Survival analysis: censoring and truncation

• Traditionally research in event history analysis has focused on situations where the interest is in a single event for each subject under study. This is called survival analysis.
• A survival time is the time elapsed from an initial event to a well-defined end-point, e.g.
  – From birth to death (time=age)
  – From birth to cancer diagnosis (time=age)
  – From cancer diagnosis to death (time=disease duration)
• A special feature of survival data is that we usually have incomplete observation of the survival times due to censoring and truncation.

The most common form of incompleteness is right censoring: for some individuals we only know that their true survival times are larger than certain censoring times.

Censoring may be due to a number of reasons: termination of study (cf. figure below), withdrawals, lost to follow-up, etc.

Survival analysis: Examples

Lung cancer data:
• 272 non-small-cell lung cancer patients in Yorkshire followed for 3 years in the early 1990s.
• Observe time to cancer death or censoring (17%).
• Covariates:
  – Age: in years
  – Sex: 0=M, 1=F
  – Activity score: 0-4
  – Anorexia: 0=absent, 1=present
  – Hoarseness: 0=absent, 1=present
  – Metastases: 0=absent, 1=present

Leukemia data:
• 1043 cases of acute myeloid leukaemia recorded in the North West Leukaemia Register between 1982 and 1998.
• Observe death or censoring (16%).
• Covariates:
  – Age: in years
  – Sex: 0=F, 1=M
  – White blood cell count: (50x10^9/L), truncated at 500
  – Deprivation score: a measure of poverty/affluence for the residential location (low is good)
Survival analysis: concepts

Uncensored survival time $T$ (assumed absolutely continuous).

Survival function: $S(t) = P(T > t)$

Hazard rate: $\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T < t + \Delta t \mid T \geq t)$

The hazard rate is the instantaneous probability of the event per unit of time.

Heuristically: $\alpha(t) \Delta t = P(T < t + \Delta t \mid T \geq t)$

Survival analysis: clustered data

Usually survival data are assumed independent. One important exception is when subjects are grouped into clusters (e.g. twins, litters, organs of an individual).

The Diabetic Retinopathy Study:
- 179 patients with diabetic retinopathy
- Laser photocoagulation randomly assigned to one eye of each patient
- Observe time to severe visual loss ("blindness") for each eye (may be censored)
- There are 179 clusters with 2 observations per cluster

Event history data

Connecting together several events for a subject as they occur over time yields event histories.

Events may be of the same type (recurrent events):
- Births for a woman
- Episodes of diarrhea for a child
- Recurrent tumours

The events may be of different types:
- Marriage, divorce, new marriage, etc.
- Cancer diagnosis, remission, relapse, death

Multistate models are used for analysis.

Survival analysis: some parametric models

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Hazard</th>
<th>Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\theta$</td>
<td>$e^{-\theta t}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\theta e^{\theta t}$</td>
<td>$e^{-(\theta) t^\theta}$</td>
</tr>
<tr>
<td>Gompertz</td>
<td>$\theta^\alpha e^{\beta t}$</td>
<td>$\exp{\theta(1-e^{\alpha \beta})}$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\theta^\alpha t^{\alpha-1} e^{-\theta t}/\Gamma(\alpha, \theta)$</td>
<td>$\Gamma(\alpha, \theta)/\Gamma(\alpha)$</td>
</tr>
</tbody>
</table>

(upper) incomplete gamma function

We will not consider parametric models any further.

Recurrent event data: examples

The bladder cancer study:
- 86 patients with superficial bladder tumours
- Tumours were removed, and the patients randomized to placebo or treatment by thiotepa
- Patients were followed up, and the recurrence of tumours were registered
Infant diarrhoea data:
- Data on episodes of diarrhoea on 926 children obtained from household surveys, October 2000 to January 2002, in Salvador, Brazil.
- Various covariates collected at beginning of study.
- Observations may be missing due to intermittent missingness and drop-out.

Follow-up information on 10 children:
- Under observation: x
- New episode: x
- Ongoing episode: x
- Drop-out: a

Recurrent event data: basic models
Two classic models for recurrent event data are the Poisson processes and the renewal processes.

In general one may obtain a model for recurrent events by specifying the intensity \( \alpha(t) \):
\[
\alpha(t) dt = P(\text{event in } [t, t+dt] \mid \text{past events})
\]
For a Poisson process \( \alpha(t) \) is a fixed function not depending on past events.
For a renewal process \( \alpha(t) = h(s(t)) \) where \( s(t) \) is the time since the last event and \( h(t) \) is a fixed function.

Multistate models: the Markov case
The survival analysis situation may be modelled by a Markov model with two states:

\[
\begin{array}{c c c}
0 & \text{Alive} & 1 \\
\alpha_{01}(t) & \text{Dead} \\
\end{array}
\]
\( \alpha_{01}(t) \) is the hazard rate or transition intensity.

With two or more causes of failure we get a model for competing risks:

\[
\begin{array}{c c c c}
0 & \text{Alive} & 1 & 2 \\
\alpha_{01}(t) & \text{Dead due to cancer} & \text{Dead cause of interest} \\
\alpha_{02}(t) & \text{Dead due to cardiovascular disease} & \text{Dead other causes} \\
\text{Dead due to other medical causes} & \text{Alcohol abuse, accidents, violence} \\
\end{array}
\]
\( \alpha_{01}(t) \) and \( \alpha_{02}(t) \) are the cause specific hazards or transition intensities (i.e. instantaneous probabilities of a transition per unit of time).
An illness-death model:

\[
\begin{array}{c}
0 \\
\text{Healthy} \\
\alpha_{01}(t) \\
\alpha_{02}(t) \\
1 \\
\text{Diseased} \\
\alpha_{12}(t) \\
2 \\
\text{Dead}
\end{array}
\]

We have a Markov process if the transition intensities do not depend on duration in a state.

We need to consider and relate models for three situations:

- Joint model for the complete event history data and the observation process
- Model for the complete event history data
- Model for the observed event history data
- Parameters of interest are defined for this model
- Statistical methods are derived and studied for this model

Explicit expressions for the transition probabilities \( P_{ij}(s,t) = P(\text{in state } h \text{ at time } t | \text{in state } h \text{ at time } s) \) are available for simple Markov models.

E.g. for competing risks (\( j = 1,2 \)):

\[
P_{0j}(s,t) = \exp\left\{-\int_s^t (\alpha_{01}(u) + \alpha_{02}(u)) \, du\right\}
\]

\[
P_{ij}(s,t) = \int_s^t P_{0j}(s,u) \alpha_{ij}(u) \, du
\]

In general there are no explicit expressions for the transition probabilities, but we will see later how the transition probability matrix may be estimated from the matrix of empirical transition intensities.

Model for complete event history data

We consider a cohort with \( n \) individuals, and let \( \alpha_i(t) \) be the intensity of the event of interest for the \( i \)-th individual:

\[
\alpha_i(t) \, dt = P(\text{event in } [t, t+dt) | H_{i-})
\]

Here \( H_i \) denotes all information on covariates and events that would have been available at \( t \) if all event histories had been completely observed (note that \( \alpha_i(t) \) equals 0 when the event cannot happen).

The \( \alpha_i(t) \) are our key model parameters, and an aim of a statistical analysis is to infer how these (and derived quantities) depend on covariates and vary over time.

Modelling event history data

We may be interested in studying simultaneously the occurrences of two or more types of events.

For now, however, we concentrate on one type of event, e.g. corresponding to a failure, a recurrent event or one specific transition in a multistate model.

When defining a general model for event history data, we need to take care of the fact the event history data are incompletely observed.

Joint model for the event history data and the observation process

We will allow the event histories to be left truncated and right censored, so an individual may enter the study at a time \( t > 0 \) (in the study time scale), and the observation of its event history may stop before death (or another terminal event).

We will also allow intermittent missingness, so that there may be intervals where events are not observed (cf. the diarrhoea data), and define the left-continuous observation process:

\[
Y_{i,obs}(t) = \begin{cases} 
1 & \text{observed “just before” time } t \\
0 & \text{not observed “just before” time } t 
\end{cases}
\]
Let $G_t$ denote all information that would have been available at $t$ on covariates, events and the observation process in the hypothetical situation where both the complete event histories and the observation process is observed.

Assume that the observation process is independent in the sense that:

$$P(\text{event in } [t, t + dt] | G_t) = P(\text{event in } [t, t + dt] | H_t)$$

Thus the likelihood of an event occurring for an individual is not influenced by whether the individual is observed or not.

**Modelling the observable data**

Denote by $F_t$ all information actually available to the researcher at $t$ (on covariates, events, entries, exits, etc).

Then for individual $i$:

$$P(\text{observe event in } [t, t + dt] | F_t) = P(\text{observe event in } [t, t + dt] | F_{t-})$$

$$= E[E[Y_{obs}(t) | \text{event in } [t, t + dt] | G_t] | F_{t-}]$$

$$= E[Y_{obs}(t)P(\text{event in } [t, t + dt] | G_t) | F_{t-}]$$

$$= E[Y_{obs}(t)P(\text{event in } [t, t + dt] | H_t) | F_{t-}]$$

$$= E[Y_{obs}(t)\alpha_i(t) | F_{t-}]$$

**Counting process formulation**

Let $N_i(t)$ count the observed occurrences of the event of interest for individual $i$ as a function of (study) time $t$.

$dN_i(t)$ is a binary variable (taking the values 0 or 1)

Thus

$$\lambda_i(t)dt = P(dN_i(t) = 1 | F_{t-}) = E(dN_i(t) | F_{t-})$$

Cumulative intensity process:

$$\Lambda_i(t) = \int_0^t \lambda_i(s)ds$$

Introduce $M_i(t) = N_i(t) - \Lambda_i(t)$

$$E(dM_i(t) | F_{t-}) = E(dN_i(t) - \lambda_i(t)dt | F_{t-})$$

$$= E(dN_i(t) | F_{t-}) - \lambda_i(t)dt$$

$$= \lambda_i(t)dt - \lambda_i(t)dt = 0$$

Thus $M_i(t)$ is a martingale.

Note: All martingales we will consider take the value 0 at $t = 0$, and hence have expected value 0 for all $t$.

At each time $t$ we have the decomposition

$$dN_i(t) = \lambda_i(t)dt + dM_i(t)$$

observation signal noise
Aggregated counting process:

\[ N(t) = \sum_{i=1}^{n} N_i(t) \]

Corresponding intensity process and martingale (by linearity of expectation):

\[ \lambda(t) = \sum_{i=1}^{n} \lambda_i(t) \quad M(t) = \sum_{i=1}^{n} M_i(t) \]

Decomposition:

\[ dN(t) = \lambda(t)dt + dM(t) \]