Survival analysis

Thomas Scheike

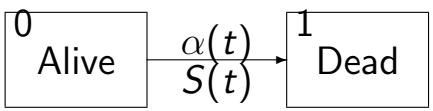
- Examples of Survival outcome
  - time to event T
    - Medicine (Death).
    - Lifetimes, Insurance, Labour market
    - Claims, time of even
- Incomplete observations
  - right censoring
  - left truncation

T = time to event

- Time scale choosen

Survival Function $P(T > t) = P(T > t)$

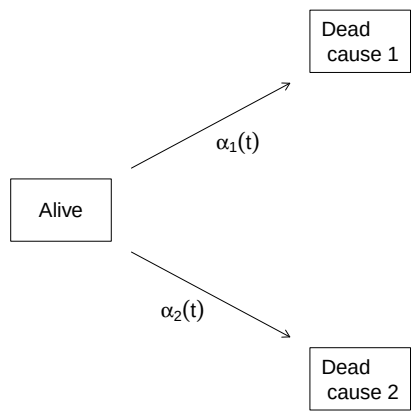
Survival model



- hazard, $\alpha(t)$, instantaneous risk of dying at time t among those that are alive (at risk).
- survival function, $S(t)$, probability of surviving t time-units.
- these two functions/quantities contain same information. That is knowing $S(\cdot)$ and $\alpha(\cdot)$ gives same information.

Competing Risks Model

If several causes of death are studied then the {*competing risks*} *model below is relevant. Here, transitions from the state 0: "alive" to the state h: "dead, cause h" is governed by* $\alpha_h(t), h = 1, \cdots, k$, *where the probability of moving to state "h" at from time t to* $t + \Delta t$

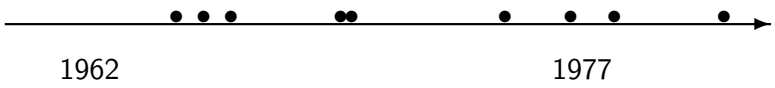


Malignant melanoma

In the period 1962-77 205 patients had their tumor removed and were followed until 1977.
At the end of 1977:

- 57 died of mgl. mel. (status=1)
- 14 died of non-related mgl. mel. (status=3)
- 134 were still alive. (status=2)

Purpose: Study effect on survival of sex, age, thickness of tumor, ulceration, etc.



Malignant melanoma

N	time	status	sex	age	year	thickness	ulcer
1	10	3	1	76	1972	6.76	1
2	30	3	1	56	1968	0.65	0
3	35	2	1	41	1977	1.34	0
4	99	3	0	71	1968	2.90	0
5	185	1	1	52	1965	12.08	1
6	204	1	1	28	1971	4.84	1
7	210	1	1	77	1972	5.16	1
8	232	3	0	60	1974	3.22	1
9	232	1	1	49	1968	12.88	1
10	279	1	0	68	1971	7.41	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
203	4688	2	0	42	1965	0.48	0
204	4926	2	0	50	1964	2.26	0
205	5565	2	0	41	1962	2.90	0

Survival

Looking at the data we note
Interest is on length of survival

- time to death from either cause
- time to death from malignant melanom
- some subjects are alive at end of follow-up

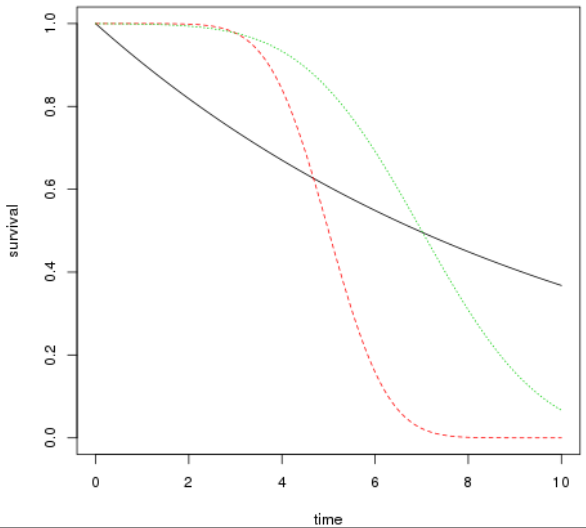
The Survival function

Instead of usual procedures for quantitative data (means), consider the survival function, and try to estimate this population parameter. Based on this can compute other measure such as median survival, and survival probability at different time points. Residual mean life, or residual restricted mean life.
Let T be a survival time. The survival function is

$$\begin{aligned} S(t) &= P(T > t) \\ &= \text{probability of being alive at time } t. \end{aligned}$$

- $S(t) \geq 0$ for alle $t \geq 0$.
- decreasing
- $S(0) = 1$
- $S(\infty) = 0$
- Followup typically limitted so do not know tail of distribution, so mean is difficult to compute

Survival function



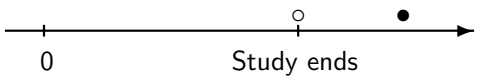
Right Censoring

Let T be the lifetime of interest. For survival data we need to apply techniques that take censoring into account. That is we do typically not observe T fully due to limitted follow-up.

- Right-censoring: Patient still alive at end of follow-up. know that T is larger than follow-up

In practice

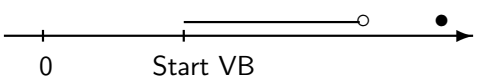
- some subjects die in follow-up time.
- some are alive at end of follow-up.
- some leave the study due to unrelated reasons.



Truncation

Recruit patients that are alive for survival study. Many cohorts of this type: Østerbro study (recruit representative cohort of people with age 50, 60, 70).

- Observe only those that are alive (and under risk) when included at a given time (typically age time-scale).
- Østerbro cohorts, time to death from from discharge, age at divorce (given marriage).
- Immortal time bias if forgotten



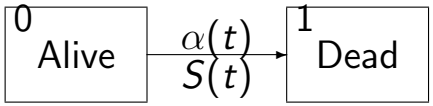
Left truncation is also called delayed entry.

Interval censoring

Time is only know to happen in interval

- special methods for this type of data
- basic quantities are the same
- requires special methods but basic principles are the same as for right-censoring, left-truncation

Survival model



- hazard, $\alpha(t)$, instantaneous risk of dying at time t for those alive
- survival function, $S(t)$, probability of surviving t time-units
- cumulative hazard $A(t) = \int_0^t \alpha(s)ds$
- $S(t) = \exp(-A(t))$

Survival model

Want to describe $S(t)$

- covariate effects on $S(t)$
- How many survive 5 years
- mean life
- residual mean life $E(T|T > t) - t$

Survival model

Hazard function

$$\alpha(t) = \frac{f(t)}{S(t)} = lim_h \frac{1}{h} P(t < T \ \leq t + h)/P(T > t)$$
$$= lim_h \frac{1}{h} P(t < T \ \leq t + h|T > t)$$

From

- $\alpha(t) = f(t)/S(t) = -D_t log(S(t))$

and $S(0) = 1$ it follows that

- $S(t) = \exp(-A(t))$

with $A(t) = \int_0^t \alpha(s)ds$ the cumulative hazard.

- $A(t)$ is easy to estimate nonparametrically

Hazard characterizes distribution.

Now right censored survival time, censored by C censoring time

- survival time $T^* \sim S(t)$
- C censoring time $C \sim S_c(t)$

Observe $T = \min(T^*, C)$ and $\delta = I(T^* \leq C)$.

- Simplest case $C \equiv \tau$.

density $f(t, \theta)$, likelihood of observed iid data

$$\prod_i f(T_i)^{\delta_i} S(T_i)^{1-\delta_i}$$

MLE can be found

- standard MLE theory applies
- clearly best to consider only $S(t, \hat{\theta})$ on $[0, \tau]$.

Survival model

Distribution of $T^* \sim S(t)$, $F(t) = P(T^* \leq t) = 1 - S(t)$.

- smooth density of $F(t) \sim f(t)$, $F(t) = \int_0^t f(s)ds$.

Most modellling done via hazard function

- $\alpha(t) = f(t)/S(t) = -D_t log(S(t))$

Likelihood now

$$\prod_i \alpha(T_i)^{\delta_i} S(T_i) = \prod_i \alpha(T_i)^{\delta_i} \exp(-\int_0^{T_i} \alpha(s)ds)$$

interpretation:

- contribution from risk time $S(T_i)$ survives $[0, T_i]$
  - exponential of hazard for risk periods
- death given survival $\alpha(T_i)$

Survival model

Important simple models

- exponential model $\alpha(t, \theta) \equiv \theta$
- piecewise exponential model, piecewise constant hazards model, $\alpha(t, \theta) = \sum_j I(t \in I_j)\theta_j$ for partioning of time-axis, I_j.
- weibull $\alpha(t, \theta) = \lambda\beta(\lambda t)^{\beta-1}$
- Comperz-Makeham (life-insurance).

With covariates, the Cox model,

$$\begin{aligned}\alpha(t, X) &= lim_h \frac{1}{h} P(t < T^* \leq t + h | X, T > t) \\ &= \lambda_0(t) \exp(X^T \beta)\end{aligned}$$

Survival model

MLE for exponetial model

$$\hat{\lambda} = \frac{D}{T_\bullet} = \frac{\text{occurrence}}{\text{exposure}}$$

with $D = \sum_i \delta_i$, and $T = \sum_i T_i$
central principle for survival analysis method, risk-time versus number of events.

Survival model , censoring

- survival time $T^* \sim S(t) \sim \alpha(t)$
- C censoring time $C \sim S_c(t) \sim \alpha_c(t)$

Observe $T = \min(T^*, C)$ and $\delta = I(T^* \leq C)$.
Need that $C \perp T^*$ to be able to estimate $S(t)$.
Likelihood

$$\prod_i \alpha(T_i)^{\delta_i} S(T_i) S_c(T_i) \alpha_c(T_i)^{1-\delta_i}$$

for the two types of observations

- $P(\delta = 0, T = t) = P(C = t, T > t) = S(T) \times S_c(T)\alpha_c(T)$
- $P(\delta = 1, T = t) = P(T = t, C > t) = \alpha(T)S(T) \times S_c(T)$

using independence

Survival model , censoring

Likelihood for parameters of interest thus becomes

$$\prod_i \alpha(T_i)^{\delta_i} S(T_i)$$

- S_c does not depend on θ
- conditioning on C_i.
- Assumed $C \perp T^*$ but can not check this in data.
- Can only estimate $S(t)$ under independent censoring.

Survival model , censoring

Hazard of T^* in observed data is

$$\alpha_o(t) = lim_h \frac{1}{h} P(t < T^* \leq t+h | T^* > t, C > t)$$

and hazard from $S(t)$ is

$$\alpha(t) = lim_h \frac{1}{h} P(t < T^* \leq t+h | T^* > t)$$

the same under independence:

$$\begin{aligned}
\alpha_o(t) &= lim_h \frac{1}{h} P(t < T \leq t+h | T^* > t, C > t) \\
&= lim_h \frac{1}{h} P(t < T^* \leq t+h, C > t) / P(T^* > t, C > t) \\
&= lim_h \frac{1}{h} P(t < T^* \leq t+h) S_c(t) / (P(T^* > t) S_c(t)) \\
&= lim_h \frac{1}{h} P(t < T^* \leq t+h) / P(T^* > t) = \alpha(t)
\end{aligned}$$

Survival model , Truncation

Another important incompleteness is that we might only see T^* for those where $T^* > L$.

- $T^* \sim \alpha(t)$
- L truncation time
- delayed entry at L.

We need that $L \perp T^*$ (possibly conditional on covariates).
We observed $T^* | T^* > L, L$ conditioning also on L, that is from conditional distribution $P^L() = P(| T^* > L, L)$.
Hazard of observed survival times at $t > L$

$$\begin{aligned}
\alpha_L(t) &= lim_h \frac{1}{h} P^L(t < T^* \leq t+h | T^* > t) \\
&= lim_h \frac{1}{h} P^L(t < T^* \leq t+h) / P^L(T^* > t) \\
&= lim_h \frac{1}{h} P(t < T^* \leq t+h) / P(T^* > t) \qquad = \alpha(t)
\end{aligned}$$

Survival model , Truncation + Censoring

- $T^* \sim \alpha(t)$
- $C \sim \alpha_c(t)$
- L truncation time
- Observe conditional on $T^* > L$.
  - $T = \min(T^*, C)$
  - $\delta = I(T^* \leq C)$.

Hazard of delayed entry T^* is the sam as the one from T^* and the likelihood of interest

$$\prod_i \alpha(T_i)^{\delta_i} \exp(-\int_{L_i}^{T_i} \alpha(s) ds)$$

since $P(T_i^* > t | T_i^* > L_i, L_i) = \exp(-\int_{L_i}^{T_i} \alpha(s) ds)$

Survival model , Truncation + Censoring

MLE for exponetial model

$$\hat{\lambda} = \frac{D}{T_\bullet} = \frac{\text{occurrence}}{\text{exposure}}$$

with $D = \sum_i \delta_i$, and $T = \sum_i (T_i - L_i)$
Repeated independent truncation and censoring also leads to MLE with

- exponential of hazard for exposure periods
- $\alpha(T_i)$ for event time
- MLE for exponential same form

Survival model, dependent censoring

- Now do not assume dependence but assume that $T, C \sim G(t, c) = P(T > t, C > c)$, bivariate survival function.

Observed hazards are

$$\alpha_o(t) = lim_h \frac{1}{h} P(t < T \ \leq t + h | T > t, C > t)$$
$$\alpha_{co}(t) = lim_h \frac{1}{h} P(t < C \ \leq t + h | T > t, C > t)$$

These hazards are one we see in data.
We see again on

- $\tilde{T} = \min(T, C)$
- $\delta = I(T \leq C)$.

Survival model, dependent censoring

likelihood

$$\alpha_o(\tilde{T}_i)^{\delta_i} \exp(-\int_0^{\tilde{T}_i} \alpha_o(s)ds) \times \alpha_{co}(\tilde{T}_i)^{1-\delta_i} \exp(-\int_0^{\tilde{T}_i} \alpha_{co}(s)ds)$$

We start by noting that if we have T and C independent and with

- $T \sim \alpha_o(t)$
- $C \sim \alpha_{co}(t)$

These have the same observed hazards and likelihood

$$\alpha_o(T_i)^{\delta_i} \exp(-\int_0^{T_i} \alpha_o(s)ds) \times \alpha_{co}(T_i)^{1-\delta_i} \exp(-\int_0^{T_i} \alpha_{co}(s)ds)$$

Survival model, dependent censoring

Conclusion we can not check independence in data.

- Any set of marginals can be observed also in independent data.
- need to construct experient so that independence is full-filled.
  - in medical studies we typically (try to) decide how the censoring is done by design.

Survival model, dependent censoring

- Now do not assume dependence but assume that $T, C \sim G(t,c) = P(T > t, C > c)$, bivariate.
- $D_1 G(t,c)$ first partial derivative after t
- $D_2 G(t,c)$ first partial derivative after c
- $G(t,c)$

Likelihood

$$(-D_1 G(\tilde{T}, \tilde{T}))^{\delta} \times (-D_2 G(\tilde{T}, \tilde{T}))^{1-\delta}$$
$$= (\frac{-D_1 G(\tilde{T}, \tilde{T})}{G(\tilde{T}, \tilde{T})})^{\delta} \times (\frac{-D_2 G(\tilde{T}, \tilde{T})}{G(\tilde{T}, \tilde{T})})^{1-\delta} \times G(\tilde{T}, \tilde{T})$$

and

$$= \frac{-D_1 G(t,t)}{G(t,t)} = \lim_h \frac{1}{h} P(T \in [t, t+h] | T > t, C > t) = \alpha_o(t)$$

Survival model, dependent censoring

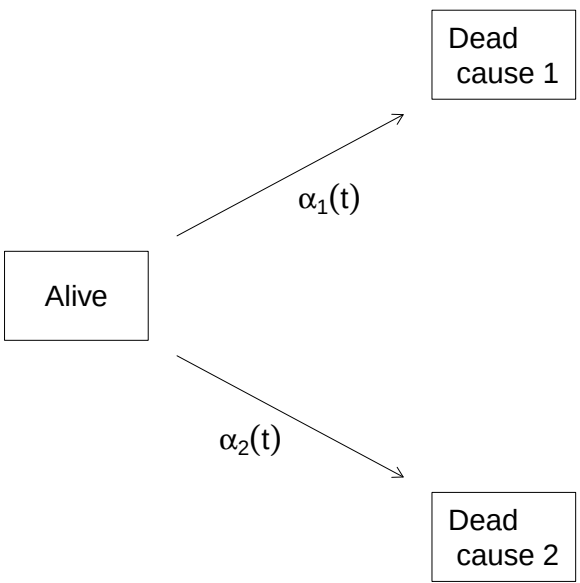
So likelihood becomes

$$\alpha_o(\tilde{T}_i)^{\delta_i} \exp(-\int_0^{\tilde{T}_i} \alpha_o(s)ds) \alpha_{co}(\tilde{T}_i)^{1-\delta_i} \exp(-\int_0^{\tilde{T}_i} \alpha_{co}(s)ds)$$

That the stochastic variable $\tilde{T} = \min(T, C)$ has hazard $\alpha_o(t) + \alpha_{co}(t)$ follows from direct calculations.
And then

$$G(t,t) = P(\min(T, C) > t) = P(T > t, C > t)$$
$$= \exp(-\int_0^t \alpha_o(s) + \alpha_{co}(s)ds)$$

Survival model, Competing risks



Competing risks

Here same methods and arguments can be used to estimate $\alpha_1(t)$ and $\alpha_2(t)$ under independent right censoring.
One way of thinking about this model, is via latent variables

- T_1 time of death of type 1.
- T_2 time of death of type 2.

We observe $T = \min(T_1, T_2)$ and might in addition have right censoring present.
Here T_1 and T_2 will typically not be independent and we are back in dependent censoring case.

Survival model, Competing risks

Observed cause specific hazards are

$$\alpha_1(t) = lim_h \frac{1}{h} P(t < T_1 \ \leq t + h | T_1 > t, T_2 > t)$$
$$\alpha_2(t) = lim_h \frac{1}{h} P(t < T_2 \ \leq t + h | T_1 > t, T_2 > t)$$

With additional censoring calculations goes as earlier and these are also seen in data.

Survival model, Summary

- survival data with incomplete observations
  - hazard unchanged by independent right censoring
  - hazard unchanged by independent left truncation
- hazard is key for modelling of survival data
  - give instantenous risk of event given at risk
  - characterizes distribution
  - survival function or other relevant measures can be computed based on hazard