SURVIVAL ANALYSIS

ARTHUR ALLIGNOL

INTRODUCTION

INTRODUCTION

Survival analysis is used to analyse data for which the time until some event occurs is of interest. E.g.:

- Time from treatment begin to death in cancer patients
- Time from unemployment to employement
- Time to credit rating change
- Time to failure for car parts

Such data generally require special techniques for analysis for two main reasons:

1. Survival data are generally not symmetrically distributed
2. At the end of the study, some individuals may not have reached the endpoint of interest. The exact survival time is thus not known. We only know that it is larger than the amount of time the individual has been in the study. Such individuals are said to be *right-censored*

THE HAZARD AND SURVIVAL FUNCTIONS

THE HAZARD AND SURVIVAL FUNCTIONS

Let $T_1, T_2, \ldots, T_n$ be i.i.d survival times of n individuals. Instead of modelling T directly, we model functions thereof.

The hazard function:

The hazard function is defined as

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \in [t, t + \Delta t) | T \geq t)}{\Delta t}$$

A more intuitive form

$$\alpha(t)\mathrm{d}t = \Pr(T \in \mathrm{d}t | T \geq t)$$

Cumulative hazard function: $A(t) = \int_0^t \alpha(u)\mathrm{d}u$

THE SURVIVAL FUNCTION

The *survival function*, $S(t)$, is given by

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

where $F(t)$ is the cdf of T

Knowing the cumulative hazard suffices to recover the survival function

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

More interesting is to see that

$$\mathrm{d}A(u) = \alpha(u)\mathrm{d}u,$$

with $\mathrm{d}A(u) = A(u) - A(u-)$. Then

$$1 - \mathrm{d}A(u) = \Pr(T \geq u + \mathrm{d}u | T \geq u),$$

which are conditional survival probabilities.

THE SURVIVAL FUNCTION

Let partition the interval $[0, t]$ into subintervals of the form

$$0 = t_0 < t_1 < \cdots < t_j = t.$$

In order to survive from 0 to t, you need to survive the intermediate subintervals, thus

$$\begin{aligned}\Pr(T > t) &\approx \Pr(T > t_1)\Pr(T > t_2|T > t_1)\ldots\Pr(T > t_j|T > t_{j-1}) \\ &\approx (1 - \Delta A(t_1))(1 - \Delta A(t_2))\ldots(1 - \Delta A(t_j)) \\ &\approx \prod_{k=1}^{j}(1 - \Delta A(t_k))\end{aligned}$$

with $\Delta A(t_j) = A(t_j) - A(t_{j-1})$.

THE SURVIVAL FUNCTION

Increasing the number of time points towards infinity and the distance between each time point toward 0, one ends up with an infinite product, namely the *product integral* and we write $\prod$. Therefore,

$$S(t) = \prod_{u \in (0,t]} (1 - \mathrm{d}A(u)) \, .$$

CENSORING AND TRUNCATION

CENSORING

Right-censoring

The most common type of censoring. It happens when, e.g., an individual is still alive at the end of the study. Then we only observe

$$T = \min(X, C),$$

with C the censoring time and X the survival time.

- *Type I censoring*: a survival time X is observed if it is no larger than a prespecified censoring time c , otherwise we only know that X exceeds c
- *Random censoring*: similar except that we consider the c's to be the observed values of the random variables C's independent of the X's
- *Independent censoring*: censored individuals at t are representative of the individuals still at risk of an event at t givan the covariates

CENSORING

Left-censoring

Left-censoring occurs when an event is known to have happened before a certain time. An example is seroconversion in HIV studies.

Interval-censoring

Happens when an event is known to occur within an interval. Typical example is development of a disease between two visits.

TRUNCATION

Left-truncation/delayed entry

Left-truncation or delayed entry arises if the observation of an individual is conditional on the occurrence of a previous event other than the main event of interest, $L < T$, with L a left-truncation/delayed entry time.

It often happens if the natural timescale is to be used but time 0 precedes study entry.

- Time scale is time from diagnosis to death but individuals are sampled at a later time point. Those who died before will never be included in the study

Right-truncation

Individuals only included in a study if $T \leq R$, R a right-truncation time.

CENSORING AND TRUNCATION

→ *Truncation* is different from *censoring* in that censoring is a missing data problem, but truncation is a sampling issue.

→ *Choice of the time scale* might induce left-truncation (e.g., age as a time-scale)

In the following, we assume the censoring/truncation to be independent.

NON-PARAMETRIC ESTIMATION

COUNTING PROCESS

Consider n individuals with survival times $T_i, i = 1, \ldots, n$.

Define

$$N_i(t) = \mathbf{I}\{T_i \leq t\}$$

and

$$Y_i(t) = \mathbf{I}\{T_i \geq t\}$$

Then $N(t) = \sum_i N_i(t)$ counts the number of events up to t and $Y(t) = \sum_i Y_i(t)$ counts the number of individuals still under observation just before t.

COUNTING PROCESS

With left-truncation/right-censoring, the data are represented by (L, T, δ),

- $L < T$
- $T = \min(X, C)$, X and C are the survival and right-censoring times, respectively
- $\delta = 1$ if $T = X$.

The counting process and at-risk process are amended to

- $N_i(t) = \mathbf{I}\{T_i \leq t, \delta = 1\}$
- $Y_i(t) = \mathbf{I}\{L_i < T_i \leq t\}$

NON-PARAMETRIC ESTIMATION

Nelson-Aalen estimator

An estimator of $\alpha(t)\mathrm{d}t$ is given by

$$\frac{\#\text{ events at } t}{\#\text{ at risk just prior to } t} = \frac{\Delta N(t)}{Y(t)}$$

The *Nelson-Aalen estimator* of the cumulative hazard $A(t) = \int \alpha(u)\mathrm{d}u$ is

$$\hat{A}(t) = \sum_{s \leq t} \frac{\Delta N(s)}{Y(s)}$$

An estimator of the variance is given by

$$\hat{\sigma}_A^2 = \sum_{s \leq t} \frac{\Delta N(s)}{Y^2(s)}$$

NON-PARAMETRIC ESTIMATION

Kaplan-Meier estimator

An estimator of the survival function $S(t)$ is obtained by plugging in $\hat{A}(t)$ in the product integral

$$\hat{S}(t) = \prod_{u \in (0,t]} \left(1 - \mathrm{d}\hat{A}(u)\right)$$

With a finite number of event times, the product integral can be written as a finite matrix product

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \Delta\hat{A}(s)\right)$$

This estimator is called the *Kaplan-Meier* or *product-limit* estimator.

Greenwood's formula leads to a variance estimator

$$\sigma_S^2 = \hat{S}^2(t) \sum_{s \leq t} \frac{\Delta N(s)}{\Delta N(s)(\Delta N(s) - Y(t))}$$

ILLUSTRATION

Comparison of medical therapies to aid smokers quit

The aim of the study was to compare a triple medication vs. nicotine patch in smokers with medical illnesses. This was a randomised clinical trial.

Patients were followed up for up to 6 months.

The primary outcome was time from randomisation to relapse (return to smoking). Individuals who remained non-smokers for 6 months were censored.

```
data(pharmacoSmoking, package = "asaur")
head(pharmacoSmoking, n = 3)
```

```
   id ttr relapse         grp age gender  race employment yearsSmoking
1  21 182       0   patchOnly  36   Male white         ft           26
2 113  14       1   patchOnly  41   Male white      other           27
3  39   5       1 combination  25 Female white      other           12
  levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
1        heavy     21-49     35-49             0              0
2        heavy     21-49     35-49             3             90
3        heavy     21-49     21-34             3             21
```

ILLUSTRATION

Channing house retirement center

The data contain information on the age of death of 462 individuals who were in residence between Jan. 1964 and July 1975.

The age at which the individuals entered the centre is also available.

These data are better analysed using the age-scale, thus are left-truncated

```
data(channing, package = "KMsurv")
head(channing, n = 3)
```

```
  obs death ageentry  age time gender
1   1     1     1042 1172  130      2
2   2     1      921 1040  119      2
3   3     1      885 1003  118      2
```

COX PROPORTIONAL HAZARDS MODEL

THE COX PROPORTIONAL HAZARDS MODEL

The Cox proportional hazards model assumes that

$$\alpha(t|\mathbf{X}) = \alpha_0(t)\exp(\beta_1 X_1 + \cdots + \beta_q X_q)$$

where $\alpha_0(t)$ is the *baseline hazard function* (an unspecified, non-negative hazard function)

The interpretation of the parameter β_j is that $\exp(\beta_j)$ gives the relative risk change associated with an increase of one unit in covariate x_j, all other explanatory variables remaining constant. $\exp(\beta_j)$ is known as an *hazard ratio*

THE COX PROPORTIONAL HAZARDS MODEL

Written in this way we see that the model forces the hazard ratio between two individuals to be *constant over time* since

$$\frac{\alpha(t|\mathbf{X}_1)}{\alpha(t|\mathbf{X}_2)} = \frac{\exp(\beta\mathbf{X}_1)}{\exp(\beta\mathbf{X}_2)}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors of covariate values for two individuals.

Estimation of β obtained through maximisation of the *partial likelihood*

The baseline cumulative hazard can be estimated with the *Breslow estimator*

$$\hat{A}_0(t) = \sum_{s \leq t} \frac{\Delta N(s)}{\sum_i Y_i(t) \exp(\beta\mathbf{X}_i)}$$

THE COX PROPORTIONAL HAZARDS MODEL

The main assumptions of the Cox model are

- linearity and additivity of the predictors wrt log hazard
- *Proportional hazard assumption:* no time by predictor interactions, i.e., the effect of X is constant over time

STRATIFIED COX MODELS

$\approx$ adjust for factors that are not modelled/we don't want to model, e.g., clinical study site.

The idea of *stratified* Cox models is to allow the form of the underlying hazard function to vary across levels of the stratification factors:

$$\alpha_j(t|\mathbf{X}) = \alpha_{0;j}(t)\exp(\beta\mathbf{X})$$

where $\alpha_{0;j}$ is the baseline hazard for the j-th strata.

INFERENCE

Standard likelihood inference tests are also available for the Cox partial likelihood.

Wald test

$$\frac{\hat{\beta}}{\mathrm{se}\hat{\beta}} \sim \mathcal{N}(0,1)$$

Likelihood ratio test

$$2\left(l(\hat{\beta}) - l(\beta^{(0)})\right) \sim \chi^2$$

MODEL DIAGNOSTIC

No easy to use residuals as in linear models...

Testing the proportional assumtion

- With a binary variable, plotting the estimated cumulative hazards within one covariate level against the respective estimate within another covariate level should approximate a straight line if the assumption holds.
- Compute a time-dependent hazard ratio, with $\mathrm{E}(\text{scaled Schoenfeld residuals}) + \beta_j \approx \beta_j(t)$. If the assumption holds, $\beta_j(t) \approx$ a straight line

Influence measures

The *Martingale residuals* can be used to detect influencial individuals (that die too early or live too long)

They call also be used to assess the structural form of the model

ILLUSTRATION

Comparison of medical therapies to aid smokers quit

The objective is now to compare the two treatment therapies in terms of relapse and identify other factors related to this outcome

TIME-DEPENDENT VARIABLES

A time-dependent variable is a variable whose value changes over the course of time, e.g.,

- Multiple measurements of some labatory value
- Treatment/intervention change

The Cox model can readily include such a covariate, say, $X(t)$

$$\alpha(t|X(t)) = \alpha_0(t)\exp(\beta X(t))$$

TIME-DEPENDENT VARIABLES

In practice, the data need to be represented in the following way

id	start (()	stop (])	delta	X
1	0	5	0	10
1	5	10	0	20
2	0	3	0	20
2	3	7	1	40
3	0	10	1	30

ILLUSTRATION

This data set is a random sample drawn from the SIR-3 study that aimed at analysing the effect of nosocomial infections on the length of ICU stay.

The sample includes information to assess the effect of nosocomial pneumonia (a time-dependent variable) on the length of stay.

The endpoint is either discharge alive from the ICU or dead in the unit. These data are censoring complete as the censoring time is known for all patients.

```
data(icu.pneu, package = "kmi")
icu.pneu[icu.pneu$id == "3517", ]
```

```
      id start stop status event pneu adm.cens.exit      age sex
29 3517     0   20      0     3    0             20 71.13786   F
30 3517    20   70      1     3    1            329 71.13786   F
```