Traditional (biostatistical) approaches to event-time prediction

Jolien Cremers

NationScale Fall Meeting: September 10th, 2021

Research questions of interest

Time until a certain event of interest takes place

- What is the probability that the event does (not) take place in the next X years?
- Are there differences in event timing between groups?
- How do certain personal, behavioral or other characteristics affect when the event takes place in an individuals' life?

Time-to-event analysis: basic quantities

- T = time of the event, random variable with density function $f(t)$
- C = censoring time
- Survival = probability of surviving up to time t, a "risk": $S(t) = P(T > t)$
- Hazard = instantaneous probability of event at time t given one has not experienced the event before, a "rate": $\lambda(t) = P(T \in (t, t + \Delta t) \mid T \leq t)$
- Hazard has a direct relation to the survival: $\lambda(t) = \frac{f(t)}{S(t)}$

Time-to-event analysis: the proportional hazards model

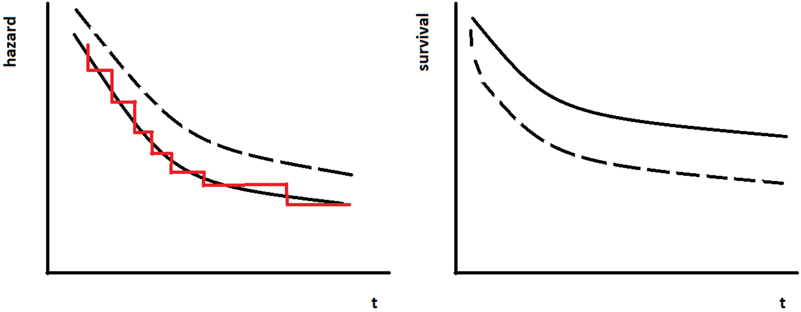
$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{s}_i(t)'\boldsymbol{\gamma})$$

- $\lambda_i(t) =$ hazard at time t for individual i
- $\lambda_0(t) =$ baseline hazard at time t
- $\boldsymbol{s}_i(t) =$ vector of covariates at time t for individual i
- $\boldsymbol{\gamma} =$ vector of coefficients

$\exp(\boldsymbol{\gamma})$, interpretation as a relative risk

Can be written as a poisson model for (piecewise constant) hazard rates. Convenient for ML methods (see e.g. Bender, Rügamer, Scheipl, & Bischl (2020)).

Visualization



Modelling issues/challenges

1. Censoring
   - Hazard can be computed for right censored data (uses individuals at risk).
   - Maybe less of a problem in register data, but immigration/emigration.
   - Under right censoring, using a logistic regression model results in biased estimates.
   - Dependent, left, interval censoring
2. Competing risks
   - An event that either hinders the observation of the event of interest or modifies the chance that the event occurs.
   - Dying twice.
   - No one-to-one relation between hazard and survival of a single event (Andersen, Geskus, Witte, & Putter (2012)).

Modelling issues/challenges (cont.)

3. Recurrent events
   - Repeated observations of the same event, e.g. doctor visits, birth of children, change of job
4. Time varying covariates
5. Time varying effects (γ_t)
6. Endogeneity of covariates
   - Bias if association between covariate and event is bidirectional.
   - Solution: joint models (economics: simultaneous equations, simulatneous hazard), with multiple outcomes (time-to-event and longitudinal)
7. Causality
8. Including lots and lots of covariates and (non-linear) associations for many individuals over a long time period

Time-to-event analysis: basic quantities (with equations)

- T time of the event, r.v. with density function $f(t)$
- C censoring time
- Probability that event has occured by time t, c.d.f.:
  $P(T \leq t) = \int_0^t f(t)dt$
- Survival = probability of surviving up to time t:
  $S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(t)dt = \exp(-\int_0^t \lambda(t)dt)$
- Hazard = instantaneous probability of event at time t given one has not experienced the event before:
  $\lambda(t) = P(T \in (t, t + \Delta t) \mid T \leq t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t)$
- Cumulative hazard, $\Lambda(t) = \int_0^t \lambda(t)dt$

References

Andersen, P. K., Geskus, R. B., Witte, T. de, & Putter, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, *41*(3), 861–870. doi:10.1093/ije/dyr213

Bender, A., Rügamer, D., Scheipl, F., & Bischl, B. (2020). A general machine learning framework for survival analysis. Retrieved from http://arxiv.org/abs/2006.15442