An introduction to survival analysis

Georg Wölflein

School of Computer Science, University of St Andrews

April 11, 2022

University of St Andrews

Contents

1 Time-to-event data

2 Survival function

3 Kaplan-Meier estimator

4 Hazard function

5 Cox's proportional hazards model

What is time-to-event (TTE) data?

We can measure **time** in:

- years
- months
- seconds

The **event** could be:

- death from disease
- product failure
- losing a customer

must be a binary variable

TTE data consists of $(\textit{time}, \overbrace{\textit{event}}^{\text{yes/no}})$ tuples.

Time-to-event (TTE) data

TTE analysis is also known as:

- survival analysis
- failure time analysis
- reliability theory (engineering)
- duration modelling (economics)
- event history analysis (sociology)

Use cases for TTE analysis:

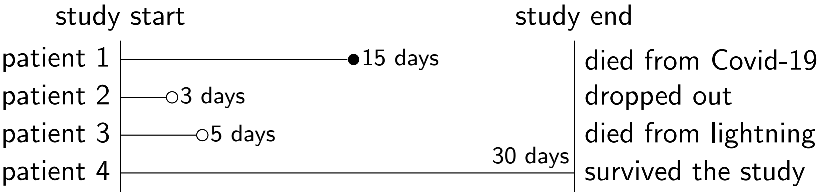
- clinical research
- customer analytics (churn)
- hardware (equipment failure)

Example: Covid-19 treatment trial

A randomised controlled trial ($n = 4$) was conducted to assess the efficacy of drug ABC in treating Covid-19. This is what happened to the patients:

patient	received ABC?	outcome
1	yes	died from Covid-19 on day 15
2	no	dropped out of the study after day 3
3	yes	died by a lightning stroke on day 5
4	no	survived the study (30 days)

Example: Covid-19 treatment trial



study start

study end

patient 1 — 15 days — died from Covid-19

patient 2 — 3 days — dropped out

patient 3 — 5 days — died from lightning

patient 4 — 30 days — survived the study

The **time** is the number of days since testing positive for Covid-19.

The **event** is whether the patient died due to Covid-19.

Time-to-event data

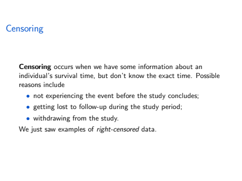
patient	time	event
1	15	yes
2	$[0,3]$	no
3	$[0,5)$	no
4	$[0,30]$	no

Censoring

Censoring occurs when we have some information about an individual's survival time, but don't know the exact time. Possible reasons include

- not experiencing the event before the study concludes;
- getting lost to follow-up during the study period;
- withdrawing from the study.

We just saw examples of *right-censored* data.



Censoring

Censoring occurs when we have some information about an individual's survival time, but don't know the exact time. Possible reasons include

- not experiencing the event before the study concludes;
- getting lost to follow-up during the study period;
- withdrawing from the study.

We just saw examples of *right-censored* data.

Left censoring happens if the individual observed the event before the start of the study. This is often very hard to deal with and therefore not included in the study.

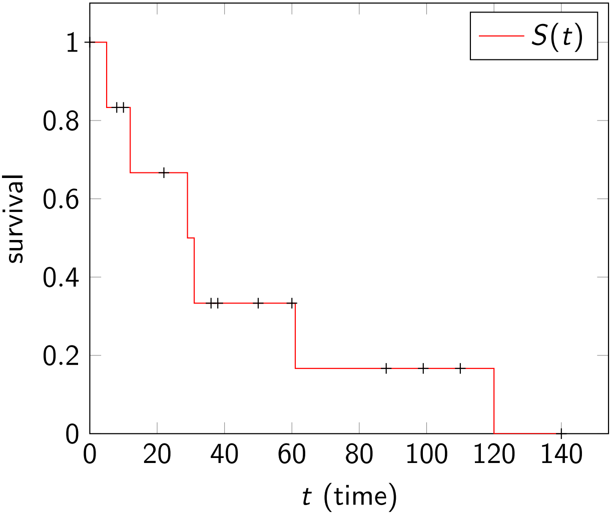
Survival function

Let T be a continuous random variable representing survival time. The **survival function** $S(t)$ is the probability that an individual will survive past time t.

Survival function

$$S(t) = \Pr\left(T > t\right)$$

Survival curve



Modelling the survival function

The **Kaplan-Meier estimator** provides a non-parametric estimate of the survival function $S(t)$ using the survival curve.

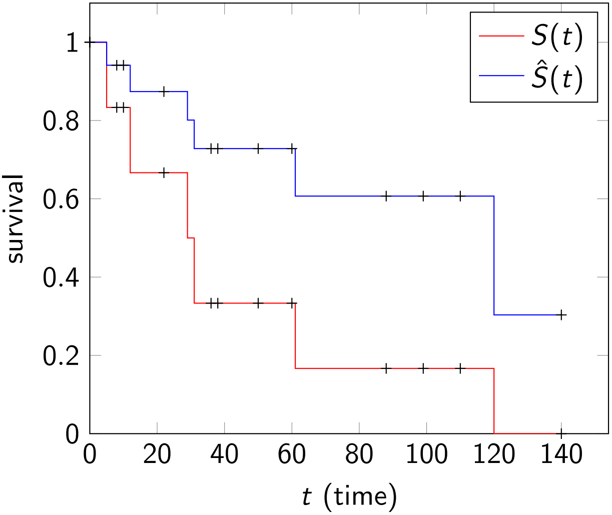
Kaplan-Meier estimator

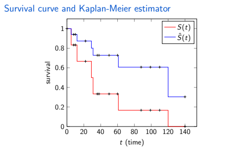
$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where

- t_i is an event time
- d_i is the number of deaths at time t_i
- n_i is the number of individuals *known to have survived* until t_i

Survival curve and Kaplan-Meier estimator





Survival curve and Kaplan-Meier estimator

- When there is no censoring, $S(t) = \hat{S}(t)$.
- Commonly used to compare two study populations.
- Does not control for covariates.

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\begin{aligned}\lambda(t) &= \lim_{dt \to 0} \frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt} \\ &= \lim_{dt \to 0} \frac{\Pr(t \leq T \leq t + dt)}{dt \cdot S(t)}\end{aligned}$$

What does survival depend on?

Recall the survival function $S(t) = \Pr(T > t)$ as the probability that an individual will survive past time t. Let's assume that $S(t)$ depends on

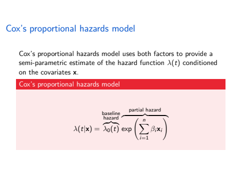
1. the **baseline hazard function** (how risk of event occurence changes over time at baseline covariates); and
2. the **effect parameters** (how hazard varies due to the covariates), also known as the *partial hazard*.

Cox's proportional hazards model

Cox's proportional hazards model uses both factors to provide a semi-parametric estimate of the hazard function $\lambda(t)$ conditioned on the covariates $\mathbf{x}$.

Cox's proportional hazards model

$$\lambda(t|\mathbf{x}) = \overbrace{\lambda_0(t)}^{\substack{\text{baseline} \\ \text{hazard}}} \overbrace{\exp\left(\sum_{i=1}^{n} \beta_i \mathbf{x}_i\right)}^{\text{partial hazard}}$$



Cox's proportional hazards model

Cox's proportional hazards model uses both factors to provide a semi-parametric estimate of the hazard function $\lambda(t)$ conditioned on the covariates $\mathbf{x}$.

Cox's proportional hazards model

$$\lambda(t|\mathbf{x}) = \overbrace{\lambda_0(t)}^{\substack{\text{baseline}\\\text{hazard}}} \overbrace{\exp\left(\sum_{i=1}^{n} \beta_i \mathbf{x}_i\right)}^{\text{partial hazard}}$$

- $\lambda_0(t)$ is a population-level baseline hazard that changes over time (for a reference individual with zeroed covariates).
- The partial hazard is a linear function of the covariates that is exponentiated. Each coefficient β_i is the relative risk associated with covariate $\mathbf{x}_i$.

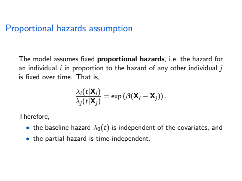
Proportional hazards assumption

The model assumes fixed **proportional hazards**, i.e. the hazard for an individual i in proportion to the hazard of any other individual j is fixed over time. That is,

$$\frac{\lambda_i(t|\mathbf{X}_i)}{\lambda_j(t|\mathbf{X}_j)} = \exp\left(\beta(\mathbf{X}_i - \mathbf{X}_j)\right).$$

Therefore,

- the baseline hazard $\lambda_0(t)$ is independent of the covariates, and
- the partial hazard is time-independent.



Proportional hazards assumption

The model assumes fixed **proportional hazards**, i.e. the hazard for an individual i in proportion to the hazard of any other individual j is fixed over time. That is,

$$\frac{\lambda_i(t|\mathbf{X}_i)}{\lambda_j(t|\mathbf{X}_j)} = \exp\left(\beta(\mathbf{X}_i - \mathbf{X}_j)\right).$$

Therefore,

- the baseline hazard $\lambda_0(t)$ is independent of the covariates, and
- the partial hazard is time-independent.

The so-called **extended Cox model** allows the partial hazard to vary with time, thus no longer satisfying the proportional hazards assumption.

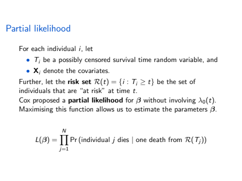
Partial likelihood

For each individual i, let

- T_i be a possibly censored survival time random variable, and
- $\mathbf{X}_i$ denote the covariates.

Further, let the **risk set** $\mathcal{R}(t) = \{i : T_i \geq t\}$ be the set of individuals that are "at risk" at time t.
Cox proposed a **partial likelihood** for $\boldsymbol{\beta}$ without involving $\lambda_0(t)$.
Maximising this function allows us to estimate the parameters $\boldsymbol{\beta}$.

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \Pr\left(\text{individual } j \text{ dies} \mid \text{one death from } \mathcal{R}(T_j)\right)$$



Partial likelihood

For each individual i, let

- T_i be a possibly censored survival time random variable, and
- $\mathbf{X}_i$ denote the covariates.

Further, let the **risk set** $\mathcal{R}(t) = \{i : T_i \geq t\}$ be the set of individuals that are "at risk" at time t.

Cox proposed a **partial likelihood** for $\boldsymbol{\beta}$ without involving $\lambda_0(t)$. Maximising this function allows us to estimate the parameters $\boldsymbol{\beta}$.

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \Pr\left(\text{individual } j \text{ dies} \mid \text{one death from } \mathcal{R}(T_j)\right)$$

- $L_j(\beta)$ is a *partial* likelihood because it considers only patients who died, not those that are censored.

Partial likelihood formula

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{j=1}^{N} \Pr\left(\text{individual } j \text{ dies} \mid \text{one death from } \mathcal{R}(T_j)\right) \\
&= \ldots \\
&= \prod_{j=1}^{N} \frac{\lambda(T_j|\mathbf{X}_j)}{\sum_{k\in\mathcal{R}(T_j)} \lambda(T_j|\mathbf{X}_k)} \\
&= \prod_{j=1}^{N} \frac{\lambda_0(T_j)\exp\left(\boldsymbol{\beta}\mathbf{X}_j\right)}{\sum_{k\in\mathcal{R}(T_j)} \lambda_0(T_j)\exp\left(\boldsymbol{\beta}\mathbf{X}_k\right)} \\
&= \prod_{j=1}^{N} \frac{\exp\left(\boldsymbol{\beta}\mathbf{X}_j\right)}{\sum_{k\in\mathcal{R}(T_j)} \exp\left(\boldsymbol{\beta}\mathbf{X}_k\right)}
\end{aligned}
$$

Parameter estimation

We can estimate the parameters $\boldsymbol{\beta}$ by minimizing the negative partial log-likelihood, i.e. $-\log L(\boldsymbol{\beta})$, by taking the partial derivatives with respect to the parameters $\boldsymbol{\beta}$ and solving for the minimum using e.g. the Newton-Raphson algorithm.

Hazard ratios

The fraction used to express the proportional hazards assumption is actually the **hazard ratio**, measuring the risk of individual i relative to individual j:

$$HR = \frac{\lambda(t|\mathbf{X}_i)}{\lambda(t|\mathbf{X}_j)} = \exp\left(\beta(\mathbf{X}_i - \mathbf{X}_j)\right).$$

We may be interested in the relative risk associated with a particular covariate c, specifically the risk of said covariate having value c_i compared to c_j. Consider two dummy individuals i and j differing only in the c^{th} covariate, i.e. $\mathbf{X}_{i,k} = \mathbf{X}_{j,k}$ for $k \neq c$. Then the relative risk associated with c_i compared to c_j is

$$HR = \exp\left(\beta_c(c_i - c_j)\right).$$

Interpretation of hazard ratios

- $HR = 1$: no effect
- $HR > 1$: increase in hazard
- $HR < 1$: reduction in hazard