Survival Models

Lecture VIII. Summary

Censored data, where does it come from?

Censored data – the value of some observations is only partially known

Censored data, where does it come from?

Censored data – the value of some observations is only partially known

Question

When can this occur?

Censored data, where does it come from?

Censored data – the value of some observations is only partially known

Question

When can this occur?

For example:

- patients in clinical trials
  - can move away (no knowledge of their future situation)
  - can refuse the treatment (dropouts)
  - can survive until the end of trial (exact death time not known)
- reliability tests – systems/details survive the test period
- indemnity limits and reinsurance agreements in non-life insurance
- surveys/questionnaires often have questions that create censored results

Censored data, where does it come from?

Censored data – the value of some observations is only partially known

Question

When can this occur?

For example:

- patients in clinical trials
  - can move away (no knowledge of their future situation)
  - can refuse the treatment (dropouts)
  - can survive until the end of trial (exact death time not known)
- reliability tests – systems/details survive the test period
- indemnity limits and reinsurance agreements in non-life insurance
- surveys/questionnaires often have questions that create censored results

$\Rightarrow$ **the censoring problem is a common problem in practice**

Censored data, analysis problem

We get wrong/biased estimates if

- we ignore the censored observations – we throw away some information!
- we treat censored observations as exact – we create a biased sample!

Censored data, analysis problem

We get wrong/biased estimates if

- we ignore the censored observations – we throw away some information!
- we treat censored observations as exact – we create a biased sample!

$\Rightarrow$ **need for different approach**

Setup and main characteristics

- T – a (continuous) nonnegative rv (failure time)
- $F(t) = \mathbb{P}(T \leq t)$ – cdf of T
- $f(t) = F'(t)$ – pdf of T

Survivor function:

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

Hazard function:

$$h(t) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}\{t \leqslant T < t + \Delta t | T \geqslant t\}}{\Delta t} = \frac{f(t)}{S(t)}$$

Cumulative hazard function:

$$H(t) = \int_0^t h(u)\, du$$

Key relations

$$h(t) = -(\ln S(t))'$$

$$H(t) = \int_0^t h(u)du = -\ln(S(t))$$

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

$$f(t) = h(t)S(t) = h(t)\exp(-H(t)) = h(t)\exp(-\int_0^t h(u)du)$$

Empirical estimates

For survival function (Kaplan-Meier estimator):

$$\hat{S}(t) = \prod_{i:y_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$

For hazard function

$$\hat{h}(y_i) = \frac{d_i}{n_i}$$

For cumulative hazard function (Nelson-Aalen estimator):

$$\hat{H}(t) = \sum_{i:y_{(i)} \leq t} \frac{d_i}{n_i}$$

Parametric models. Location-scale family of distributions

Main characteristics:

- distribution of time T has two parameters, scale λ and shape α
- in log-time, $Y = \ln(T)$, the distribution has two parameters, location $\mu = -\ln(\lambda)$, scale $\sigma = \frac{1}{\alpha}$
- each rv Y can be expressed as $Y = \mu + \sigma Z$, where Z is the standard member, i.e. $\mu = 0$ $(\lambda = 1)$ and $\sigma = 1$ $(\alpha = 1)$
- the models built on T are log-linear

Members of this class:

T	$Y = \ln(T)$
Weibull	extreme minimum value
log-normal	normal
log-logistic	logistic

Cox PH model and AFT model

Cox proportional hazards PH model:

$$h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta})$$

- is built on hazard
- is semiparametric
- is more robust

Accelerated failure time (AFT) model

$$S(t|\underline{x}) = S_0^*(\exp(-\underline{x}'\underline{\beta}^*)t)$$

- is built on failure time T
- is parametric
- is less robust

Model fitting and diagnostics

General idea of most visual tests:

- find a linear relationship that holds if assumptions are fulfilled
- check if this linearity holds for given data

For example

- Weibull model: $\ln(t)$ vs $\ln(-\ln(S(t)))$
- log-logistic model: $\ln(t)$ vs $\left(-\ln \frac{S(t)}{1-S(t)}\right)$
- PH assumption: $\ln(-\ln S(t|\underline{x}))$ vs $\ln(-\ln S_0(t))$

More tools:

- Akaike information criterion (AIC)
- different residuals to test PH assumption
    - Cox-Snell
    - martingale
    - deviance
    - Schoenfeld

Multistate models

Usable in various practical scenarios:

- illness-death model
- competing risks model
- progressive model
- ...

Main points

- modelled as (non-homogeneous) continuous time Markov chains
- hazard function as transition intensity
- Cox PH idea applicable here as well – Cox-Markov model
- empirical estimates for survival and transition probabilities (Aalen-Johansen estimators)