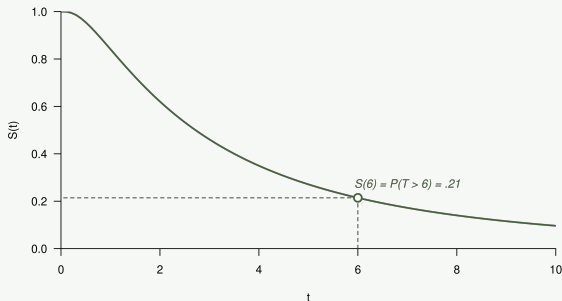$$\lambda(t) = \Lambda'(t)$$

Survival function

$$S(t) = P(T > t) = 1 - F(t)$$

Example:
If event = death, then $S(t)$ = prob. that death occurs after time t.

Properties:

- ▶ $S(0) = 1$, $S(\infty) = 0$.
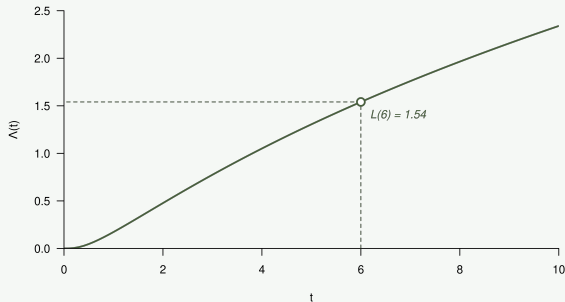- ▶ Non-increasing function of t.

Cumulative hazard function

$$\Lambda(t) = -\log(S(t))$$

Idea:
Accumulated risk up until time t.

Properties:
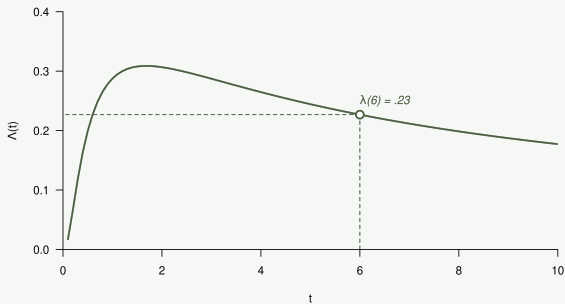
- ▶ $\Lambda(0) = 0$.
- ▶ Non-decreasing function of t.

Hazard function

$$\lambda(t) = \Lambda'(t)$$

Idea:
Instantaneous event rate at time t.

Relation between the three functions

All functions are related:
Any two functions can be derived from the third function.

► The three functions are equivalent ways of describing the same
  random variable (T = time until event).

More generally, all the following functions give mathematically
equivalent specifications of the distribution of T:

► $F(t)$: Distribution function
► $f(t)$: Density function
► $S(t)$: Survival function
► $\lambda(t)$: Hazard function
► $\Lambda(t)$: Cumulative hazard function.

Examples

Next are two primary examples of parametric survival distributions:

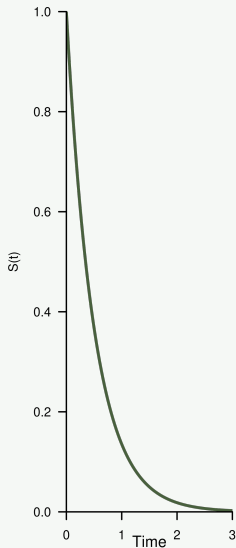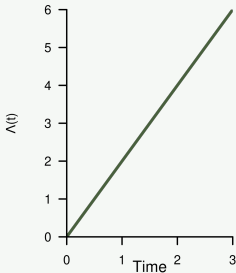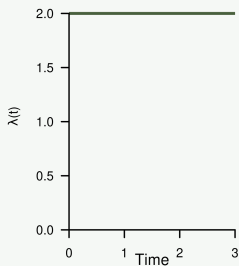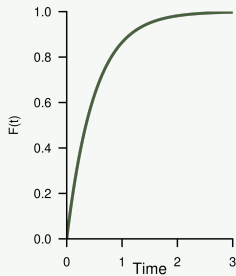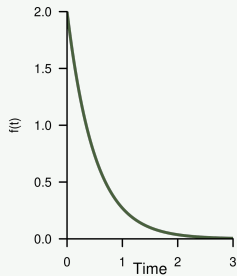▶ the exponential distribution;
▶ the Weibull distribution.

These models (still) include no covariates, thus:

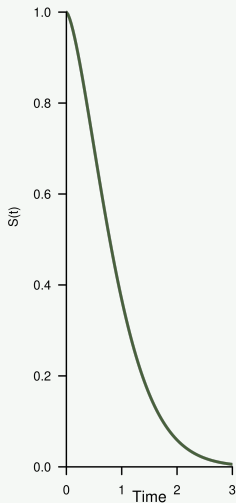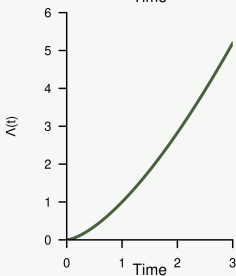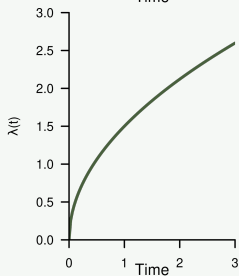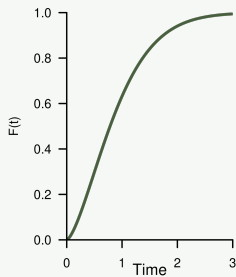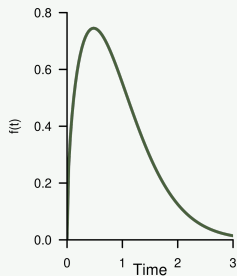▶ Each subject in the sample is assumed to have the same distribution of T.

No formulas.
Instead: Let's plot.

Exponential survival distribution

Weibull survival distribution (I)

Weibull survival distribution (II)
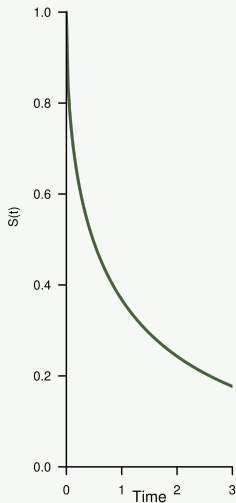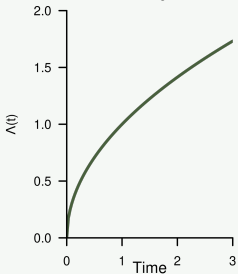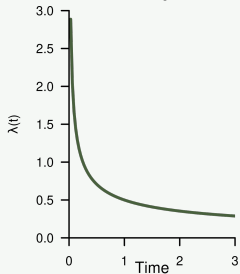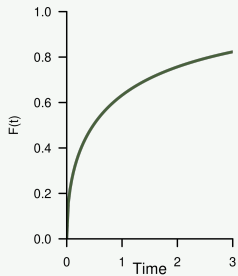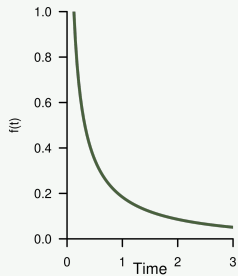
Quantiles

Q: What is the time by which $(100q)\%$ of the population will fail?

A: Value t_q such that $F(t_q) = q$, or, equiv., $S(t_q) = 1 - q$.
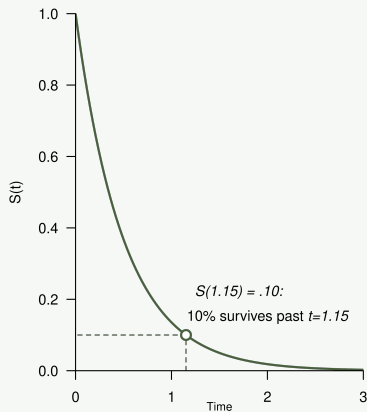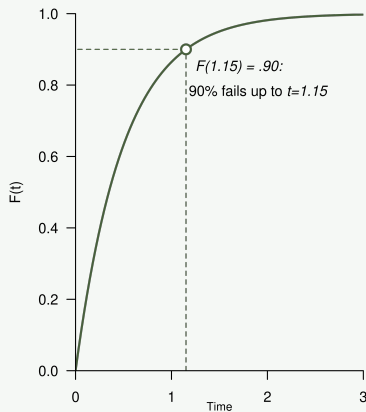
In particular, median survival time = $t_{.50}$.

Expected failure time

(Note: T is skewed, so the mean is not the best summary. Better use median.)

Q: What is the expected failure time?

A: It is the area under the survival function.
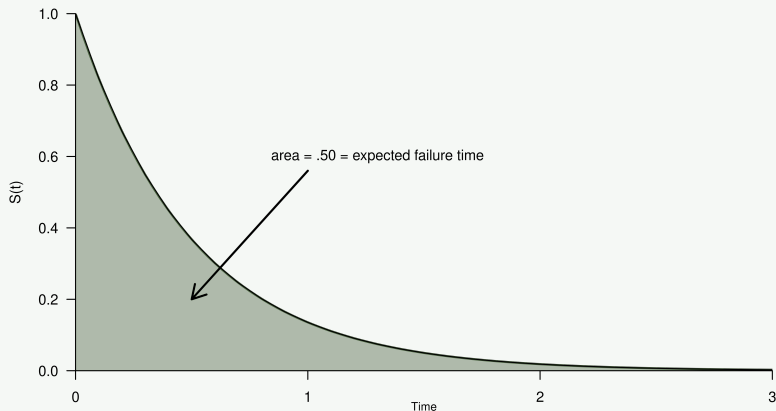
Various estimation approaches

There are several options available to estimate the survival function (and friends).

Here we will briefly go through only a few:

- ► Not parametric and homogeneous (i.e., without predictors):
  - ✓ Kaplan-Meier estimator
  - ✓ Altschuler-Nelson estimator
- ► Parametric:
  - ✓ Homogeneous (i.e., no predictors): Exponential, Weibull, normal, logistic, log-normal, log-logistic,...
  - ✓ Proportional hazards models
- ► Semi-parametric:
  - ✓ Cox proportional hazards regression model

After a brief intro to each, I will use them all on an empirical dataset.

Kaplan-Meier estimator

- ▶ Also known as the *product-limit* estimator.
- ▶ Non parametric, and super simple to do even manually.
- ▶ Key ingredient: *Conditional probabilites*.

Assume $t = 0, 1, 2, \ldots$
We have that $S(0) = P(T > 0) = 1$. For $t \geq 1$ we then have that

$$P(T > t | T > t - 1) = \frac{P(T > t, T > t - 1)}{P(T > t - 1)} = \frac{P(T > t)}{P(T > t - 1)}$$

and so

$$P(T > t) = P(T > t - 1) \times P(T > t | T > t - 1),$$

or in terms of the survival function,

$$S(t) = S(t - 1) \times P(T > t | T > t - 1)$$

$$\boxed{S(t) = S(t - 1) \times (1 - P(T \leq t | T > t - 1))}$$

Kaplan-Meier estimator – Example

Data: Seven subjects; failure times $T = 1, 3, 3, 3+, 6+, 9, 10+$.

Day	No. subjects at risk	Deaths	Censored	$S(t) = S(t-1) \times \times (1 - P(T \leq t \mid T > t-1))$
1	7	1	0	$1 \times (1 - 1/7) = 6/7$
3	$7 - (1+0) = 6$	2	1	$6/7 \times (1 - 2/6) = 4/7$
6	$6 - (2+1) = 3$	0	1	$4/7 \times (1 - 0/3) = 4/7$
9	$3 - (0+1) = 2$	1	0	$4/7 \times (1 - 1/2) = 2/7$
10	$2 - (1+0) = 1$	0	1	$2/7 \times (1 - 0/1) = 2/7$

Hence:

$$S(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 6/7 = .86, & 1 \leq t < 3 \\ 4/7 = .57, & 3 \leq t < 9 \\ 2/7 = .29, & 9 \leq t < 10 \\ \text{undefined}^*, & t \geq 10 \end{cases}.$$

*Not everyone failed by $t = 10$, so we cannot tell what happened after that.

Kaplan-Meier estimator – Example
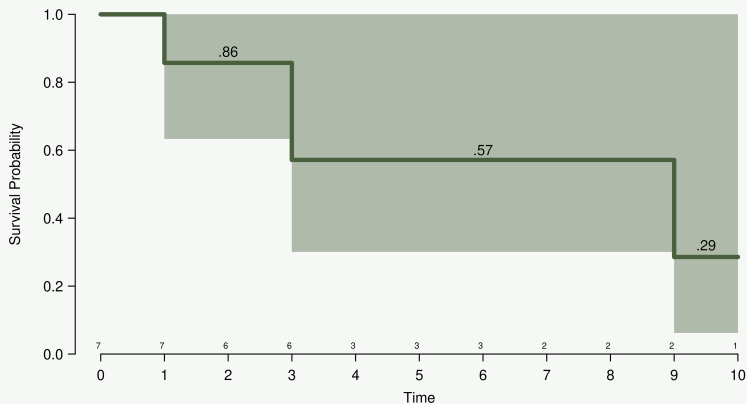
Altschuler-Nelson estimator

- ▶ Non parametric, also simple.
- ▶ Similar to Kaplan-Meier, but based on $\Lambda(t)$.

Recall that $\Lambda(t) = $ accumulated risk up until time t.

Hence it makes sense to estimate $\Lambda(t)$ by

$$\widehat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\# \text{ failures at } t_i}{\# \text{ subjects at risk at } t_i}.$$

Then,

$$\widehat{S}(t) = \exp(-\widehat{\Lambda}(t)).$$

Interesting property: $\sum_i \widehat{\Lambda}(Y_i) = $ total number of events.

Altschuler-Nelson estimator – Example

Data: Seven subjects; failure times $T = 1, 3, 3, 3+, 6+, 9, 10+$.

Day	No. subjects at risk	Deaths	Censored	$\Lambda(t)$
1	7	1	0	$1/7$
3	$7 - (1+0) = 6$	2	1	$1/7 + 2/6 = 10/21$
6	$6 - (2+1) = 3$	0	1	$10/21 + 0/3 = 10/21$
9	$3 - (0+1) = 2$	1	0	$10/21 + 1/2 = 41/42$
10	$2 - (1+0) = 1$	0	1	$41/42 + 0/1 = 41/42$
		$\sum_i = 4$		$\sum_i = 4$

Hence:

$$S(t) = \exp(-\Lambda(t)) = \begin{cases} \exp(0) = 1, & 0 \le t < 1 \\ \exp(-1/7) = .87, & 1 \le t < 3 \\ \exp(-10/21) = .62, & 3 \le t < 9 \\ \exp(-41/42) = .38, & 9 \le t < 10 \\ \text{undefined}, & t \ge 10 \end{cases} .$$

Altschuler-Nelson estimator – Example
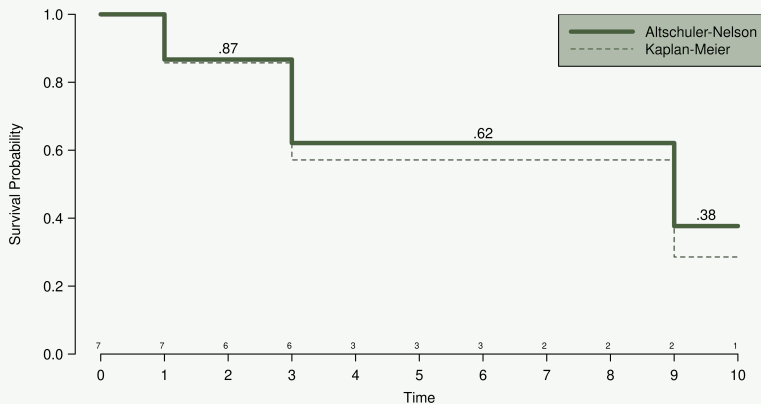
Homogeneous parametric models

Q: How about *continuous*, parametric, counterparts to KM and AN?
Still incorporating no predictors?

A: There are really *a lot* of possibilities.

Most common examples:

- ▶ Exponential
- ▶ Weibull
- ▶ Normal
- ▶ Logistic
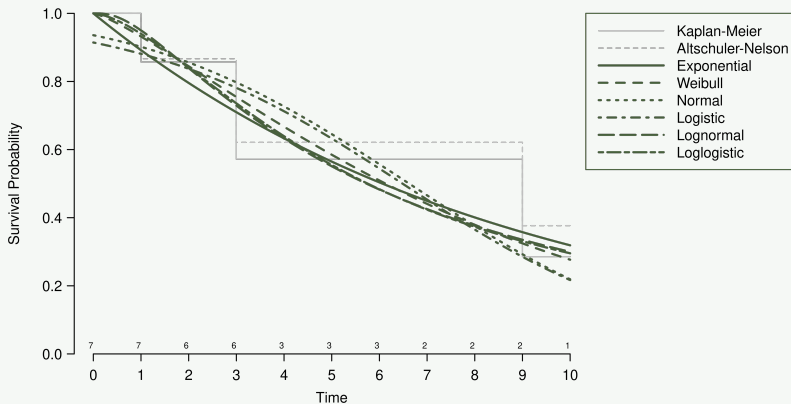- ▶ Log-normal
- ▶ Log-logistic
- ▶ ...

My advice:
Just fit several of these and compare.
There is no 'best' model, it depends on the data.

Homogeneous parametric models

Data: $T = 1, 3, 3, 3+, 6+, 9, 10+.$

Assessing model fit

I like Harrell's take on this:

▶ To assess model fit, use graphical methods
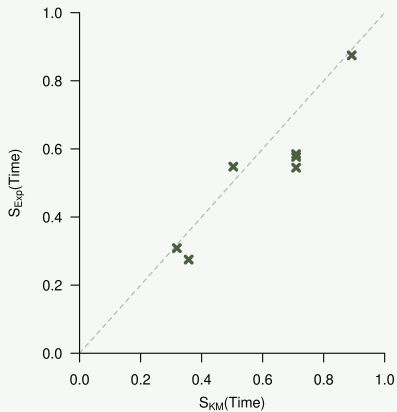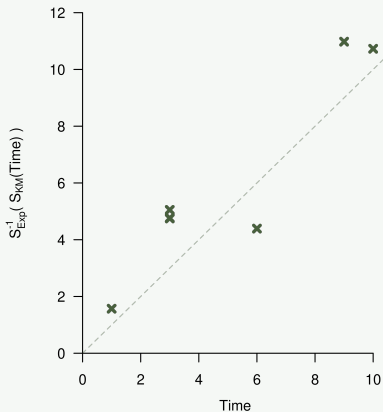   (and no tests; yeah, that's right!).

We show an example:
Assess the fit of the exponential model.

Two plotting options, akin to *QQ*-plots:

▶ Plot $S_{\text{Exp}}^{-1}(S_{\text{KM}}(T))$ versus T;
▶ Plot $S_{\text{Exp}}(T)$ versus $S_{\text{KM}}(T)$.

Assessing model fit

Data: $T = 1, 3, 3, 3+, 6+, 9, 10+$.

Parametric proportional hazards model

First model until now that allows incorporating predictor variables $X = \{X_1, X_2, \ldots, X_k\}$.

▶ X_i can be continuous, dichotomous, polytomous, etc.

The proportional hazards (PH) model generalizes the hazard function $\lambda(t)$:

$$\lambda(t|X) = \lambda(t)\overbrace{\exp(\underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k}_{X\beta})}^{\text{relative hazard function}} = \lambda(t)\exp(X\beta)$$

▶ $\lambda(t|X)$ = hazard function for T given the predictors X.
▶ $\lambda(t)$ = 'underlying' hazard function (for a subject with $X\beta = 0$).
▶ $\exp(X\beta)$ describes the *relative* effects of the predictors.

Note: The intercept β_0 may be omitted (kind of 'absorbed' into $\lambda(t)$).

Parametric proportional hazards model

$$\lambda(t|X) = \lambda(t)\exp(X\beta)$$

Here are the 'friends':

$$\Lambda(t|X) = \Lambda(t)\exp(X\beta)$$
$$S(t|X) = S(t)^{\exp(X\beta)}$$

- $\Lambda(t) =$ 'underlying' cumulative hazard function (for a subject with $X\beta = 0$).
- $S(t) =$ 'underlying' survival function (for a subject with $X\beta = 0$).

Parametric proportional hazards model

It is easiest to consider the log-model versions:

$$\log \lambda(t|X) = \log \lambda(t) + X\beta$$
$$\log \Lambda(t|X) = \log \Lambda(t) + X\beta$$
$$\log S(t|X) = \underbrace{\log S(t)}_{\text{time}} \times \underbrace{\exp(X\beta)}_{\text{predictors}}$$

► Observe that we separated the time and the predictors components.

Important consequence due to the separability of t and X:

► The effect of X is assumed to be the same at all values of t.
► I.e.: We assume no $t \times X$ interaction effect.

Parametric proportional hazards model

How to interpret regression coefficient β_j $(j = 1, \ldots, k)$?

$$\log \lambda(t|X) = \log \lambda(t) + (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k)$$
$$\log \Lambda(t|X) = \underbrace{\log \Lambda(t)}_{\text{time}} + \underbrace{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k)}_{\text{predictors}}$$

▶ Additive interpretation:
  ✓ $\log \lambda(t|X)$ increases by β_j units when X_j increases by 1 unit at any time point t, holding all the other predictors constant:

  $$\log \lambda(t|\ldots, X_j + 1, \ldots) = \log \lambda(t|\ldots, X_j, \ldots) + \beta_j.$$

  ✓ Same for $\log \lambda(t|X)$.

Parametric proportional hazards model

How to interpret regression coefficient β_j $(j = 1, \ldots, k)$?

$$\lambda(t|X) = \lambda(t) \quad \times \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k)$$
$$\Lambda(t|X) = \Lambda(t) \quad \times \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k)$$
$$\log S(t|X) = \underbrace{\log S(t)}_{\text{time}} \times \underbrace{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k)}_{\text{predictors}}$$

▶ Multiplicative interpretation:
  ✓ $\lambda(t|X)$ is multiplied by $\exp(\beta_j)$ units when X_j increases by 1 unit at any time point t, holding all the other predictors constant:

$$\frac{\lambda(t| \ldots, X_j + 1, \ldots)}{\lambda(t| \ldots, X_j, \ldots)} = \exp(\beta_j).$$

  ✓ Same for $\lambda(t|X)$.
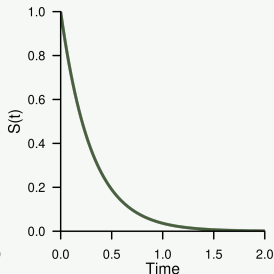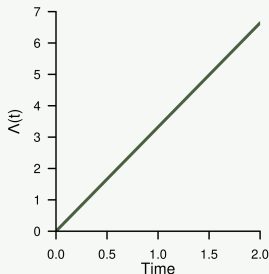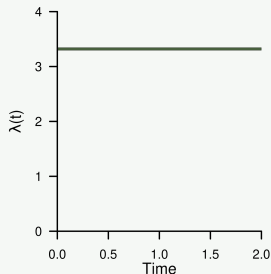 ✓ Same for $\log S(t|X)$.

Example: Exponential PH survival model

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k.$$

$$\lambda(t|X) = \exp(X\beta)$$
$$\Lambda(t|X) = t \exp(X\beta)$$
$$S(t|X) = \exp(-t)^{\exp(X\beta)}$$

Example: Weibull PH survival model

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_x X_k.$$

$$\lambda(t|X) = \gamma t^{\gamma-1} \exp(X\beta)$$
$$\Lambda(t|X) = t^\gamma \exp(X\beta)$$
$$S(t|X) = \exp(-t^\gamma)^{\exp(X\beta)}$$